CS231N Final Report: Investigating a Shared Embedding Space for Image to Audio Object-Centric Data

Suzannah Wistreich
Stanford University
suzannah@stanford.edu

Abstract

We present a self-supervised method to learn a shared embedding space for audio and image data for everyday objects. Our approach centers around the use of pretrained models AST (Audio Spectrogram Transformer) and DINOv2. We train AST with frozen DINOv2, and project their embeddings into a shared space to improve cross-sensory retrieval tasks. We utilize 100,000 image-audio pairs from ObjectFolder and 3,000 test points from ObjectFolderReal to evaluate our model. We find that our method shows significant results for inter-object retrieval tasks, but currently performs at-chance for intra-object retrieval. We investigate this performance disparity in the Discussion. Additionally, we test and address our model's ability to perform sim-to-real learning.

1. Introduction

In this report, we explore the integration of audio and image data into a shared multimodal embedding space, with an emphasis on object-centric data. The identification and use of everyday objects is often essential for interacting in one's environment, and therefore is a topic of interest for many embodied artificial intelligence tasks. Humans meaningfully learn about objects via multimodal data (rather than disjointed observations taken in isolation), so we hypothesize that efficiently learning this correspondence can greatly improve a model's performance on intuitive object recognition, and has uses for downstream robotics tasks.

To first evaluate baseline AST's ability to classify our audio of interest, we utilize the provided top-k classification inference metrics. To implement this joint embedding space, we took two independent models, AST [10] and DINOv2 [1] , and trained AST with frozen DINOv2 such that their learned features are projected to a shared embedding space. Then, after the integration of AST and DINOv2, we evaluate the joint embedding space on a variety of cross-sensory retrieval tasks.

1.1. Problem Statement

The primary goal of this project is to integrate the embedding of audio and image data from well-performing models into a shared multimodal embedding space for object-centric applications. Specifically, the task is to utilize a pretrained Audio Spectrogram Transformer (AST), in conjunction with the frozen DINOv2 model. The pretrained AST we utilize was trained on AudioSet [6] , a large-scale dataset with diverse audio-text label pairings. The evaluation of this shared embedding space will be assessed by the ability to facilitate effective cross-modal retrieval.

1.2. Method Inputs and Outputs

The inputs to our model consist of approximately 100,000 image-audio data pairs. These pairs are from the ObjectFolder dataset [4], developed in Stanford's Vision and Learning Laboratory, which contains high-quality neural (simulated) visual, auditory, and tactile representations of everyday objects. We then use a paired AST and DinoV2 architecture to output a shared embedding space of audio and visual data. However, to evaulate these embedding outputs, we utilize a variety of cross-sensory retrieval inference tasks. The outputs of these cross-sensory retrieval tasks are the mean Average Precision (mAP) between the embeddings of DINOv2's visual features and AST's impact audio features.

We define the cross-model retrieval task general as follows: given either the audio or image embedding of a specific object and point on that object, identify the embedding of the other modality of interest corresponding to the *same point* of the same object. There are two configurations for our cross-sensory retrieval tasks:

- **Inter-Object Retrieval:** Retrieving the correct modality embedding for a single point from each of 100 different validation objects.

- **Intra-Object Retrieval:** Retrieving the correct modality embedding for k points, each from *within* a single object.

  Please see *Methods* for more implementation details.

In this paper, we present the results of our joint-embedding Inter-Object and Intra-Object retrieval tasks. We also present results of the single-modality AST method, which serves as a baseline and is detailed below.

1.3. Baseline Method

The Audio Spectrogram Transformer model provides a comprehensive method to run inference on an arbitrary amount of datapoints, provided that the researcher prepares audio and text-label pairs. We run inference on approximately 90,000 datapoints from ObjectFolder, and present the result of the baseline method in *Experiments Results*.

2. Related Work

2.1. Single-Modality Learning

Multimodal work is frequently built upon the architectures of single-modality learning. Here we highlight the two main previous works that we build our joint audio-vision work upon: AST (Audio Spectrogram Transformer) and DINOv2.

The AST model [10] is introduced by Y. Gong et al., presenting the first convolutional-free approach to learning audio embeddings. AST has shown strong performance in audio classification across a default set of 527 classes, indicating strong embeddings and the ability to learn more fine-grained audio features. However, the current AST framework is suited for audio-text label relationships and classification. For the integration across more modalities, such as vision and tactile, we have the opportunity to utilize AST's learned embeddings and integrate them with visual data for a deeper multimodal understanding, rather than relying on bridging these modailities via text labels.

DINOv2 [1] , developed by Facebook Research, is a self-supervised algorithm for learning visual features. DINOv2 uses teacher-student architecture to extract image features very efficiently, without relying on labeled data. DINOv2 utilizes a Vision Transformer (ViT) as its main architecture, and is widely used as a stable backbone for various computer vision tasks. For this reason, we employ DINOv2 as our principle visual model, and use it off-the-shelf without additional training.

2.2. Multimodal Supervised Learning

We begin by examining multimodal visual models with supervised approaches. PolyVit [14], introduced by Likhosherstov et al., cotrains on image, audio, and video data with a single transformer and shared parameters. PolyVit takes a supervised approach by training on a variety of robust annotated datasets such as CIFAR-10 [13], Kinetics 400 [12], and Audioset [6].

OMNIVORE [9], a method by Girdhar, R., Singh, M., Ravi, N., et al. trains a single vision model with multiple visual modalities: images, videos, and single-view 3D. This model utilized a transformer-based architecture and excels at classification tasks for the three visual data formats used. They also found that cross-modal correspondences emerged without explicit training, indicating the model's ability for generalization across new modality pairings. This zero-shot capability would be particularly desirable for down-stream embodied AI tasks, and is something we hope to emulate. Like PolyVit, OMNIVORE is a supervised approach by training on labled datasets such as ImageNet [2], Kinetics-400 [12], and SUN RGB-D [17].

Recently, the Touch-Vision-Language (TVL) Model [3], developed by Fu, Datta, Huang, Panitch, Drake, et al., was presented as a unique approach to integrate tactile data with vision and language. They utilized human annotations and generated pseudo-label, which was proved to achieve significant multimodal understanding.

While these approaches achieve state-of-the-art results, we are interested in exploring self-supervised approaches without explicit labels. This partly arises from a necessity, as object-centric labeled data is relatively scarce.

2.3. Multimodal Self-Supervised Learning

In addition to multimodal supervised learning methods, there have been various approaches to self-supervised learning to jointly embed features such as audio, vision, tactile, etc.

Nagrani et al. [16] present a novel video-mining pipeline to create a weakly-labeled dataset for video-to-audio and caption data. This is a clever approach to address a gap between image-captioned and video-captioned data, but still relies on some form of captioning.

As an extension of OMNIVORE, OmniMAE [7] utilizes a novel approach of training a ViT with a masked autoencoder (MAE) for images and videos. This masking allows for efficient data processing and does not require any labeled data, making it a fully self-supervised strategy.

AudioCLIP [11] extends the CLIP model to integrate audio and image learning along with image and text. They utilize a blend of CLIP ViT and ResNeXt architectures to make this tri-modal correspondence and is fully self-supervised.

Lit (Locked-image Tuning) [18], by Zhai et al., also performs contrastive self-supervised learning for image and text modalities. Additionally, the authors highlight that freezing a strong image-encoder led to optimal results. This, along with ImageBind, encouraged us to freeze DINOv2 in our own methods.

Perhaps most similar to our method is the ImageBind framework [8], developed at Meta by R. Girdhar et al. ImageBind presents a novel approach to creating a shared embedding space across multiple modalities of data including images, audio, text, audio, and video. This research our most relevant predecessor as it addresses the core task of self-supervised multimodal learning via joint embedding spaces. By this method, ImageBind is able to perform a wide range of retrieval and zero or few-shot generation tasks, displaying the ability to utilize joint embeddings between modalities it was not directly trained on. However, ImageBind seems to underperform significantly with object-centric data. We hypothesize that this is due to the model's training on a broad range of category types, which is useful for general-purpose models. However, it would follow that for object-specific tasks, there was likely an insufficient amount of object training data for ImageBind to perform well on these tasks. This motivates our project's use of a multimodal embedding space, with a emphasis on object-centric data.

3. Methods

3.1. Integration of AST and DINOv2

3.1.1 Pretraining and Architecture Details

The AST model is pretrained on AudioSet [6], a robust audio dataset which allows AST's pretrained weights to handle a wide variety of inputs. The model uses a frequency stride of 10 and a time stride of 10 with overlapping patches, allowing for more fine-grained feature extraction for object impact sounds. AST proccesses audio data by adapting the Vision Transformer architecture to audio spectrograms. This strategy treats the time-frequency representation essentially how image pixels are treated in vision tasks.

DINOv2 is fully-frozen during our training process, as DINOv2 is a very strong off-the-shelf model and allows for consistent visual features over training. DINOv2 also uses a Vision Transformer as its backbone architecture. DINOv2's frozen model involved training a student network to predict the output of a teacher network, allowing the student to learn consistent image features over multiple views of the same image. Images must be re-sized and normalized prior to passing to DINOv2.

In general, a ViT operates by extending the transformer natural language processing architecture to image recognition tasks. Images are first divided into sub-patches (which can have overlap, depending on the stride), then flattened and lineraly embedded. To still encode positional information of the pixels, positional embeddings are added to the inputs. These patches are processed by multiple self-attention mechanisms, and are very useful for vision-based tasks such as natural images or spectrograms.

3.1.2 Training

AST processes audio waveforms into mel spectrograms, then passes these spectrograms through the AST model to generate embeddings. Frozen DINOv2 simultaneously processes images from the training set. Training pairs of audio-text data are pulled from rendered directories of ObjectFolder data. AST's weights are updated during the training process and optimized with AdamW, with the intention of improving from the AST baseline presented in *Experiments Results*.

3.2. Evaluation Implementation

3.2.1 Inter-Object Retrieval

Inter-object retrieval is implemented by first processing embeddings from two different modalities (audio and vision) for the same datapoint pair. The cosine similarity scores are then computed between these embeddings, defined as:

$$\mathbf{x}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|}, \quad \mathbf{x}_2 = \frac{\mathbf{x}_2}{\|\mathbf{x}_2\|}$$

Finally, an evaluation of mean Average Precision (mAP) is done to measure retrieval performance. mAP is calcuated as:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{rank}_i + 1}$$

Where N is the number of queries and the ranki is the rank position of the correct class for the i-th query.

The algorithm iterates over a DataLoader we designed to load audio and image data pairs. The DataLoader serves batches of audio-visual data pairs. The DataLoader is instructed to load $n-way$ objects, with only one audio-vision data pair each.

3.2.2 Intra-Object Retrieval

Intra-object retrieval is implemented similarly as inter-object retrieval, but instructs the DataLoader to only load point pairs from within a single object.

3.3. Loss

We optimize a combined InfoNCE loss function to encourage corresponding audio and image embeddings to be close together in the shared embedding space. This allows dissimilar pairs to be further apart in the embedding space as well. InfoNCE loss is defined as follows:

$$\mathcal{L}_{\text{InfoNCE}} = -\log\left(\frac{\exp(\text{sim}(u, v)/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(u, v_k)/\tau)}\right)$$

Where u and v are the embeddings of positive pairs, v_k are the embeddings of negative samples, τ is the temperature parameter, and sim is a cosine similarity function.

3.4. Baseline Method

To contrast the mAP of our proposed method, we employ the provided Baseline AST method. We performed inference on ObjectFolder data *without* the additional training on object-centric data, to get a baseline mAP for audio-text label classification. The off-the-shelf AST used for this task is the same pretrained version we use on our proposed method.

3.5. Evaluation

We evaluate the baseline method via results from the provided AST classification task implementation. This implementation takes an arbitrary number of audio-text label pairs, and performs inference on these pairs, returning mAP, AUC, and d-prime metrics for the top-k resulting class labels. At least 2 correct labels must be supplied for each validation point, from a total of 527 classes. The results of this baseline are in *Experiments Results, 5.3*.

We evaluate our proposed model on the cross-sensory retrieval tasks as described in Section 2. The results of the method implemented for this milestone (Intra-Object Retrieval) as well as a comparison method that was already completed (Inter-Object Retrieval) are presented in Section 5.

4. Dataset

4.1. Simulated and Real Datasets

The primary dataset used for this project is R. Gao et al.'s ObjectFolder [4] , published from Stanford's Vision and Learning Laboratory. ObjectFolder has approximately 100,000 simulated datapoints with visual, auditory, and tactile information. The dataset is composed of 1,000 distinct everyday objects, with approximately 100 datapoints per object. We use 90% of ObjectFolder data for training, (approximately 90,000 points across 90 objects), and 10% for validation (approximately 10,000 points across 10 objects). Note that we separate our train and validation data by object: an object's datapoints are either fully in the training set or fully in the validation set.
Finally, we present testing results from a related dataset, Gao et al.'s ObjectFolderReal [5]. ObjectFolderReal, also published from Stanford's Vision and Learning Laboratory, has approximately 3,000 real datapoints of visual, auditory, and tactile information. There are 100 objects in this dataset, with about 30 points per object. We selected this as our testing set for its quality of being similar to ObjectFolder yet unseen by our model. We are also interested in our model's ability to perform sim-to-real transfer.

4.2. Preprocessing

4.2.1 Rendering of ObjectFolder Real Data

The ObjectFolder data is neural data (simulated), meaning that preprocessing involves utilizing rendering scripts for the waveform and image data. The waveform data were already rendered in our SVL cluster. Waveforms are coverted to spectrograms with 44.1kHZ and 256 mel-frequency bins.

For the images, we had to create a rendering script for the images to meet desired specifications. First, ObjectFolderReal supplied point and normal information for each datapoint impact. Therefore, we wanted to generate a textured rendering of an object in space pointing at the point of impact contact, with the camera positioned along the normal to that point. We also wanted the camera distance from the object's surface to be the same distance as the ObjectFolder data for consistency. Our images are rendered at 512x512 pixels. Rendering image and spectrogram samples are given in Figure 1 and 2.

Figure 1: ObjectFolderReal Rendering Examples

4.2.2 Compatibility with AST

AST incorporates the preprocessing of waveform to spectrogram within its own pipeline, but both AST and DINOv2 require some normalization of inputs before passing to their respective models. AST data is normalized to a mean of -3.739 and standard deviation of 6.697., so we incorporate this step to our preprocessing pipeline.

To run the AST baseline method, which was classification inference on audio-text label pairs, it was also required to map audios to one of 527 default AST text-label classes. As this inference was simply used as a baseline, we hand-mapped object impact sounds to classes based on the object's material. ObjectFolder has objects from 7 different material classes, the mapping rule for this is presented in Table 1.

5. Experiments & Results

5.1. Hyperparameters

We utilize a learning rate of $1e-5$ and AST's pre-set weight decay of $5e-7$. We experimented with a vari-

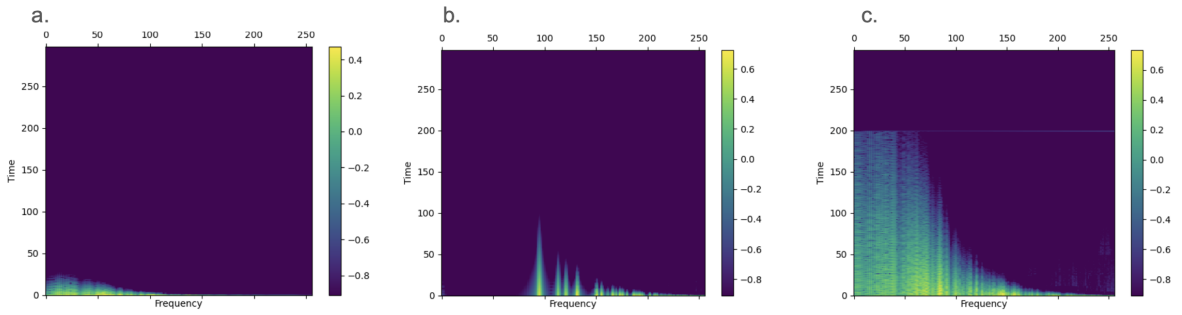

Figure 2: Spectrogram Rendering Examples.
2a Spectrogram of a wooden spatula. **2b** Spectrogram of a ceramic bowl. **2c** Spectrogram of an iron skillet.

Material	AST Class Label
Ceramic	clank, clink, dishes pots pans
Wood	wood, wood block
Glass	glass, clank, clink
Iron	clang, cowbell
Plastic	clatter, thunk
Polycarbonate	clatter, thunk
Steel	clang, reverb

Table 1: Mapping of materials to AST class labels

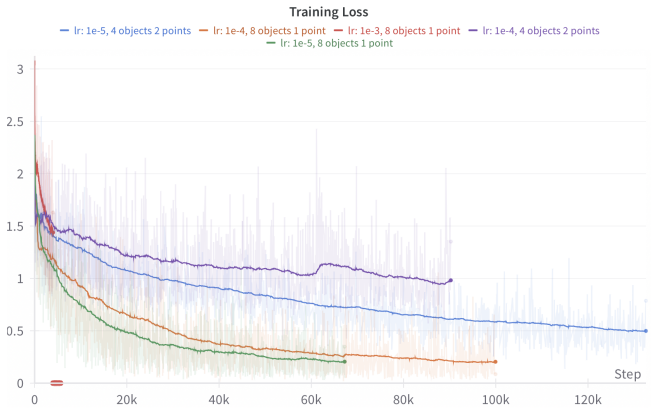

Figure 3: Visualization of Training Loss with Various Hyperparameters

ety of learning rates, and found that learning rates greater than $1e-3$ lead to exploding gradients. Values of $1e-4$ and $1e-5$ both seemed to converge, but $1e-5$ had significant improvements for training with 4 objects and two points each, so we selected $1e-5$ as our final learning rate. Figure 3 visualizes these details.

Due to the computational needs of the AST model, we were limited to 8 points per training batch. However, we experimented with how to split these 8 points per batch: either 8 objects with 1 point each, or 4 objects with 2 points each. We found that 8 objects with 1 point each led to more significant convergence, as detailed in Figure 3. We did not perform cross-validation.

5.2. Optimization

We use AdamW [15] for our optimizer, with the learning rate and weight decay specified above, to update the AST weights by the computed gradient from our loss function. This is the same optimizer used by the pre-trained AST model, as AdamW often yields better training loss and generalization to unseen data.

5.3. Baseline Classification Results

AST's off-the-shelf classification reports the metrics of mAP, AUC, and d-prime. mAP (mean Average Precision) measures the average precision across recall and retrieval tasks. In AST's baseline case, this is the average precision across recall levels for each class label. AUC is the area under the ROC curve, and is a measure of the performance of binary classification. d-prime is a measure of the model's ability to distinguish between the target and noise.
From running a validation sample of 300 ObjectFolder datapoints on pretrained AST, we got a audio-label classification mAP of 0.162, AUC of 0.523, and d-prime of 0.074. Without any additional training, this mAP showed some promis-

ing result of using AST's embeddings, especially when considering our the model's ability to predict our loosely-labeled data out of 527 class label. This encouraged us to perform Experiments 1-3 after joint AST and DINOv2 training.

5.4. Multimodal Cross Sensory Retrieval Results

5.4.1 Experiment 1: Inter-Object Retrieval

We continue with the primary metric of mAP. Our first experiment was Inter-Object Retrieval. We perform cross-sensory retrieval with either 20, 50, and 100 objects, with one point per object. Results for Experiment 1 as compared to chance performance are tabled in Table 2.

Retrieval Type	Ours (%)	Random (%)
(20-way)		
DINOv2 → AST	48.98	18.0
AST → Dino	49.48	18.0
(50-way)		
DINOv2 → AST	25.82	10.0
AST → Dino	26.81	10.0
(100-way)		
DINOv2 → AST	24.1	5.19
AST → Dino	25.5	5.19

Table 2: Inter-Object Audio ↔ DINOv2 Cross-Modal Retrieval

5.4.2 Experiment 2: Inter-Object Retrieval, with Spectrogram Averaging

We were also interested in how inter-object retrieval would perform when treating audio as more of a global versus local feature. To test this, we sampled n points per object, took the *average* of the audio spectrogram from these points, then performed the same inter-object retrieval task as defined above.

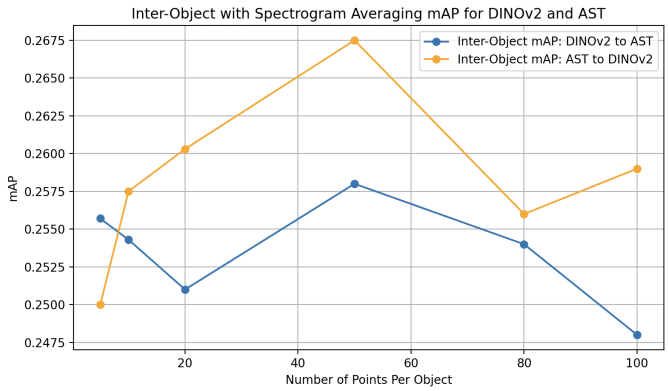

Figure 4: Inter-Object mAP (with Spectrogram Averaging) versus Number of Points per Object

We fixed the task to 100 objects, but found that the number of points sampled per object seemed to influence mAP. Our findings for this experiment are detailed above in Figure 4.

5.4.3 Experiment 3: Intra-Object Retrieval

We perform intra-object retrieval, as defined in *Methods*, across varying numbers of points. These results are tabulated in Table 3.

Retrieval Type	Ours (%)	Random (%)
(10-way)		
DINOv2 → AST	29.6	29.02
AST → Dino	28.8	29.02
(20-way)		
DINOv2 → AST	18.2	18.0
AST → Dino	17.3	18.0
(100-way)		
DINOv2 → AST	5.21	5.19
AST → Dino	5.38	5.19

5.4.4 Experiment 4: Inter-Object Sim-to-Real Testing

Finally, we test our model's ability for sim-to-real learning transfer. As intra-object retrieval appeared at-chance performance, we just perform Experiment 4 on inter-object retrieval. Results for 2, 5, 10, 20, and 100-way retrieval are visualized below in Figure 5.

5.5. Discussion

5.5.1 Inter-Object Discussion

In Experiment 1, we find that our method is quite effective for inter-object retrieval, achieving mAP scores that are significantly above chance for 20, 50, and 100-way object

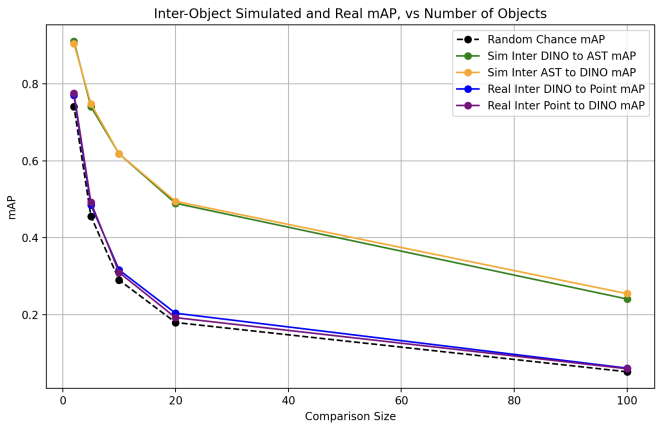

Figure 5: Inter-Object Sim-To-Real Testing

retrieval tasks. We believe that this is because of our large-quantity of robust training data that has a variety of everyday objects. Additionally, when visualizing the ObjectFolder spectrograms, it seems that the material of an object is a main predictor in spectrogram features. This qualitative observation can be seen in Figure 2. Additionally, we believe that we encountered some over-fitting to our training data, as we achieved very small loss values for train loss, but after a significant amount of training (1̃00k steps), our validation loss began to diverge. However, all inter-object results are from the validation set and are still significant, so we did not mitigate the overfitting for now.

We find that performing inter-object retrieval with spectrogram-averaging (Experiment 2) yields comparably significant results as Experiment 1. We also find that a ”medium” amount of points per object (eg. 20-80 points) yielded the highest mAP of 26.8, and was slightly higher than Experiment 1. This finding aligns with our intuition: if we take the spectrogram average over a few object points (eg. 1-10 points), that is essentially equivalent to Experiment 1. However, taking the average over too many intra-object spectrograms could muddle the signal too much, especially for objects with more variable spectrograms. Again citing Figure 2, materials with more reverb (such as iron and steel) tended to have more variability in their spectrograms.

5.5.2 Intra-Object Discussion Investigation

For Experiment 3, we find that our method is at-chance, and it seems that our joint audio-image embeddings are not sufficiently fine-grained for intra-object retrieval. After some reflection, while this is a harder task than inter-object cross-sensory retrieval, we concluded that humans are able to still semi-reliably discriminate different impacts audio locations on the same object. This prompted us to investigate the quality of our spectrograms before processing by AST. Our hypothesis was that the spectrograms were likely not detailed enough to capture more subtle differences between intra-object audios.

We found that our original spectrograms had two main sub-optimal features. First, due to the smaller number of timesteps used in our impact-audio waveforms, we found that only 1/3 of the spectrogram actually contained the audio signal, and 2/3's was padded by AST's default configurations. Second, we found that the original spectrograms used 128 mel-frequency bins, which often did not sufficiently capture the differences between intra-object spectrograms. To confirmed this insufficiency, we plotted the L1 differences between two intra-object points for a few sample objects. These failing points are highlighted in Figure 6 and 7.

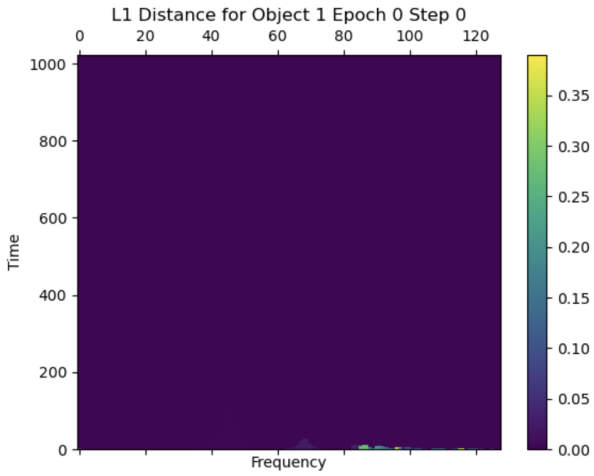

Figure 7: L1 Distance for Spectrogram Samples from Figure 6

After investigating these failing points, we adjusted the expected time steps for AST, and increased mel-frequency bins from 128 to 256. Our next step is to re-train and run Experiment 3 with this new configuration.

5.5.3 Sim-To-Real Gap

Finally, it seems that our current model is unsuccessful at bridging the sim-to-real learning gap, as inter-object performance on ObjectFolderReal data is only marginally above chance. We hypothesize this might be in-part due to the spectrogram deficiencies described in the previous section. We also aim to investigate if the lower-fidelity of simulated versus real data becomes more apparent in spectrogram visualizations than waveform audios. If this is the case, then AST could face difficulties in cross-sensory retrieval of real

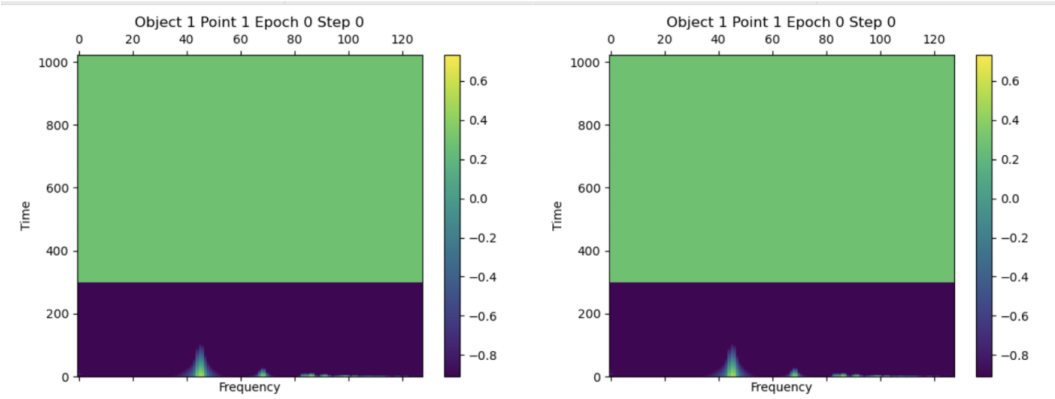

Figure 6: Intra-Object Spectrogram sample, featuring AST default padding and insufficient variation

data with higher-fidelity spectrograms.

6. Conclusion & Future Work

In this study, we successfully designed and evaluated model for shared embeddings between audio and image data, focusing on object-centric data. We integrated AST for audio processing and frozen DINOv2 for images into a joint model, which proved to be highly effective for inter-ojbect cross sensory retrieval tasks. Our results show significant improvement over baseline single-modality classification, which highlights the potential for this approach to be used for more intuitive object recognition and downstream robotics tasks.

Our future work will include re-training our architecture with the improved spectrogram configuration, with the intention of improving intra-object cross-sensory retrieval performance. Additionally, we will investigate the sim-to-real gap our model faces, as described in the Discussion. In a larger context, this work is part of a larger work, ObjectBind, is an ongoing project in multimodal object-centric foundation models. Objectbind aims to integrate modalities such as images, mesh, audio, and tactile data into a shared embedding space. Our hope for this project is to develop an all-inclusive model with intuitive object-recognition abilities, as the identification and interaction with objects is fundamental for humans and embodied agents alike to perform everyday tasks.

7. Contributions & Acknowledgements

This project is part of a larger work, ObjectBind, is an ongoing project in multimodal object-centric foundation models. This research is part of Stanford Vision and Learning Laboratory, and I am advised by Dr. Jiajun Wu. My PhD mentor is Samuel Clarke (noted as S.C. in contributions). My CS231n project focuses on the audio-to-image work for Objectbind.

Author Information
Suzannah Wistreich (S.W.) is the sole author of this paper.

S.W. implemented the code for inter-object with spectrogram averaging and intra-object retrieval code (Experiments 2 and 3).
S.W. wrote the script for the rendering of ObjectFolderReal data, and did the investigation of how to make intra-object retireval more reliable.
S.C. implemented the inter-object code (Experiment 1). Our training script was already in-place and derived from our mesh-to-image work, which was written by S.C. with modifications by S.W.

We thank AST (Audio Spectrogram Transformer) and DINOv2 for the use of their pretrained models.

- AST Repository: https://github.com/YuanGongND/ast.git
- DINOv2 Repository: https://github.com/facebookresearch/dinov2.git

We thank the Stanford Vision and Learning Laboratory for their job scheduling and GPU usage.

References

[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 1, 2

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 2

[3] L. Fu, G. Datta, H. Huang, W. C.-H. Panitch, J. Drake, J. Ortiz, M. Mukadam, M. Lambeta, R. Calandra, and K. Gold-

berg. A touch, vision, and language dataset for multimodal alignment. *arXiv preprint arXiv:2402.13232*, 2024. 2

[4] R. Gao, Y.-Y. Chang, S. Mall, L. Fei-Fei, and J. Wu. Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations. In *Conference on Robot Learning (CoRL)*, 2021. 1, 4

[5] R. Gao, Y. Dou, H. Li, T. Agarwal, J. Bohg, Y. Li, L. Fei-Fei, and J. Wu. The objectfolder benchmark: Multisensory learning with neural and real objects. *arXiv preprint arXiv:2306.00956*, 2023. 4

[6] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audioset: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 1, 2, 3

[7] R. Girdhar, A. El-Nouby, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Omnimae: Single model masked pretraining on images and videos. *arXiv preprint arXiv:2206.08356*, 2023. 2

[8] R. Girdhar et al. Imagebind: One embedding space to bind them all. *arXiv preprint arXiv:2105.05665*, 2021. 3

[9] R. Girdhar, M. Singh, N. Ravi, A. Joulin, I. Misra, and P. Goyal. Omnivore: A single model for many visual modalities. *arXiv preprint arXiv:2201.08377*, 2022. 2

[10] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass. Ast: Audio spectrogram transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10699–10709, 2022. 1, 2

[11] A. Guzhov, F. Raue, J. Hees, and A. Dengel. Audioclip: Extending clip to image, text, and audio. *arXiv preprint arXiv:2106.13043*, 2021. 2

[12] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 2

[13] A. Krizhevsky. Learning multiple layers of features from tiny images. 2009. 2

[14] V. Likhosherstov, A. Arnab, K. Choromanski, M. Lucic, Y. Tay, A. Weller, and M. Dehghani. Polyvit: Co-training vision transformers on images, videos, and audio. *arXiv preprint arXiv:2111.12993*, 2022. 2

[15] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019. 5

[16] A. Nagrani, P. H. Seo, B. Seybold, A. Hauth, S. Manen, C. Sun, and C. Schmid. Learning audio-video modalities from image captions. *arXiv preprint arXiv:2204.00679*, 2022. 2

[17] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 2

[18] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer. Locked-image tuning: Training text models to see better. *arXiv preprint arXiv:2111.07991*, 2022. 2