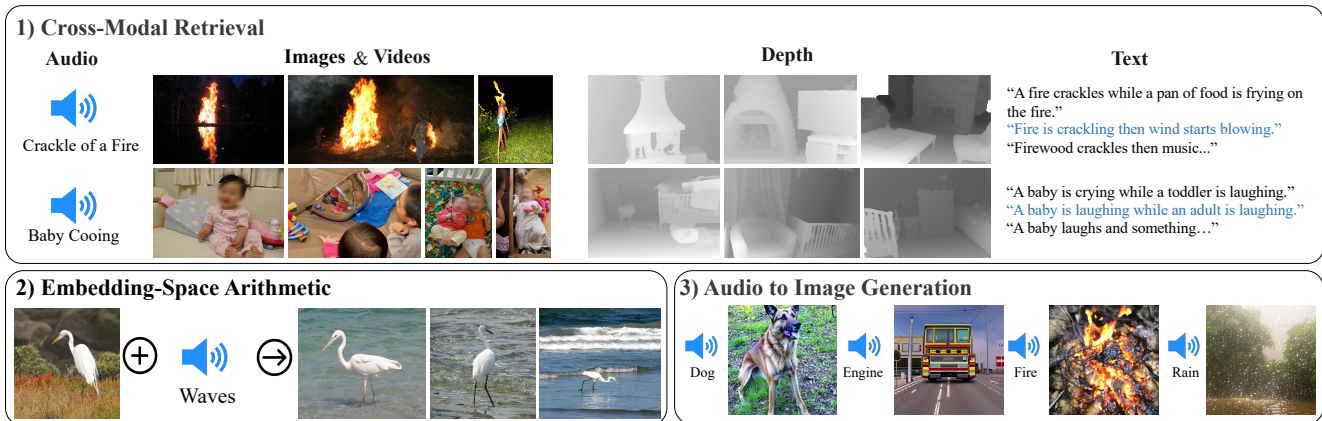


# IMAGEBIND: One Embedding Space To Bind Them All

Rohit Girdhar\*    Alaaeldin El-Nouby\*    Zhuang Liu    Mannat Singh  
Kalyan Vasudev Alwala    Armand Joulin    Ishan Misra\*  
FAIR, Meta AI

<https://facebookresearch.github.io/ImageBind>



**Figure 1.** IMAGEBIND’s joint embedding space enables novel multimodal capabilities. By aligning six modalities’ embedding into a common space, IMAGEBIND enables: **1)** Cross-Modal Retrieval, which shows *emergent* alignment of modalities such as audio, depth or text, that aren’t observed together. **2)** Adding embeddings from different modalities naturally composes their semantics. And **3)** Audio-to-Image generation, by using our audio embeddings with a pre-trained DALLE-2 [60] decoder designed to work with CLIP text embeddings.

## Abstract

We present IMAGEBIND, an approach to learn a joint embedding across six different modalities - images, text, audio, depth, thermal, and IMU data. We show that all combinations of paired data are not necessary to train such a joint embedding, and only image-paired data is sufficient to bind the modalities together. IMAGEBIND can leverage recent large scale vision-language models, and extends their zero-shot capabilities to new modalities just by using their natural pairing with images. It enables novel emergent applications ‘out-of-the-box’ including cross-modal retrieval, composing modalities with arithmetic, cross-modal detection and generation. The emergent capabilities improve with the strength of the image encoder and we set a new state-of-the-art on emergent zero-shot recognition tasks across modalities, outperforming specialist supervised models. Finally, we show strong few-shot recognition results outperforming prior work, and that IMAGEBIND serves as a new way to evaluate vision models for visual and non-visual tasks.

\*Equal technical contribution.

## 1. Introduction

A single image can bind together many experiences – an image of a beach can remind us of the sound of waves, the texture of the sand, a breeze, or even inspire a poem. This ‘binding’ property of images offers many sources of supervision to learn visual features, by aligning them with any of the sensory experiences associated with images. Ideally, for a single joint embedding space, visual features should be learned by aligning to all of these sensors. However, this requires acquiring all types and combinations of paired data with the same set of images, which is infeasible.

Recently, many methods learn image features aligned with text [1, 30, 45, 59, 63, 80, 81], audio [3, 4, 49, 54, 55, 68] etc. These methods use a single pair of modalities or, at best, a few visual modalities. However, the final embeddings are limited to the pairs of modalities used for training. Thus, video-audio embeddings cannot directly be used for image-text tasks and vice versa. A major obstacle in learning a true joint embedding is the absence of large quantities of multimodal data where all modalities are present together.

In this paper, we present **IMAGEBIND**, which learns a single shared representation space by leveraging multiple types of image-paired data. It does not need datasets where all modalities co-occur with each other. Instead, we leverage the binding property of images and we show that just aligning each modality’s embedding to image embeddings leads to an emergent alignment across all of the modalities. In practice, **IMAGEBIND** leverages web-scale (image, text) paired data and combines it with naturally occurring paired data such as (video, audio), (image, depth) *etc.* to learn a single joint embedding space. This allows **IMAGEBIND** to implicitly align the text embeddings to other modalities such as audio, depth *etc.*, enabling zero-shot recognition capabilities on that modality without explicit semantic or textual pairing. Moreover, we show that it can be initialized with large-scale vision-language models such as CLIP [59], thereby leveraging the rich image and text representations of these models. Thus, **IMAGEBIND** can be applied to a variety of different modalities and tasks with little training.

We use large-scale image-text paired data along with naturally paired ‘self-supervised’ data across four new modalities - audio, depth, thermal, and Inertial Measurement Unit (IMU) readings – and show strong emergent zero-shot classification and retrieval performance on tasks for each of these modalities. These emergent properties improve as the underlying image representation is made stronger. On audio classification and retrieval benchmarks, **IMAGEBIND**’s emergent zero-shot classification matches or outperforms specialist models trained with direct audio-text supervision on benchmarks like ESC, Clotho, AudioCaps. **IMAGEBIND** representations also outperform specialist supervised models on few-shot evaluation benchmarks. Finally, we show that **IMAGEBIND**’s joint embeddings can be used for a wide variety of compositional tasks as illustrated in Figure 1, including cross-modal retrieval, combining embeddings via arithmetic, detecting audio sources in images, and generating images given audio input.

## 2. Related Work

**IMAGEBIND** builds upon several advances in vision-language, multimodal, and self-supervised research.

**Language Image Pre-training.** Training images jointly with linguistic signals like words or sentences has been shown to be an effective method for zero-shot, open-vocabulary recognition and text to image retrieval [13, 17, 37, 66]. Language as supervision can further be used for learning strong video representations [2, 46, 47]. Joulin et al. [33] show that using large-scale image dataset with noisy captions yields strong visual features. Recently, CLIP [59], ALIGN [30] and Florence [81] collect large collections of image and text pairs and train models to embed image and language inputs in a joint space using contrastive learning, exhibiting impressive zero-shot performance. CoCa [80]

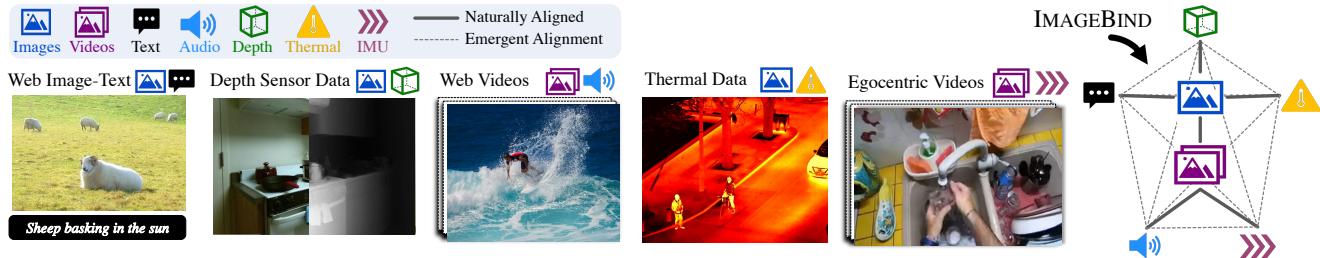
adds an image captioning objective on top of the contrastive loss for improved performance. Flamingo [1] handles arbitrarily interleaved images and texts, and achieves state of the art on many few-shot learning benchmarks. LiT [82] adopts contrastive training for fine-tuning and observes freezing image encoders works the best. This prior line of works mostly considers image and text, while our work enables zero-shot recognition on multiple modalities.

**Multi-Modal Learning.** Our work binds multiple modality representations in a joint embedding space. Prior works explored joint training of multiple modalities in a supervised [20, 41] or self-supervised contexts [3, 19, 49, 68, 72]. The success of image and language pre-training methods such as CLIP has inspired approaches that revisits learning deep semantic representations through matching other modalities with linguistic inputs. Various methods adapt CLIP to extract semantically strong video representations [14, 42, 44, 77]. Most related to our method, Nagrani et al. [50] create a weakly-labeled dataset for paired video-audio and captions that allows for training multi-modal video-audio encoder to match textual features resulting in strong audio and video retrieval and captioning performance. AudioCLIP [26] adds audio as an additional modality into a CLIP framework, enabling zero-shot audio classification. In contrast, **IMAGEBIND** does not require explicit paired data between all modalities and instead leverages image as a natural weak supervision for unifying modalities.

**Feature Alignment** Pre-trained CLIP models have been utilized as teachers to supervise other models due to the strength of its visual representations [43, 57, 73]. Moreover, CLIP joint image and text embedding space has also been leveraged for a variety of zero-shot tasks like detection [23, 86], segmentation [40], mesh animation [79] *etc.* showing the power of joint embedding spaces. PointCLIP [83] finds a pre-trained CLIP encoder can be used for 3D recognition by projecting a point cloud to a number of 2D depth map views, which in turn are encoded using CLIP visual encoder. In multilingual neural machine translation, a similar phenomenon to the emergence behavior of **IMAGEBIND** is commonly observed and utilized: if languages are trained in the same latent space through learned implicit bridging, translation can be done between language pairs on which no paired data is provided [32, 39].

## 3. Method

Our goal is to learn a single joint embedding space for all modalities by using images to bind them together. We align each modality’s embedding to image embeddings, such as text to image using web data and IMU to video using video data captured from egocentric cameras with IMU. We show that the resulting embedding space has a powerful emergent zero-shot behavior that automatically associates pairs of modalities without seeing any training data for that spe-



**Figure 2. IMAGEBIND overview.** Different modalities occur naturally aligned in different data sources, for instance images+text and video+audio in web data, depth or thermal information with images, IMU data in videos captured with egocentric cameras, *etc.* IMAGEBIND links all these modalities in a common embedding space, enabling new emergent alignments and capabilities.

cific pair. We illustrate our approach in Figure 2.

### 3.1. Preliminaries

**Aligning specific pairs of modalities.** Contrastive learning [27] is a general technique for learning an embedding space by using pairs of related examples (positives) and unrelated examples (negatives). Using pairs of aligned observations, contrastive learning can **align pairs of modalities** such as (image, text) [59], (audio, text) [26], (image, depth) [68], (video, audio) [49] *etc.* However, in each case, the joint embeddings are trained and evaluated using the same pairs of modalities. Thus, (video, audio) embeddings are not directly applicable for text-based tasks while (image, text) embeddings cannot be applied for audio tasks.

**Zero-shot image classification using text prompts.** CLIP [59] popularized a ‘zero-shot’ classification task based on an aligned (image, text) embedding space. This involves constructing a list of text descriptions that describe the classes in a dataset. An input image is classified based on its similarity to the text descriptions in the embedding space. Unlocking such zero-shot classification for other modalities requires specifically training using paired text data, *e.g.*, (audio, text) [26] or (point-clouds, text) [83]. In contrast, IMAGEBIND unlocks zero-shot classification for modalities *without* paired text data.

### 3.2. Binding modalities with images

IMAGEBIND uses pairs of modalities  $(\mathcal{I}, \mathcal{M})$ , where  $\mathcal{I}$  represents images and  $\mathcal{M}$  is another modality, to learn a single joint embedding. We use large-scale web datasets with (image, text) pairings that span a wide range of semantic concepts. Additionally, we use the natural, self-supervised pairing of other modalities – audio, depth, thermal, and Inertial Measurement Unit (IMU) – with images.

Consider the pair of modalities  $(\mathcal{I}, \mathcal{M})$  with aligned observations. Given an image  $\mathbf{I}_i$  and its corresponding observation in the other modality  $\mathbf{M}_i$ , we encode them into normalized embeddings:  $\mathbf{q}_i = f(\mathbf{I}_i)$  and  $\mathbf{k}_i = g(\mathbf{M}_i)$  where  $f, g$  are deep networks. The embeddings and the encoders

are optimized using an InfoNCE [53] loss:

$$L_{\mathcal{I}, \mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_i / \tau)}{\exp(\mathbf{q}_i^\top \mathbf{k}_i / \tau) + \sum_{j \neq i} \exp(\mathbf{q}_i^\top \mathbf{k}_j / \tau)}, \quad (1)$$

where  $\tau$  is a scalar temperature that controls the smoothness of the softmax distribution and  $j$  denotes unrelated observations, also called ‘negatives’. We follow [74] and consider every example  $j \neq i$  in the mini-batch to be a negative. The loss makes the embeddings  $\mathbf{q}_i$  and  $\mathbf{k}_i$  closer in the joint embedding space, and thus aligns  $\mathcal{I}$  and  $\mathcal{M}$ . In practice, we use a symmetric loss  $L_{\mathcal{I}, \mathcal{M}} + L_{\mathcal{M}, \mathcal{I}}$ .

**Emergent alignment of unseen pairs of modalities.** IMAGEBIND uses modalities paired with images, *i.e.*, pairs of the form  $(\mathcal{I}, \mathcal{M})$  to align each the embeddings from each modality  $\mathcal{M}$  to those from images. We observe an emergent behavior in the embedding space that aligns two pairs of modalities  $(\mathcal{M}_1, \mathcal{M}_2)$  even though we only train using the pairs  $(\mathcal{I}, \mathcal{M}_1)$  and  $(\mathcal{I}, \mathcal{M}_2)$ . This behavior allows us to perform a wide variety of zero-shot and cross-modal retrieval tasks without training for them. We achieve state-of-the-art zero-shot text-audio classification results without observing a single sample of paired (audio, text).

### 3.3. Implementation Details

IMAGEBIND is conceptually simple and can be implemented in many different ways. We deliberately choose a vanilla implementation that is flexible and allows for an effective study and easy adoption. In § 5, we present design decisions that are critical for good emergent ‘binding’.

**Encoding modalities.** We use a Transformer architecture [71] for all the modality encoders. We use the Vision Transformer (ViT) [12] for images. Following [19], we use the same encoder for images and videos. We temporally inflate [7] the patch projection layer of the ViT and use 2 frame video clips sampled from 2 seconds. We follow [21] for encoding audio and convert a 2 second audio sampled at 16kHz into spectrograms using 128 mel-spectrogram bins. As the spectrogram is also a 2D signal like an image, we use a ViT with a patch size of 16 and stride 10. We treat thermal images and depth images as one-channel images and

| Dataset                         | Task          | #cls | Metric | #test |
|---------------------------------|---------------|------|--------|-------|
| Audioset Audio-only (AS-A) [18] | Audio cls.    | 527  | mAP    | 19048 |
| ESC 5-folds (ESC) [58]          | Audio cls.    | 50   | Acc    | 400   |
| Clotho (Clotho) [16]            | Retrieval     | -    | Recall | 1045  |
| AudioCaps (AudioCaps) [36]      | Retrieval     | -    | Recall | 796   |
| VGGSound (VGGs) [8]             | Audio cls.    | 309  | Acc    | 14073 |
| SUN Depth-only (SUN-D) [67]     | Scene cls.    | 19   | Acc    | 4660  |
| NYU-v2 Depth-only (NYU-D) [64]  | Scene cls.    | 10   | Acc    | 653   |
| LLVIP (LLVIP) [31]              | Person cls.   | 2    | Acc    | 15809 |
| Ego4D (Ego4D) [22]              | Scenario cls. | 108  | Acc    | 68865 |

**Table 1. Emergent zero-shot classification datasets** for **audio**, **depth**, **thermal**, and **Inertial Measurement Unit (IMU)** modalities. We evaluate IMAGEBIND without training for any of these tasks and without training on paired text data for these modalities. For each dataset, we report the task (classification or retrieval), number of classes (#cls), metric for evaluation (Accuracy or mean Average Precision), and the number of test samples (#test).

also use a ViT to encode them. We follow [20] to convert depth into disparity maps for scale invariance. We extract the IMU signal consisting of accelerometer and gyroscope measurements across the  $X$ ,  $Y$ , and  $Z$  axes. We use 5 second clips resulting in 2K time step IMU readings which are projected using a 1D convolution with a kernel size of 8. The resulting sequence is encoded using a Transformer. Finally, we follow the text encoder design from CLIP [59].

We use separate encoders for images, text, audio, thermal images, depth images, and IMU. We add a modality-specific linear projection head on each encoder to obtain a fixed size  $d$  dimensional embedding, that is normalized and used in the InfoNCE loss from Eq 1. In addition to ease of learning, this setup allows us to also initialize a subset of the encoders using pretrained models, *e.g.*, the image and text encoder using CLIP [59] or OpenCLIP [29].

## 4. Experiments

We first describe the main experimental setup and provide full details in the supplement.

**Naturally paired modalities and datasets.** We use IMAGEBIND on six modalities - image/video, text, audio, depth, thermal images, and IMU. As described in § 3.3, we treat videos as 2 frame images and process them the same as images. For the naturally available paired data, we use the (video, audio) pairs from the Audioset dataset [18], (image, depth) pairs from the SUN RGB-D dataset [67], (image, thermal) pairs from the LLVIP dataset [31] and (video, IMU) pairs from the Ego4D dataset [22]. For these pairs of modalities, we do not use any extra supervision like class labels, text *etc.* Since SUN RGB-D and LLVIP are relatively small, we follow [20] and replicate them 50 $\times$  for training.

**Large scale image-text pairs.** We leverage image-text supervision from large-scale web data [59]. For ease of experimentation, we use pretrained models that are trained on billions of (image, text) pairs. Specifically, we use the

pretrained vision (ViT-H 630M params) and text encoders (302M params) from OpenCLIP [29] in our experiments.

**Encoders for each modality.** We convert audio into 2D mel-spectrograms [21], and thermal and depth modalities into 1 channel images and use ViT-B, ViT-S encoders respectively. The image and text encoders are kept frozen during the IMAGEBIND training and the audio, depth, thermal, and IMU encoders are updated.







**Emergent zero-shot vs. zero-shot.** Methods such as CLIP [59], AudioCLIP [26] *etc.* train with modality pairs, (image, text) and (audio, text), to demonstrate zero-shot classification using text-prompts for the same modality. In contrast, IMAGEBIND binds modalities together using only image-paired data. Thus, just by training on (image, text) and (image, audio), IMAGEBIND can perform zero-shot classification of audio using text prompts. As we do not directly train for this ability, we term it *emergent* zero-shot classification to distinguish it from methods that specifically train using paired text-supervision for all modalities.

**Evaluation on downstream tasks.** We comprehensively evaluate IMAGEBIND on a many different downstream tasks using different protocols. We summarize the main datasets used for evaluation in Table 1.

### 4.1. Emergent zero-shot classification

We evaluate IMAGEBIND on emergent zero-shot classification and use the text prompt templates from [59] (full details in Appendix B). We report the results in Table 2. Each task measures IMAGEBIND’s ability to associate text embeddings to the other modalities without observing them together during training. Given the novelty of our problem setting, there are no “fair” baselines to compare IMAGEBIND with. Nevertheless, we compare to prior work that uses text paired with certain modalities (*e.g.* audio [26, 50]), and for certain “visual-like” modalities such as depth and thermal, we use the CLIP model directly. We also report the best reported supervised upper bound per benchmark.

IMAGEBIND achieves a high emergent zero-shot classification performance. On each benchmark, IMAGEBIND achieves strong gains and even compares favorably to supervised specialist models trained for the specific modality and task. These results demonstrate that IMAGEBIND aligns the modalities and implicitly transfers the text supervision associated with images to other modalities like audio. In particular, IMAGEBIND shows strong alignment for non-visual modalities like audio and IMU suggesting that their naturally available pairing with images is a powerful source of supervision. For completeness, we also report the standard zero-shot image (ImageNet [62] - IN1K, Places-365 [85] - P365) and video (Kinetics400 [34] - K400, MSR-VTT 1k-A [76] - MSR-VTT) tasks. As the image & text encoders are initialized (and frozen) using OpenCLIP, these results match those of OpenCLIP.

|               |  |  |  |  |  |  |                        |           |                        |       |       |
|---------------|---|---|---|---|---|---|------------------------|-----------|------------------------|-------|-------|
|               | IN1K  | P365  | K400  | MSR-VTT   | NYU-D   | SUN-D   | AS-A                   | VGGS      | ESC                    | LLVIP | Ego4D |
| Random        | 0.1   | 0.27  | 0.25  | 0.1   | 10.0  | 5.26  | 0.62                   | 0.32      | 2.75                   | 50.0  | 0.9   |
| IMAGEBIND     | 77.7  | 45.4  | 50.0  | 36.1  | 54.0  | 35.1  | 17.6                   | 27.8      | 66.9                   | 63.4  | 25.0  |
| Text Paired   | -   | -   | -   | -   | 41.9*   | 25.4*   | 28.4 <sup>†</sup> [26] | -         | 68.6 <sup>†</sup> [26] | -     | -     |
| Absolute SOTA | 91.0 [80]   | 60.7 [65]   | 89.9 [78]   | 57.7 [77]   | 76.7 [20]   | 64.9 [20]   | 49.6 [38]              | 52.5 [35] | 97.0 [9]               | -     | -     |

**Table 2. Emergent zero-shot classification** of IMAGEBIND using text prompts highlighted in blue. IMAGEBIND aligns images with text, depth, audio, thermal and IMU modalities. The resulting embedding space can associate text embeddings with the non-image modalities, and leads to strong emergent zero-shot classification. We show strong performance even on non-visual modalities such as audio and IMU. We compare to ‘Text Paired’ baselines wherever possible, which trains with paired text data for that modality. \*We use the OpenCLIP ViT-H [29] on depth rendered as grayscale images. <sup>†</sup>[26] that uses AS class names as supervision during training, and hence is not “zero-shot”. Overall, IMAGEBIND shows strong emergent zero-shot performance, even compared to such upper bounds. We also report the absolute state-of-the-art (SOTA) on each dataset for reference, which typically uses additional supervision, model ensembles *etc.* We report the top-1 classification accuracy for all datasets except MSR-VTT (Recall@1) and Audioset Audio-only (mAP).

|  | Emergent | Clotho     |             | AudioCaps  |             | ESC         |
|--|----------|------------|-------------|------------|-------------|-------------|
|  |          | R@1        | R@10        | R@1        | R@10        | Top-1       |
| <i>Uses audio and text supervision</i> |          |            |             |            |             |             |
| AudioCLIP [26]                         | ✗        | -          | -           | -          | -           | <b>68.6</b> |
| <i>Uses audio and text loss</i>        |          |            |             |            |             |             |
| AVFIC [50]                             | ✗        | 3.0        | 17.5        | 8.7        | 37.7        | -           |
| <i>No audio and text supervision</i>   |          |            |             |            |             |             |
| IMAGEBIND                              | ✓        | <b>6.0</b> | <b>28.4</b> | <b>9.3</b> | <b>42.3</b> | 66.9        |
| <i>Supervised</i>                      |          |            |             |            |             |             |
| AVFIC finetuned [50]                   | ✗        | 8.4        | 38.6        | -          | -           | -           |
| ARNLQ [52]                             | ✗        | 12.6       | 45.4        | 24.3       | 72.1        | -           |

**Table 3. Emergent zero-shot audio retrieval and classification.** We compare IMAGEBIND to prior work on zero-shot audio retrieval and audio classification. Without using audio-specific supervision, IMAGEBIND outperforms prior methods on zero-shot retrieval and has comparable performance on the classification task. IMAGEBIND’s emergent zero-shot performance approaches those of specialist supervised models.

## 4.2. Comparison to prior work

We now compare IMAGEBIND against prior work in zero-shot retrieval and classification tasks.

**Zero-shot text to audio retrieval and classification.** Unlike IMAGEBIND, prior work trains using paired data for that modality, *e.g.*, AudioCLIP [26] uses (audio, text) supervision and AVFIC [51] uses automatically mined (audio, text) pairs. We compare their zero-shot text to audio retrieval and classification performance to IMAGEBIND’s emergent retrieval and classification in Table 3.

IMAGEBIND significantly outperforms prior work on the audio text retrieval benchmarks. On the Clotho dataset, IMAGEBIND has double the performance of AVFIC despite not using any text pairing for audio during training. Compared to the supervised AudioCLIP model, IMAGEBIND achieves comparable audio classification performance on ESC. Note that AudioCLIP uses class names from AudioSet as text targets for audio-text training, hence is referred to as ‘su-

|                 | Modality | Emergent | MSR-VTT    |             |             |
|-----------------|----------|----------|------------|-------------|-------------|
|                 |          |          | R@1        | R@5         | R@10        |
| MIL-NCE [48]    | V        | ✗        | 8.6        | 16.9        | 25.8        |
| SupportSet [56] | V        | ✗        | 10.4       | 22.2        | 30.0        |
| FIT [5]         | V        | ✗        | 15.4       | 33.6        | 44.1        |
| AVFIC [50]      | A+V      | ✗        | 19.4       | 39.5        | 50.3        |
| IMAGEBIND       | A        | ✓        | <b>6.8</b> | <b>18.5</b> | <b>27.2</b> |
| IMAGEBIND       | A+V      | ✗        | 36.8       | 61.8        | 70.0        |

**Table 4. Zero-shot text based retrieval** on MSR-VTT 1K-A. We compare IMAGEBIND’s emergent retrieval performance using audio and observe that it performs favorably to methods that use the stronger video modality for retrieval.

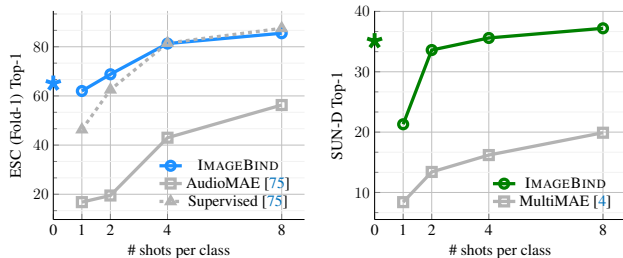
pervised’. IMAGEBIND’s strong performance on all three benchmarks validates its ability to align the audio and text modalities using images as a bridge.

**Text to audio and video retrieval.** We use the MSR-VTT 1k-A benchmark to evaluate the text to audio and video retrieval performance in Table 4. Only using audio, IMAGEBIND achieves strong emergent retrieval performance compared to the video retrieval performance of prior work like MIL-NCE. The text to video performance for our model is strong (36.1% R@1 in Table 2) as it uses OpenCLIP’s vision and text encoders and outperforms many prior methods. However, combining the audio and video modalities further boosts performance showing the utility of IMAGEBIND’s features over an already strong retrieval model.

## 4.3. Few-shot classification

We now evaluate the label-efficiency of IMAGEBIND by evaluating on few-shot classification. We use the audio and depth encoders from IMAGEBIND and evaluate them on audio and depth classification respectively in Figure 3. For  $\geq 1$ -shot results, we follow [49, 59] and train linear classifiers on fixed features (details in Appendix B).

On few-shot audio classification (Figure 3 left), we compare with (1) self-supervised AudioMAE model trained



**Figure 3. Few-shot classification on audio and depth.** We report the emergent zero-shot classification performance on each benchmark (denoted by  $\star$ ). We train linear classifiers on fixed features for the  $\geq 1$ -shot case. **(Left)** In all settings, IMAGEBIND outperforms the self-supervised AudioMAE model. IMAGEBIND even outperforms a supervised AudioMAE model upto 4 shot learning showing its strong generalization. **(Right)** We compare with the MultiMAE model trained with images, depth, and semantic segmentation masks. IMAGEBIND outperforms MultiMAE across all few-shot settings on few-shot depth classification.

on audio from Audioset and (2) a supervised AudioMAE model finetuned on audio classification. Both baselines use the same capacity ViT-B audio encoder as IMAGEBIND. IMAGEBIND significantly outperforms the AudioMAE model on all settings with gains of  $\sim 40\%$  accuracy in top-1 accuracy on  $\leq 4$ -shot classification. IMAGEBIND also matches or outperforms the supervised model on  $\geq 1$ -shot classification. IMAGEBIND’s emergent zero-shot performance surpasses the supervised  $\leq 2$ -shot performance.

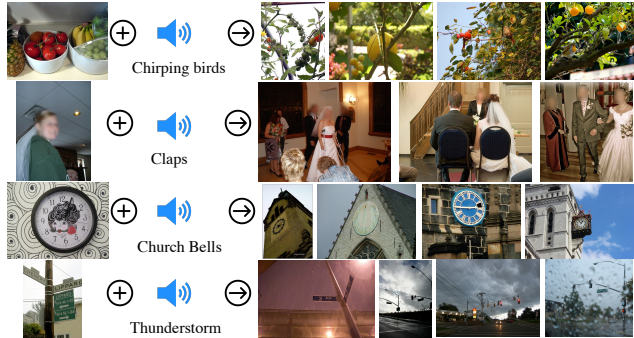
For few-shot depth classification, we compare with the multimodal MultiMAE [4] ViT-B/16 model trained on images, depth, and semantic segmentation data. IMAGEBIND significantly outperforms MultiMAE across all the few-shot settings. Altogether, these results show the strong generalization of IMAGEBIND audio and depth features trained with image alignment.

#### 4.4. Analysis and Applications

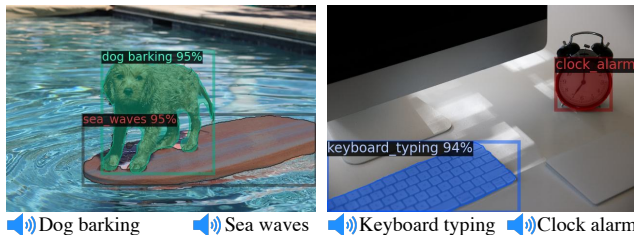
**Multimodal embedding space arithmetic.** We study whether IMAGEBIND’s embeddings can be used to compose information across modalities. In Figure 4, we show image retrievals obtained by adding together image and audio embeddings. The joint embedding space allows for us to compose two embeddings: *e.g.*, image of fruits on a table + sound of chirping birds and retrieve an image that contains both these concepts, *i.e.*, fruits on trees with birds. Such *emergent compositionality* whereby semantic content from different modalities can be composed will likely enable a rich variety of compositional tasks.

Without re-training, we can ‘upgrade’ existing vision models that use CLIP embeddings to use IMAGEBIND embeddings from other modalities such as audio.

**Upgrading text-based detectors to audio-based.** We use a



**Figure 4. Embedding space arithmetic** where we add image and audio embeddings, and use them for image retrieval. The composed embeddings naturally capture semantics from different modalities. Embeddings from an image of fruits + the sound of birds retrieves images of birds surrounded by fruits.



**Figure 5. Object detection with audio queries.** Simply replacing Detic [86]’s CLIP-based ‘class’ embeddings with our audio embeddings leads to an object detector promptable with audio. This requires no re-training of any model.

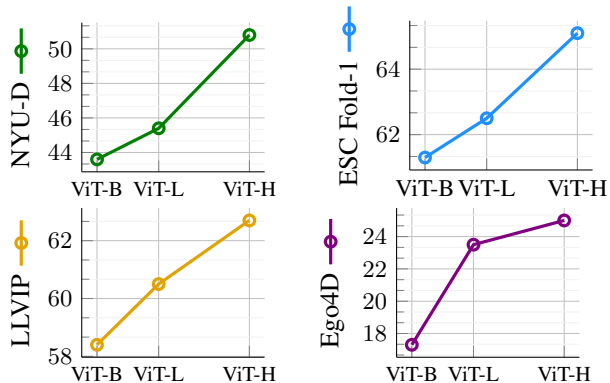
pretrained text-based detection model, Detic [86], and simply replace its CLIP-based ‘class’ (text) embeddings with IMAGEBIND’s audio embeddings. Without training, this creates an ‘audio’-based detector that can detect and segment objects based on audio prompts. As shown in Figure 5, we can prompt the detector with the barking sound of a dog to localize a dog.

#### Upgrading text-based diffusion models to audio-based.

We use a pretrained DALLE-2 [60] diffusion model (private reimplementation) and replace its prompt embeddings by our audio embeddings. In Figure 1, we observe that we can repurpose the diffusion model to generate plausible images using different types of sounds.

### 5. Ablation Study

We investigate various design choices for learning a joint embedding space for different modalities. Since the ablation experimental setup is similar to § 4, we only note the main differences (full details in Appendix C). We report results on the ESC fold-1 for the ablation study. We use a ViT-B encoder for the image, audio, depth, and thermal modalities by default and train them for 16 epochs (*vs.* 32 epochs



**Figure 6. Scaling the image encoder** size while keeping the other modality encoders’ size fixed. We measure the performance on the emergent zero-shot classification of depth, audio, thermal, and IMU modalities. Scaling the image encoder significantly improves the zero-shot classification results suggesting that a stronger visual representation improves the ‘binding’ of modalities.

in § 4). For IMU we use a lightweight 6 layer encoder with 512 dimensional width and 8 heads, and train it for 8 epochs. The text encoder follows [59] and is a twelve layer Transformer with a width of 512 dimensions. We initialize the image and text encoder from the CLIP model [59].

### 5.1. Scaling the Image Encoder

The central idea in IMAGEBIND is aligning the embeddings of all modalities to image embeddings. Thus, the image embeddings plays a central role in the emergent alignment of unseen modalities and we study their effect on the emergent zero-shot performance. We vary the size of the image encoder and train an encoder for the depth, audio *etc.* modalities to match the image representation. To isolate the effect of the image representation, we fix the size of the other modality encoders. We use the pretrained CLIP (ViT-B and ViT-L) and OpenCLIP (ViT-H) image and text encoders for this experiment. Our results in Figure 6 show that IMAGEBIND’s emergent zero-shot performance on all modalities improves with better visual features. For depth and audio classification, the stronger ViT-H *vs.* the ViT-B image encoder, provides a gain of 7% and 4% respectively. Thus, stronger visual features can improve recognition performance even on non-visual modalities.

### 5.2. Training Loss and Architecture

We study the effect of the training design choices on the emergent zero-shot classification. We focus on two modalities with different characteristics - depth which is visual and spatial, and audio which is non-visual and has a temporal component. We found that studying these diverse modalities led to robust and transferable design decisions.

**Contrastive loss temperature.** We study the effect of the

temperature  $\tau$  (Eq 1) in Table 5a. We experiment with a learnable temperature initialized to 0.07 (parametrized in the log-scale) following [59] *vs.* various values of fixed temperatures. Unlike [59], we observe that a fixed temperature is best for depth, audio and IMU classification. Additionally, we see that a higher temperature is better for training the depth, thermal, and IMU encoders, whereas a lower temperature works best for the audio modality.

**Projection head.** We vary the projection head used for each encoder from a linear layer to an MLP with 768 hidden dimensions. The results in Table 5b show that a linear projection performs better for both modalities. This is in contrast to standard self-supervised methods like SimCLR [10] whose performance improves with MLP projection heads.

**Training epochs.** We vary the number training epochs and report the classification performance in Table 5c. Longer training consistently improves the emergent zero-shot performance for both modalities across all datasets.

**Data augmentation for paired images.** During IMAGEBIND training, we augment images either using basic augmentation (cropping, color jitter) or strong augmentation that additionally applies RandAugment [11] and RandErase [84]. We specify the augmentation parameters in Appendix C. Stronger augmentation helps depth classification when training on the small number of (image, depth) pairs from the SUN RGB-D dataset. However, for audio, strongly augmenting the video makes the task too challenging, leading to a significant drop of 34% on ESC.

**Depth specific design choices.** We vary the type of spatial crops used for training in Table 5e. Following CMC [68], we use two unaligned random crops from the corresponding image and depth pair *vs.* our default choice of using spatially aligned random crops. Contrary to CMC, we observe that random cropping severely degrades performance: more than 10% on SUN-D. Unlike vanilla self-supervised learning, our image representations learned from image-text pairs are more semantic and thus spatially misaligned crops hurt performance. In Table 5f, we observe that RandErase [84] boosts performance on depth classification.

**Audio specific design choices.** We train for video-audio alignment using temporally aligned samples or unaligned samples and measure the final performance in Table 5g. Similar to the depth classification observation, temporally aligned samples lead to better performance. Table 5h shows that using frequency masking augmentation for audio also provides a small boost in performance.

**Capacity of the audio and depth encoders** and their impact of the classification performance is reported in Table 6. A smaller encoder for depth improves performance presumably because of the relatively small size of the (image, depth) dataset. Conversely, we observe that larger audio encoder improves the performance, particularly when paired with a high capacity image encoder.

|  |                          |                             |                |   |                |                                |                  |
|--|--------------------------|-----------------------------|----------------|---|----------------|--------------------------------|------------------|
| Temp →                                 | Learn 0.05 0.07 0.2 1.0  | Proj head →                 | Linear MLP     | Epochs →                                | 16 32 64       | Data aug →                     | Basic Strong     |
| SUN-D                                  | 24.1 27.0 27.3 26.7 28.0 | SUN-D                       | 26.7 26.5      | SUN-D                                   | 26.7 27.9 29.9 | SUN-D                          | 25.4 26.7        |
| ESC                                    | 54.8 56.7 52.4 45.4 24.3 | ESC                         | 56.7 51.0      | ESC                                     | 56.7 61.3 62.9 | ESC                            | 56.7 22.6        |
| <b>(a) Temperature for loss.</b>       |                          | <b>(b) Projection Head.</b> |                | <b>(c) Training epochs.</b>             |                | <b>(d) Data aug for image.</b> |                  |
| Spatial align →                        | None Aligned             | Data aug →                  | None RandErase | Temporal align →                        | None Aligned   | Data aug →                     | Basic +Freq mask |
| SUN-D                                  | 16.0 26.7                | SUN-D                       | 24.2 26.7      | ESC                                     | 55.7 56.7      | ESC                            | 56.5 56.7        |
| <b>(e) Spatial alignment of depth.</b> |                          | <b>(f) Depth data aug.</b>  |                | <b>(g) Temporal alignment of audio.</b> |                | <b>(h) Audio data aug.</b>     |                  |

**Table 5. Training loss and architecture** design decisions and their impact on emergent zero-shot classification. Settings for results in § 4 highlighted in gray. **(a)** A fixed temperature in the contrastive loss outperforms a learnable one for all modalities. **(b)** A linear projection head for computing the depth or audio embedding works better than an MLP head. **(c)** Longer training improves the zero-shot classification performance for both modalities. **(d)** Stronger image augmentation improves depth classification while basic augmentation significantly improves audio classification. **(e, f)** Using spatially aligned image and depth crops when training IMAGEBIND significantly improves performance. Similarly, RandErase augmentation is critical to good zero-shot classification on depth. **(g, h)** Temporally aligned audio and video matching gives improved performance and using frequency augmentation for audio gives a slight improvement.

| Image Encoder | Audio Encoder (ESC) |             | Depth Encoder (SUN) |       |
|---------------|---------------------|-------------|---------------------|-------|
|               | ViT-S               | ViT-B       | ViT-S               | ViT-B |
| ViT-B         | 52.8                | 56.7        | 30.7                | 26.7  |
| ViT-H         | 54.8                | <b>60.3</b> | <b>33.3</b>         | 29.5  |

**Table 6. Capacity of the audio and depth encoders** and their impact on performance. A stronger image encoder improves performance for both audio and depth tasks. As the number of (image, depth) pairs is small, a smaller encoder improves performance for depth. For audio classification, a larger encoder is better.

| Batch size → | 512  | 1k   | 2k   | 4k   |
|--------------|------|------|------|------|
| NYU-D        | 47.3 | 46.5 | 43.0 | 39.9 |
| ESC          | 39.4 | 53.9 | 56.7 | 53.9 |

**Table 7. Effect of scaling batch size.** We found the optimal batch size for contrastive loss varied by the modality. For image-depth task, a smaller batch size was better, likely due to the small size and limited diversity of the original dataset. For audio-video task where we have a lot more positive and negative audio-video pairs, using a large batch size lead to better results.

**Effect of batch size.** In Table 7 we evaluate the effect of batch size on the representation learned. As shown, the batch size can vary across modalities depending on the size and complexity of the corresponding pretraining datasets.

**IMAGEBIND to evaluate pretrained vision models** in Table 8. We initialize the vision encoder using a pretrained model and keep it fixed. We use image-paired data to align and train text, audio, and depth encoders (full details in Appendix B). Compared to the supervised DeiT model, the self-supervised DINO model is better at emergent zero-shot classification on both depth and audio modalities. Moreover, the emergent zero-shot performance is not correlated with the pure vision performance on ImageNet suggesting that these tasks measure different properties. IMAGEBIND can serve as a valuable tool to measure vision models’ strength on multimodal applications.

|           | IN1K              | VGG5 ESC  | SUN-D | NYU-D |
|-----------|-------------------|-----------|-------|-------|
| DINO [6]  | 64.4              | 17.2 44.7 | 26.8  | 48.8  |
| DeiT [70] | 74.4 <sup>†</sup> | 9.6 25.0  | 25.2  | 48.0  |

**Table 8. IMAGEBIND as an evaluation tool.** We initialize (and fix) the image encoder with different methods and align other modalities. IMAGEBIND measures the impact of visual features on multimodal tasks. <sup>†</sup> trained with IN1K supervision.

## 6. Discussion and Limitations

IMAGEBIND is a simple and practical way to train a joint embedding space using only image alignment. Our method leads to emergent alignment across all modalities which can be measured using cross-modal retrieval and text-based zero-shot tasks. We enable a rich set of compositional multimodal tasks across different modalities, show a way to evaluate pretrained vision models for non-vision tasks and ‘upgrade’ models like Detic and DALLE-2 to use using audio. There are multiple ways to further improve IMAGEBIND. Our image alignment loss can be enriched by using other alignment data, for instance other modalities paired with text, or with each other (*e.g.* audio with IMU). Our embeddings are trained without a specific downstream task, and thus lag the performance of specialist models. More research into adapting general purpose embeddings for each task, including structured prediction tasks such as detection will be beneficial. Finally, new benchmarks, *e.g.* our emergent zero-shot task to measure emergent abilities of multimodal models, would help create exciting new applications. Our model is a research prototype and cannot be readily used for real world applications ( Appendix F).

**Acknowledgements:** Authors would like to thank Uriel Singer, Adam Polyak and Naman Goyal for their help with the DALLE-2 experiments, and the entire Meta AI team for many helpful discussions.



## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1, 2
- [2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2020. 2
- [3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. 1, 2
- [4] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. *arXiv preprint arXiv:2204.01678*, 2022. 1, 6
- [5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *arXiv preprint arXiv:2104.00650*, 2021. 5
- [6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 8
- [7] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 3
- [8] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 4, 12
- [9] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP*, 2022. 5
- [10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 7, 14
- [11] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR*, 2020. 7
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3
- [13] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 2
- [14] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 2
- [15] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022. 13
- [16] Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *ACM international conference on Multimedia*, 2013. 4, 12
- [17] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 2
- [18] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 4, 12
- [19] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. *arXiv preprint arXiv:2206.08356*, 2022. 2, 3
- [20] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A Single Model for Many Visual Modalities. In *CVPR*, 2022. 2, 4, 5, 12
- [21] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Interspeech*, 2021. 3, 4, 13
- [22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 4, 12
- [23] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2
- [24] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *CVPR*, 2013. 13
- [25] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, 2014. 13
- [26] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021. 2, 3, 4, 5
- [27] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 3
- [28] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 14
- [29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 4, 5
- [30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 2
- [31] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *ICCV*, 2021. 4, 12, 13
- [32] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. 2
- [33] Armand Joulin, Laurens van der Maaten, Allan Jabri, and

- Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016. 2
- [34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, AMustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 4
- [35] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP*, 2021. 5
- [36] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL*, 2019. 4, 12
- [37] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2
- [38] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Interspeech*, 2022. 5
- [39] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017. 2
- [40] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranfl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 2
- [41] Valerii Likhoshesterov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv preprint arXiv:2111.12993*, 2021. 2
- [42] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *ECCV*, 2022. 2
- [43] Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. *arXiv preprint arXiv:2209.03917*, 2022. 2
- [44] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. corr abs/2104.08860 (2021). *arXiv preprint arXiv:2104.08860*, 2021. 2
- [45] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 1
- [46] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 2
- [47] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *ICCV*, 2019. 2
- [48] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 5
- [49] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *CVPR*, 2021. 1, 2, 3, 5
- [50] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *ECCV*, 2022. 2, 4, 5
- [51] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, 2021. 5
- [52] Andreea-Maria Oncescu, A Koepke, Joao F Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries. *arXiv preprint arXiv:2105.02192*, 2021. 5, 12
- [53] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *NeurIPS*, 2018. 3
- [54] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 1
- [55] Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. In *ICCV*, 2021. 1
- [56] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metz, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 5
- [57] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2
- [58] Karol J Piczak. Esc: Dataset for environmental sound classification. In *ACM MM*, 2015. 4, 12
- [59] 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 *ICML*, 2021. 1, 2, 3, 4, 5, 7, 13, 14, 15
- [60] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 6, 13
- [61] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 13
- [62] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 4
- [63] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1
- [64] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012. 4, 12, 13

- [65] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *CVPR*, 2022. 5
- [66] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2014. 2
- [67] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 4, 12
- [68] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 1, 2, 3, 7
- [69] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 13
- [70] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 8
- [71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3
- [72] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. BEVT: Bert pretraining of video transformers. In *CVPR*, 2022. 2
- [73] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 2
- [74] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3
- [75] Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *NeurIPS*, 2022. 6
- [76] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 4
- [77] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 2, 5
- [78] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, 2022. 5
- [79] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *ECCV*, 2022. 2
- [80] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1, 2, 5
- [81] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1, 2
- [82] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 2
- [83] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, 2022. 2, 3
- [84] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 7
- [85] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NeurIPS*, 2014. 4
- [86] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2, 6, 13

## A. Datasets and Metrics

**Audioset (AS)** [18]. This dataset is used for both training and evaluation. It contains 10s videos from YouTube annotated into 527 classes. It consists of 3 pre-defined splits, the balanced split with about 20K videos, test split with 18K videos, and an unbalanced training split with about 2M videos. For **training**, we use the 2M unbalanced set without any labels, and only use it for audio-video matching. For **zero-shot evaluation** in Table 2, we use the test set and compute logits for each class using the textual class names along with the templates as described later in Appendix B.3. The metric used is top-1 accuracy.

**ESC-50 (ESC)** [58]. We use this dataset for evaluating the learned representations in a zero-shot manner. The task here is “Environmental Sound Classification” (ESC). It consists of 2000 5s audio clips classified into 50 classes. It has pre-defined 5 fold evaluation, each consisting of 400 test audio clips. In this work, we compute 0-shot predictions on the evaluation set for each fold and report the 5-fold average performance. For ablations we use only the first fold for computational ease. The metric used is top-1 accuracy.

**Clotho (Clotho)** [16]. This is a dataset of audio from the Freesound platform with textual descriptions. It consists of a dev and test set of 2893 and 1045 audio clips respectively, with each clip associated with 5 descriptions. We consider the text→audio retrieval task, and consider each of the 5 associated captions as a separate test query and retrieve from the set of audio clips. The metric used is  $\text{recall}@K$ , where a given test query is assumed to be correctly solved if the ground truth audio is retrieved within the top- $K$  retrieved audio clips.

**AudioCaps (AudioCaps)** [36]. This is a dataset of audio-visual clips from YouTube accompanied by textual descriptions. It consists of clips from the Audioset dataset as described earlier. We use the splits as provided in [52],<sup>1</sup> which removes clips that overlap with the VGGSound dataset. We end up with 48198 training, 418 validation and 796 test clips. We only use the test set for zero-shot evaluation of our model. The task is text→audio retrieval, and evaluation is performed using  $\text{recall}@K$ .

**VGGSound (VGS)** [8]. This dataset contains about 200K video clips of 10s length, annotated with 309 sound classes consisting of human actions, sound-emitting objects and human-object interactions. We only use the audio from the test set (with 14073 clips) for 0-shot classification. The evaluation is done using the top-1 accuracy metric.

**SUN RGB-D (SUN)**. We use the registered RGB and Depth maps provided in the SUN RGB-D [67] dataset train set (~5K pairs) for training our model. We follow [20] to post process the depth maps in two steps - 1) we use in-filled

depth values and 2) convert them to disparity for scale normalization. This dataset is only used in training, so we do not use any metadata or class labels.

**SUN Depth-only (SUN-D)**. We use only the ~5K depth maps from the val split of the SUN RGB-D [67] dataset and denote them as SUN Depth-only. This dataset is only used for evaluation and we do not use the RGB images. We process the depth maps similar to SUN RGB-D (in-filled depth, converted to disparity). We use the 19 scene classes in the dataset and use their class names for constructing the zero-shot classification templates.

**NYU-v2 Depth-only (NYU-D)**. We use the 794 val set depth maps from the NYU-v2 Depth-only [64] dataset for evaluation only. We post-process the depth similar to SUN Depth-only. We use the 10 scene class names in the dataset. The 10th scene class, called ‘other’, correspond to 18 different semantic classes – [‘basement’, ‘cafe’, ‘computer lab’, ‘conference room’, ‘dINETTE’, ‘exercise room’, ‘foyer’, ‘furniture store’, ‘home storage’, ‘indoor balcony’, ‘laundry room’, ‘office kitchen’, ‘playroom’, ‘printer room’, ‘reception room’, ‘student lounge’, ‘study’, ‘study room’]. For zero-shot evaluation, we compute the cosine similarity of the 10th class as the maximum cosine similarity among these 18 classnames.

**LLVIP (LLVIP)**. The LLVIP dataset [31] consists of RGB image and Thermal (infrared low-light) image pairs. The dataset was collected in an outdoor setting using fixed cameras observing street scenes and contains RGB images taken in a low-light paired with infrared images (8~14um frequency). The RGB thermal pairs are registered in the dataset release. For training, we use the train set with 12025 RGB image and thermal pairs. For evaluation, we use the val set with 3463 pairs of RGB and thermal images. Since the original dataset is designed for detection, we post process it for a binary classification task. We crop out pedestrian bounding boxes and random bounding boxes (same aspect ratio and size as pedestrian) to create a balanced set of 15809 total boxes (7931 ‘person’ boxes). For zero-shot classification, we use the following class names for the ‘person’ class - [‘person’, ‘man’, ‘woman’, ‘people’], and [‘street’, ‘road’, ‘car’, ‘light’, ‘tree’] for the background class.

**Ego4D (Ego4D)** [22]. For the Ego4D dataset, we consider the task of scenario classification. There are 108 unique scenarios present in the 9,645 videos of the Ego4D dataset. We filter out all videos annotated with more than one scenario which yields 7,485 videos with a single scenario assigned. For each video, We select all time-stamps that contains a synchronized IMU signal as well as aligned narrations. We sample 5 second clips around each time-stamp. The dataset is split randomly such that we have 510,142 clips for train-

<sup>1</sup>[https://www.robots.ox.ac.uk/~vgg/research/audio-retrieval/resources/benchmark-files/AudioCaps\\_retrieval\\_dataset.tar.gz](https://www.robots.ox.ac.uk/~vgg/research/audio-retrieval/resources/benchmark-files/AudioCaps_retrieval_dataset.tar.gz)

ing, and 68,865 clips for testing. During training we only use the video frames and their corresponding IMU signal. We use the test split to measure zero-shot scenario classification performance, where each clip of IMU signal is assigned the video-level scenario label as its ground-truth.

### A.1. Data Representations

We use the standard RGB and RGBT representations for **images and videos**. For videos, we use 2-frame clips, inspired from recent work on ViT-style video architectures [15, 69], where a video patch is  $2 \times 16 \times 16 (T \times H \times W)$ . We inflate the visual encoder’s weights to work with spatiotemporal patches and at inference time we aggregate features over multiple 2-frame clips. Hence, we can use models trained on image-text data directly on videos.

We used a single-channel image for the **thermal data** since it is the natural form in which current infrared thermal sensors return data [31]. For **single-view depth**, we experimented with different encodings – absolute depth [64] as returned by sensors like the Kinect, inverse depth [61], disparity [61], and HHA [24, 25]. Overall, we found that disparity representation (which is a single-channel image) worked the best. For **audio** we use the raw waveform processed into mel-spectrograms [21], as described in the main text. For **IMU** we use a  $6 \times T$  tensor to represent the sequence of IMU sensor readings over time.

## B. Evaluation details

We now describe the evaluation setups used in this work.

### B.1. Inference implementation details

**Audio/Video:** For both these temporal modalities (whether operated upon together during pre-training or separately during inference), we sample fixed length clips to operate on. During training, we randomly sample a clip, typically 2s in length. At inference time, we uniformly sample multiple clips to cover the full length of the input sample. For instance, for 5s ESC videos, we would sample  $\lceil \frac{5}{2} \rceil = 3$  clips. For video clips, we sample a fixed number of frames from each clip. For audio, we process each raw audio waveform by sampling it at 16KHz followed by extracting a log mel spectrogram with 128 frequency bins using a 25ms Hamming window with hop length of 10ms. Hence, for a  $t$  second audio we get a  $128 \times 100t$  dimensional input.

**IMU:** For IMU, we sample fixed length clips of 5 seconds, centered around time-stamps that are aligned with narrations. For each clip, we get a  $6 \times 2000$  dimensional input and we measure the zero-shot performance for scenario classification using each clip as an independent testing sample.

### B.2. Few-shot evaluation details

For the few-shot results in Figures 3 using the ESC and SUN datasets, we sampled  $k$  training samples per class,

where  $k \in \{1, 2, 4, 8\}$ . We fix the  $k$  samples such that our model and the baselines use exactly the same samples during training. For all few-shot evaluations, including the baselines, we freeze the encoder parameters and only train a linear classifier.

**Audio:** For audio few-shot training with ESC, our model and the baselines are trained using AdamW with a learning rate of  $1.6 \times 10^{-3}$  and weight decay of 0.05 for 50 epochs.

**Depth:** For depth few-shot training with SUN, our model and the baselines are trained using AdamW with a learning rate of  $10^{-2}$  and no weight decay for 60 epochs.

### B.3. Zero-shot evaluation details

**Query Templates.** For all evaluations, we use the default set of templates from CLIP [59].<sup>2</sup> Note that we use the same templates for non visual modalities like audio and depth as well since we only use semantic/textual supervision associated with images.

### B.4. Qualitative evaluation details

**Cross-modal nearest neighbors.** We perform the retrieval on the embedding feature after temperature scaling. The nearest neighbors are computed using cosine distance. In Figure 1, we show retrievals for audio from ESC, image retrievals from IN1K and COCO, depth from SUN-D, and text from AudioCaps.

**Embedding arithmetic.** For arithmetic, we again use the embedding features after temperature scaling. We  $\ell_2$  normalize the features and sum the embeddings after scaling them by 0.5. We use the combined feature to perform nearest neighbor retrieval using cosine distance, as described above. In Figure 1, we show combination of images and audio from IN1K and ESC, and show retrievals from IN1K.

**Audio→Image Generation.** For generating images from audio clips, we rely on an in-house reproduced implementation of DALLE-2 [60]. In DALLE-2, to produce images from text prompts, the image generation model relies on text embeddings produced by the pre-trained CLIP-L/14 text encoder. Since IMAGEBIND naturally aligns CLIP’s-embedding space to that of other modalities proposed in the paper, we can upgrade the DALLE-2 model to generate images by prompting it with these new unseen modalities. We achieve zero-shot audio to image generation with DALLE-2 by simply using the temperature-scaled audio embeddings generated by IMAGEBIND’s audio encoder as a proxy for the CLIP’s text embeddings in the DALLE-2’s image generation model.

**Detecting objects using audio.** We extract all audio descriptors from the validation set of ESC using an IMAGEBIND ViT-B/32 encoder, yielding 400 descriptors in total. We use an off-the-shelf CLIP-based Detic [86] model and

<sup>2</sup>[https://github.com/openai/CLIP/blob/main/notebooks/Prompt\\_Engineering\\_for\\_ImageNet.ipynb](https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb)

use the audio descriptors as the classifier for Detic in place of CLIP text-based ‘class’ embeddings. We use a score threshold of 0.9 for the qualitative results in Figure 5.

## C. Pretraining details

### C.1. Best setup

In Table 9 we detail the hyperparameters used to pre-train each of the models reported in Table 4. Our experiments were done on 32GB V100 or 40GB A100 GPUs.

| Config                 | AS                              | SUN        | LLVIP      | Ego4D |
|------------------------|---------------------------------|------------|------------|-------|
| Vision encoder         | ViT-Huge                        |            |            |       |
| embedding dim.         | 768                             | 384        | 768        | 512   |
| number of heads        | 12                              | 8          | 12         | 8     |
| number of layers       | 12                              | 12         | 12         | 6     |
| Optimizer              | AdamW                           |            |            |       |
| Optimizer Momentum     | $\beta_1 = 0.9, \beta_2 = 0.95$ |            |            |       |
| Peak learning rate     | 1.6e-3                          | 1.6e-3     | 5e-4       | 5e-4  |
| Weight decay           | 0.2                             | 0.2        | 0.05       | 0.5   |
| Batch size             | 2048                            | 512        | 512        | 512   |
| Gradient clipping      | 1.0                             | 1.0        | 5.0        | 1.0   |
| Warmup epochs          | 2                               |            |            |       |
| Sample replication     | 1.25                            | 50         | 25         | 1.0   |
| Total epochs           | 64                              | 64         | 64         | 8     |
| Stoch. Depth [28]      | 0.1                             | 0.0        | 0.0        | 0.7   |
| Temperature            | 0.05                            | 0.2        | 0.1        | 0.2   |
| Augmentations:         |                                 |            |            |       |
| RandomResizedCrop size | -                               | 224px      |            | -     |
| interpolation          | -                               | Bilinear   | Bilinear   | -     |
| RandomHorizontalFlip   | -                               | $p = 0.5$  | $p = 0.5$  | -     |
| RandomEraser           | -                               | $p = 0.25$ | $p = 0.25$ | -     |
| RandAugment            | -                               | 9/0.5      | 9/0.5      | -     |
| Color Jitter           | -                               | 0.4        | 0.4        | -     |
| Frequency masking      | 12                              | -          | -          | -     |

Table 9. Pretraining hyperparameters

**Contrastive loss batch size vs. modalities.** While contrastive losses do require larger batch size, this requirement didn’t increase with the number of modalities. As noted in Appendix B, our experiments (Table 2) sample a mini-batch of one pair of modalities at a time: batch size of 2K for (video, audio), and 512 for (image, depth), (image, thermal), and (video, IMU). These batch sizes are smaller than the >32K batch sizes used in prior work [10, 59].

**Combining modalities.** In Table 4, we show results with combining the audio and video modalities. We combine them by extracting embeddings from both modalities per sample and computing a linear combinations of those embeddings. We used a weight of 0.95 for video and 0.05 for audio for this combination, which was found to perform the best.



Figure 7. IMU retrievals. Given a text query, we show some IMU retrievals and corresponding video frames.

### C.2. Ablation setup

The following setup was used for our evaluations in § 5. Different from the best setup, all ablation experiments uses ViT-Base both for the vision and the modality-specific encoders. The models are trained for 16 epochs, unless mentioned otherwise.

For Table 5b, the differences between the linear and MLP heads are detailed below: The MLP head did not improve performance in our experiments.

|        |   |
|--------|---|
| Linear | Linear(in.dim, out.dim)                               |
| MLP    | Linear(in.dim, in.dim), GELU, Linear(in.dim, out.dim) |

## D. Additional Results

**Qualitative results.** We show additional results (along with audio) in the accompanying video.

**Practical applications of disparate modalities.** In general, a shared embedding space enables a variety of different cross-modal search and retrieval applications. *e.g.*, since IMU sensors are ubiquitous (in phones, AR/VR headsets, health trackers), IMAGEBIND can allow a user to search an IMU database using text queries (without training with IMU-text pairs). IMU-based text search has applications in healthcare/activity search. For instance, in Figure 7 we show examples of IMU (and accompanying video) retrieval given textual search query. The retrieved IMU sample, shown as 3-channel Accelerometer (Acc) and Gyroscope (Gyro) recording, matches the text query.

## E. Additional Ablations

**Design choices in losses.** Since the modality-specific encoders are trained to align with a frozen image encoder, we tried using a  $\ell_2$  regression objective. For ZS SUN top-1 accuracy, we observed that regression led to good performance as the sole objective (25.17%) or jointly with contrastive (29.04%). However, it did not improve over using only the contrastive objective (31.74%).

## F. Ethical considerations

IMAGEBIND learns a joint embedding for multiple modalities. Such an embedding is intended to associate semantically related concepts from different modalities. However, such an embedding may also create unintentional associations. Thus, joint embedding models, including IMAGEBIND must be studied carefully with a lens towards measuring such associations, and their implications. IMAGEBIND leverages the image-text embeddings learned by a pretrained model on large web-based data which has biases as documented in different studies [59]. For learning joint embeddings for other modalities such as audio, thermal, depth, and IMU we leverage datasets mentioned in Appendix A. These joint embeddings are thus limited to the concepts present in the datasets. For example, the thermal datasets we used are limited to outdoor street scenes, while the depth datasets are limited to indoor scenes.