Connecting the Dots between Audio and Text without Parallel Data through Visual Knowledge Transfer

Anonymous ACL submission

Abstract

Machines that can represent and describe environmental soundscapes have practical potential, e.g., for audio tagging and captioning. Prevailing learning paradigms of audio-text connections have been relying on parallel audio-text data, which is, however, scarcely available on the web. We propose VIP~ANT that induces **A**udio-**T**ext alignment without using any parallel audio-text data. Our key idea is to share the image modality between bi-modal image-text representations and bi-modal image-audio representations; the image modality functions as a pivot and connects audio and text in a tri-modal embedding space implicitly.

In a difficult zero-shot setting with no paired audio-text data, our model demonstrates state-of-the-art zero-shot performance on the ESC50 and US8K audio classification tasks, and even surpasses the supervised state of the art for Clotho caption retrieval (with audio queries) by 2.2% R@1. We further investigate cases of minimal audio-text supervision, finding that, e.g., just a few hundred supervised audio-text pairs increase the zero-shot audio classification accuracy by 8% on US8K. However, to match human parity on some zero-shot tasks, our empirical scaling experiments suggest that we would need about $2^{21} \approx$ 2M supervised audio-caption pairs. Our work opens up new avenues for learning audio-text connections with little to no parallel audio-text data.

1 Introduction

Environmental sound provides rich perspectives on the physical world. For example, if we hear: *joyful laughing, a playful scream, and a splash;* we not only can visualize literal objects / actions that might have given rise to the audio scene, but also, we can reason about plausible higher-level facets, e.g., a child speeding down a water slide at a water park, splashing through the water (see Figure 1).

Machines capable of parsing, representing, and describing such environmental sound hold practical promise. For example, according to the National Association of the Deaf's captioning guide, accessible audio caption generation systems should go beyond speech recognition (i.e., identifying speakers and transcribing the literal content of their speech) and provide the textual description of all the sound effects, e.g., "a large group of people talking excitedly at a party" in order to provide the full information contained in that audio.[1]

Figure 1: VIP~ANT pivots audio and text via visual imagination.

The dominant paradigm for studying *machine hearing* (Lyon, 2010) has been through human-annotated audio-text data, where text is either free-form audio descriptions ("the sound of heavy rain") or tagsets (Salamon et al., 2014; Gemmeke et al., 2017; Kim et al., 2019; Drossos et al., 2020). While naturally aligned audio-text data could be sourced from audio-tag co-occurrences (Font et al., 2013) and from video captioning data (Rohrbach et al., 2015; Xu et al., 2016; Oncescu et al., 2021a), they are either not sufficiently related to environmental sound or limited in their scale and coverage.

In this paper, we study large-scale audio-text alignment without paired audio-text (AT) data. Inspired by pivot-based models for unsupervised machine translation (Wu and Wang, 2007; Utiyama

[1] nad.org's captioning guide; Gernsbacher (2015) discusses the benefits of video captions beyond d/Deaf users.

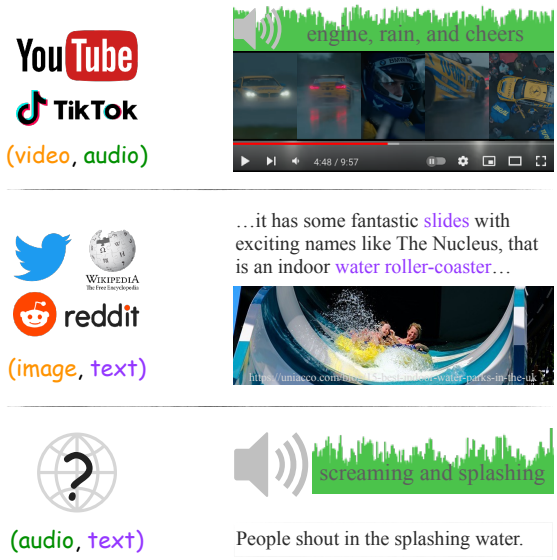


Figure 2: Video-audio and image-text co-occurrences are abundantly available on the web to support the learning of video-audio alignment and image-text alignment (e.g., via large-scale video-audio and image-text pre-training), but audio-text co-occurrences are not.

and Isahara, 2007), we propose VIP~ANT, short for Visually Pivoted Audio and(N) Text. VIP~ANT uses images as a pivot modality to connect audio and text. It parallels our motivating example: hearing a sound, humans can visually *imagine* the associated situation and literally *describe* it. Pivoting is practically viable because there are abundantly available image-text (VT) and video-audio (VA) co-occurrences on the web (see Figure 2). We can use them to supervise an image-text alignment model and an image-audio (VA) alignment model; sharing the image modality between the two alignment models will link audio and text implicitly.

Besides the fully unsupervised *pivoting* model VIP~ANT, we consider improving it with two cases of varying AT supervision. (1) *unsupervised curation:* whereby noisy AT pairs are explicitly mined from the pivoting model and serve as additional training data, and (2) *few-shot curation:* whereby a small number of human-annotated AT pairs are made available at training time.

We quantify the quality of the AT alignments via zero-shot audio-text retrieval and zero-shot audio classification. On the Clotho caption retrieval task (Drossos et al., 2020), without any parallel AT data, VIP~ANT surpasses the supervised state of the art by 2.2% R@1; on zero-shot audio classification tasks, it establishes new state of the arts, achieving 57.1% accuracy on ESC50 (Piczak, 2015) and 44.7% accuracy on US8K (Salamon et al., 2014). We also show that *unsupervised curation*, i.e., mining noisy pairs from the pivoting model, can surprisingly increase performance further (e.g., +5.7% on ESC50 and +9.3% on US8K). Finally, we find that *few-shot curation* with only a few hundred supervised AT pairs during pre-training increases the zero-shot audio classification accuracy by 8% on US8K. However, for ESC-50, according to empirical scaling laws we demonstrate, it would require around $2^{21} \approx$ 2M aligned audio-text pairs for the zero-shot model to match human parity on ESC50 under our setup, which is an order-of-magnitude more than the largest currently-available audio-text corpus of Kim et al. (2019).

2 Related work

Supervised audio representation learning. While automatic speech recognition has been a core focus of the audio processing community, environment sound classification has emerged as a new challenge and is drawing more attention (Salamon et al., 2014; Piczak, 2015; Gemmeke et al., 2017). Some prior work in learning sound event representations are supervised by category labels (Dai et al., 2017; Boddapati et al., 2017; Kumar et al., 2018; Guzhov et al., 2021b; Gong et al., 2021). Others use weaker forms of supervision for tagging (Kumar and Raj, 2017; Kong et al., 2018) and localization (McFee et al., 2018; Kim and Pardo, 2019).

Learning audio representations from visual imagination. There are two main paradigms for using visual information to derive audio representations. In the two-stage setup, an image encoder is first pre-trained; these weights are used as initialization of the supervised audio model (Guzhov et al., 2021b; Gong et al., 2021). The other adopts contrastive learning: it exploits the image-audio alignment inherent in videos and learns audio and image / video representations jointly (Korbar et al., 2018; Wang et al., 2021; Nagrani et al., 2021). We use insights from both directions by (1) using CLIP's image encoder, which has been pre-trained on image-text pairs (Radford et al., 2021), to initialize an audio encoder and (2) using contrastive pre-training on image-audio pairs. Throughout training, we do not require any labeled images or audio.

Tri-modal learning of audio-text alignment. Our work extends recent work that generalizes the

Model	AE Init.	Objective	AT Supervision	VT Alignment	Zero-shot AT Retrieval
MMV (Alayrac et al., 2020)	Random	$\mathcal{L}_{bi\text{-}bi}$	None	Trainable	✗
VATT (Akbari et al., 2021)	Random	$\mathcal{L}_{bi\text{-}bi}$	None	Trainable	✗
AudioCLIP (Guzhov et al., 2021a)	ImageNet	$\mathcal{L}_{tri}$	2M Audio Tags	Trainable	✗
Wav2CLIP (Wu et al., 2021)	Random	$\mathcal{L}_{bi\text{-}bi}$	None	Frozen	✗
VIP~ANT (ours)	Image CLIP	$\mathcal{L}_{bi\text{-}bi}$	None	Frozen	✓
VIP~ANT +AT (ours)	Image CLIP	$\mathcal{L}_{bi\text{-}bi}$	Caption Curation	Frozen	✓

Table 1: Survey of recent prior work studying for tri-modal (images, audio, and text) representation learning. AE is short for **A**udio **E**ncoder. Some work experiments with more than one objective, we report the best or the one it advocates. Importantly, we report zero-shot audio-text retrieval between audio and full-sentence text descriptions, along with scaling laws associated with that setup.

bi-modal contrastive learning to a tri-modal setting (Alayrac et al., 2020; Akbari et al., 2021). While they also connect audio and text implicitly by using images as a pivot, the quality of this audio-text alignment has rarely been studied. To our knowledge, we present the first comprehensive evaluation of the inferred audio-text alignment via zero-shot retrieval / classification.

The work closest to ours are AudioCLIP (Guzhov et al., 2021a) and Wav2CLIP (Wu et al., 2021). AudioCLIP's pre-training setup is similar to ours, but requires human-annotated textual labels of audio, while ours does not. Wav2CLIP is concurrent with our work; while similar-in-spirit, our model not only performs significantly better, but also, we more closely explore methods for improving audio-text alignment specifically.

Pivot-based alignment models. The pivoting idea for alignment learning can date back to Brown et al. (1991). Language pivots (Wu and Wang, 2007; Utiyama and Isahara, 2007) and image pivots (Specia et al., 2016; Hitschler et al., 2016; Nakayama and Nishida, 2017) have been explored in machine translation. Pivot-based models have also been shown to be helpful in learning image-text alignment (Li et al., 2020). We focus on the tri-modal case.

3 Model

We first formalize tri-modal learning by assuming available co-occurrence data for every pair of modalities, and present bi-bi-modal pre-training as an alternative when there is no paired audio-text data (§ 3.1). Then we implement VIP~ANT via bi-bi-modal pre-training (§ 3.2) and describe model variants for cases of varying AT supervision (§ 3.3).

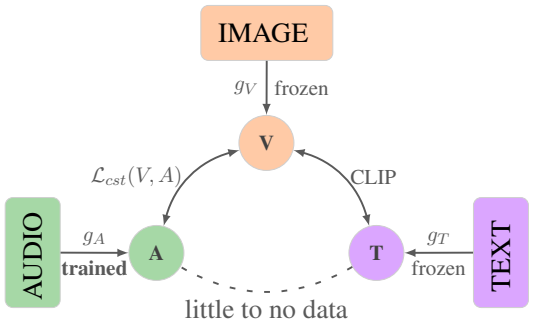


Figure 3: Learning paradigm of VIP~ANT.

3.1 Tri-modal representation learning

Tri-modal representation learning between images, audio, and text aims to derive representations from co-occurrence patterns among the three modalities (Alayrac et al., 2020; Akbari et al., 2021). We consider a simple tri-modal representation space, which relies on encoding functions $g_V : V \rightarrow \mathbf{V}$, $g_A : A \rightarrow \mathbf{A}$, and $g_T : T \rightarrow \mathbf{T}$ to map images v, audios a, and text t ($v \in V, a \in A$, and $t \in T$), respectively, to a shared vector space: $\boldsymbol{v}, \boldsymbol{a}, \boldsymbol{t} \in \mathbb{R}^d$ ($\boldsymbol{v} \in \mathbf{V}, \boldsymbol{a} \in \mathbf{A}$, and $\boldsymbol{t} \in \mathbf{T}$). Instead of pre-specifying the precise semantics of this continuous space, vector similarities across modalities are optimized to reconstruct co-occurrence patterns in training corpora, i.e., two vectors should have a higher dot product if they are more likely to co-occur. We use contrastive learning with the InfoNCE loss (Sohn, 2016; van den Oord et al., 2018):

$$\mathcal{L}_{cst}(A, B) = \sum_i \frac{\exp s(\boldsymbol{a}^{(i)}, \boldsymbol{b}^{(i)})}{\sum_{\boldsymbol{a}} \exp s(\boldsymbol{a}, \boldsymbol{b}^{(i)})} + \frac{\exp s(\boldsymbol{a}^{(i)}, \boldsymbol{b}^{(i)})}{\sum_{\boldsymbol{b}} \exp s(\boldsymbol{a}^{(i)}, \boldsymbol{b})}, \quad (1)$$

where A, B are two sets of data points from two different modal domains, respectively; $\boldsymbol{a}^{(i)}, \boldsymbol{b}^{(i)}$ are vector representations of the co-occuring pair

$(a^{(i)}, b^{(i)})$ which are encoded by $g_A(a^{(i)})$ and $g_B(b^{(i)})$, respectively; $s(\boldsymbol{a}, \boldsymbol{b})$ computes the similarity between $\boldsymbol{a}$ and $\boldsymbol{b}$, which we take to be scaled cosine similarity.

If we had access to co-occurrence data between all pairs of modalities, we could optimize the tri-modal loss:

$$\mathcal{L}_{tri}(V, A, T) = \mathcal{L}_{cst}(V, A) + \mathcal{L}_{cst}(A, T) + \mathcal{L}_{cst}(V, T)\,. \quad (2)$$

But, differently from image-text and image-audio pairs, which are abundantly available on the web, audio-text data is scarce. Instead, tri-modal representation learning minimizes a "bi-bi-modal" loss:

$$\mathcal{L}_{bi\text{-}bi}(V, A, T) = \mathcal{L}_{cst}(V, A) + \mathcal{L}_{cst}(V, T)\,. \quad (3)$$

3.2 Visually pivoted audio and text

We propose VIP~ANT, which aligns audio and text via visual images. Our model capitalizes on the availability of VA and VT pairs. It follows the bi-bi-modal learning paradigm (see Equation 3) to learn a tri-modal representation space. The image encoder is shared between the VA alignment model (i.e., $\mathcal{L}_{cst}(V, A)$) and the VT alignment model (i.e., $\mathcal{L}_{cst}(V, T)$) and thus connects audio and text in the tri-modal embedding space implicitly.

Image and text encoders. Instead of learning g_V and g_T from scratch, we build on a pre-trained CLIP model, which has been pre-trained on WebImageText (WIT), a dataset of 400 million image-text pairs gathered from the internet (Radford et al., 2021). CLIP has been shown highly performant on VT tasks, e.g., zero-shot image classification. We use the ViT-B/32 model in this work, which consists of a 12-layer vision Transformer (ViT) and a 12-layer language Transformer (Vaswani et al., 2017; Dosovitskiy et al., 2021). Given CLIP's strong VT alignment, we use its image encoder as g_V and text encoder as g_T. During learning, g_V and g_T are kept frozen and thus the joint VT representation space is untouched (see Figure 3). We minimize only the first loss term of Equation 3:

$$\min_{\Theta_A} \mathcal{L}_{cst}(V, A)\,, \quad (4)$$

where Θ_A are the trainable parameters of the audio encoder g_A.

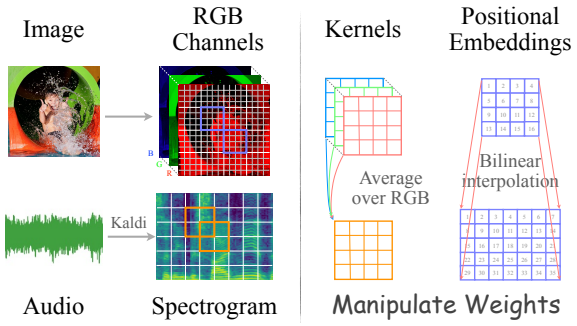


Figure 4: **Left:** three-channel image versus one-channel Spectrogram features of audio. We use ViT (Dosovitskiy et al., 2021) to encode images and audio. ViT uses a convolution layer to encode non-overlapped image patches into a sequence of image tokens, but for audio we modify the convolution stride to allow for overlaps between neighbor patches.
Right: adapting the convolution layer of ViT for audio encoding. For simplicity's sake, we omit the output channels of kernel weights and positional embeddings.

Audio encoder. Our audio encoder has the same vision Transformer architecture as CLIP's image encoder (ViT-B/32). In § 4, we show that initializing the audio encoder with CLIP's visual weights significantly improves convergence speed and accuracy. The architectural modifications which enable the use of visual CLIP's architecture for audio are (Figure 4 for an illustration):

(1) We customize the convolution stride to allow for overlaps between neighbor patches of Spectrogram features of audio.
(2) In the input embedding layer, we average the kernel weights of the convolution layer along the input channel to account for 1-channel Mel-filter bank features of audio (*cf.* RGB channels of images).
(3) We up-sample the 2-dimensional positional embeddings of image tokens to account for longer audio token sequences.

Image-audio pre-training. We conduct VA pre-training on AudioSet (AS; Gemmeke et al. (2017)).VA co-occurrence gathering, audio pre-processing, model hyperparameters, and training setups can be found in Appendix C. We measure the VA pre-training performance by retrieval precision and recall. Audio is relevant if it has the same set[2] of labels as the image query, and vice versa. Figure 5 illustrates the top-1 retrieval performance

[2]Recall that each audio clip in AudioSet is annotated with multiple labels.

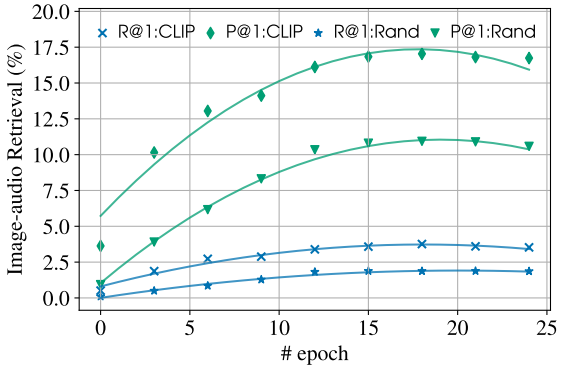


Figure 5: Image → Audio retrieval performance per image-audio pre-training epoch, evaluated on the AS balanced training set. "CLIP" and "Rand" indicates that the audio encoder is initialized from CLIP's image encoder and has random initialization, respectively.

with images as the query (similar trends are observed when using audio as the query). Compared with random initialization, initializing the audio encoder from CLIP's image encoder leads to faster convergence and better VA alignment. As we will see, this performance on VA retrieval transfers to downstream AT tasks.

3.3 Unsupervised and few-shot curation

To improve the AT alignments beyond pivoting, we consider curating audio-text pairs, and then performing an additional fine-tuning step by training the audio encoder with the AT loss, i.e., $\mathcal{L}_{cst}(A, T)$.[3] During AT fine-tuning, we keep the text encoder g_T frozen and only fine-tune the audio encoder.

Unsupervised curation. We consider explicitly mining AT pairs from the unsupervised pivoting model. Because this method *requires no human supervision* we refer to it as "unsupervised curation." Concretely, for each video segment in AudioSet, we extract a video frame, and input that frame to the original CLIP image encoder. Then, we encode a large set of candidate captions, and perform Image → Text retrieval over them by using the CLIP text encoder. The top candidate captions according to cosine similarity are then paired with the audio that corresponds to the original video clip.

We consider multiple caption sources to search over. As noted by Kim et al. (2019), captions for images and captions for environmental audio are significantly different in focus. We consider two vision-focused caption sets: (1) MSCOCO (Lin et al., 2014) captions (VC); and (2) because MSCOCO captions are limited to 80 object categories, we generate free-captions from GPT-J (Wang and Komatsuzaki, 2021) conditioned on MSCOCO captions as a prompt (FC). We additionally consider audio-focused captions from the training set of AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) (AC).[4] As a baseline, we also consider a random caption alignment, which assigns a random caption from AC to each clip (instead of pivoting on images). The bottom half of Table 2 summarizes different ways of curating AT pairs without additional supervision.

Few-shot curation. For comparison to our unsupervised methods, we also explore the effect of incorporating limited amounts of AT supervision, specifically, via captions from AudioCaps (GC) and textual labels of AudioCaps (GL).

4 Audio-text experiments

We use two types of tasks to evaluate the quality of the AT alignments learned by our model: AT retrieval and zero-shot audio classification.

AT retrieval. We conduct audio-text retrieval on two audio captioning datasets:

(1) **AudioCaps** (Kim et al., 2019) builds on AudioSet (Gemmeke et al., 2017) and provides captions for a subset of audio clips in AudioSet (sourced from YouTube). As we have pre-trained the audio encoder on AudioSet, we consider audio-text retrieval on AudioCaps as *in-domain* evaluation.
(2) **Clotho** (Drossos et al., 2020) consists of audio clips which have a duration of 15-30 seconds and come from Freesound (Font et al., 2013). It has a different sound source from AudioCaps and is used for *out-of-domain* evaluation.

We study the out-of-domain generalizability of our models by applying them to Clotho directly, without further fine-tuning on it.[5]

Zero-shot audio classification. We consider the following three widely used datasets for audio classification.

[3] Since our goal is to improve AT alignment, we primarily focus on AT fine-tuning; nonetheless, we compare AT fine-tuning to full VAT fine-tuning as in Equation 2 in Appendix F.

[4] We do not use the *alignment* of these captions — just the captions themselves.

[5] Clotho audio clips (15-30s) are longer than our pre-training audio clips (10s). See Appendix E for adaptation details.

Supervised	**GL**	**G**old textual **L**abels are used to construct AL pairs. (120816 aligned pairs)
	example	*Gurgling*
	GC	**G**old **C**aptions from AudioCaps provide an upper bound on the quality of AL alignment. (44118 aligned pairs)
	example	*Children screaming in the background as the sound of water flowing by.*
Unsupervised	**AC**	**A**udio-focused **C**aptions originate from the training captions of AudioCaps and Clotho. We perform caption retrieval by using CLIP and the prompt "the sound of". (1080078 aligned pairs)
	example	*A balloon is rubbed quickly and slowly to make squeaking sounds.*
	FC	**F**ree **C**aptions are generated by priming GPT-J with MSCOCO captions. We perform caption retrieval by using CLIP and the prompt "a photo of". (1224621 aligned pairs)
	example	*The blue colored person is jumping on the white and yellow beach ball.*
	VC	**V**ision-focused **C**aptions originate from MSCOCO. We perform caption retrieval by using CLIP and the prompt "a photo of". (1172276 aligned pairs)
	example	*A sky view looking at a large parachute in the sky.*
	RC	**R**andom **C**aptions indicates that we break the gold AL alignment in AudioCaps by randomly sampling a caption for each audio clip. They are used as a lower bound on the quality of AL alignment. (44118 aligned pairs)
	example	*A whoosh sound is heard loudly as a car revs its engines.*

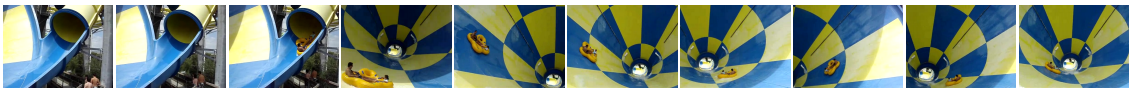


Table 2: Different ways of curating AT pairs. *Gurgling* is described as "the bubbling sound of water flowing through a narrow constriction, such as from a bottle with a narrow neck". The example comes from this YouTube video: 1O7-QuhweZE.

Model	AudioCaps				Clotho			
	Text→Audio		Audio→Text		Text→Audio		Audio→Text	
	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
Supervised SOTA	18.0	62.0	21.0	62.7	4.0	25.4	4.8	25.8
VA-Rand	1.3	7.3	5.6	24.5	1.3	7.5	3.2	13.5
VIP~ANT	0.8	7.9	10.1	38.1	1.9	9.5	7.0	25.6
+AT w/ GL	12.4	52.9	13.0	51.2	6.7	29.0	6.8	27.0
+AT w/ GC	**27.7**	**78.0**	**34.3**	**79.7**	**11.1**	**40.5**	**11.8**	**41.0**
+AT w/ AC	9.9	45.6	15.2	52.9	6.7	29.1	7.1	30.7
+AT w/ FC	8.9	41.5	14.7	50.0	6.5	27.7	7.8	29.7
+AT w/ VC	6.9	35.7	13.5	49.4	5.5	25.6	7.6	28.2
+AT w/ RC	3.8	19.9	10.7	38.1	3.5	16.9	5.5	24.9
OracleAV-CLIP	4.8	27.8	6.6	31.2				

Table 3: Audio caption retrieval performance (%) on AudioCaps test set and Clotho evaluation set. "Supervised SOTA" indicates the supervised state of the art from Oncescu et al. (2021b). OracleAV-CLIP: we replace audio with the corresponding image and evaluate image-text retrieval performance of CLIP (Radford et al., 2021). For each Clotho audio clip, we extract an audio clip which has a duration of at most 18 seconds and up-sample the positional embeddings accordingly. VA-Rand and VIP~ANT indicates that, in VA pre-training, the audio encoder is initialized randomly and from CLIP, respectively. We further fine-tune VIP~ANT on AT data, which is curated using different ways: GL, GC, AC, FC, VC, and RC (see Table 2 for details).

(1) **ESC50** (Piczak, 2015) contains 2000 audio clips from 50 classes. Each audio clip has a duration of 5 seconds and a single textual label. We follow the standard k-fold data splits.
(2) **US8K** (Salamon et al., 2014) contains 8732 audio clips from 10 classes. Each audio clip has a duration less than 4 seconds and a single textual label. We follow the standard k-fold data splits.
(3) **AudioSet** (Gemmeke et al., 2017) is a benchmark dataset for multi-label classification. AudioSet provides balanced and unbalanced training sets. The balanced set consists of 22-thousand au-

	Model	ESC50	US8K	AS
	Supervised	95.7$_{\pm 1.4}$	86.0$_{\pm 2.8}$	37.9
Zero-shot	VA-Rand	37.6(33.0)	41.9(38.1)	1.7(2.0)
	VIP~ANT	57.1(49.9)	44.7(37.8)	2.6(2.8)
	+AT w/ GL	67.2(64.5)	62.6(61.0)	15.4(**18.9**)
	+AT w/ GC	**69.5**(64.2)	**71.9**(67.1)	13.3(13.6)
	AudioCLIP	69.4	65.3	
	+AT w/ AC	62.8(55.7)	54.0(47.0)	11.6(12.3)
	+AT w/ FC	62.5(58.0)	52.7(50.0)	11.2(12.2)
	+AT w/ VC	61.9(58.0)	52.7(50.3)	8.9(10.7)
	+AT w/ RC	51.6(36.1)	42.3(28.5)	4.1(4.6)
	Wav2CLIP	41.4	40.4	

Table 4: Zero-shot audio classification accuracies (%) on ESC50 and US8K and mAPs (%) on AudioSet (AS). "Supervised" indicates that we fine-tune VIP~ANT for supervised audio classification. In the zero-shot setting, we use a prompt '*the sound of*' by default; accuracies in the parenthesis are obtained without using the prompt. "+AT" means that we fine-tune VIP~ANT on AT pairs curated through different ways. AudioCLIP is pre-trainined using the 2 million textual labels of AudioSet; +AT w/ GL and +AT w/ GC are trained with only 44K labels / captions. Wav2CLIP is most directly comparable to our fully unsupervised pivoting model VIP~ANT.

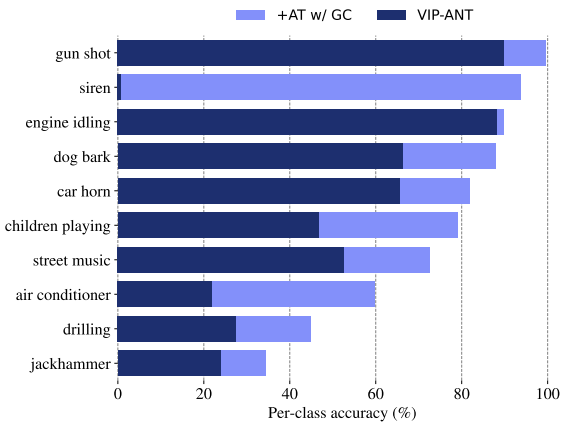


Figure 6: Per-class accuracy on US8K.

dio clips and the unbalanced set contains around 2 million audio clips. It also provides 20-thousand balanced audio clips for evaluation (more data statistics can be found in Table 5 in Appendix A).

For each audio clip $\boldsymbol{a}$, we predict the label $\boldsymbol{t}$ with the closest cosine similarity in the tri-modal space:

$$\arg\max_i \cos(\boldsymbol{t}^{(i)}, \boldsymbol{a}) \,. \tag{5}$$

4.1 Main results

Our prediction results for AT retrieval are given in Table 3 and for zero-shot classification in Table 4 (Appendix G contains qualitative results of the tri-modal representations).

Initializing with visual CLIP weights helps. Comparing VA-Rand to VIP~ANT, we see accuracy increases in all classification and retrieval setups. For example, on AudioCaps, VIP~ANT outperforms VA-Rand by 4.5% R@1 and 13.6% R@10. This confirms that the findings of Gong et al. (2021) carry-over to unsupervised audio pre-training.

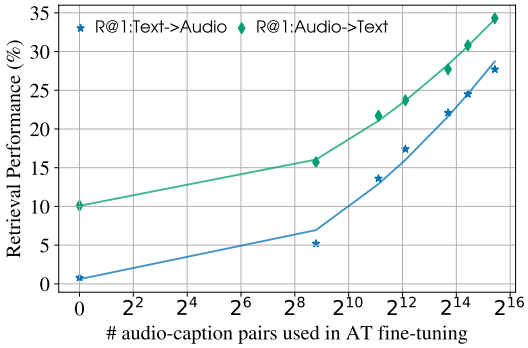
Pivoting works well for Audio $\rightarrow$ Text. VIP~ANT, exhibits surprisingly strong performance on AT retrieval tasks and zero-shot classification. For example, it outperforms the supervised baseline (Oncescu et al., 2021b) by 2.2% R@1 for text retrieval, without being trained or fine-tuned on Clotho, and without ever having seen an aligned AT pair.

Prompting (usually) helps. Inspired by the zero-shot image classification setups of CLIP (Radford et al., 2021), we prefix textual labels with a prompt in zero-shot audio classification. We empirically find that the prompt '*the sound of*' works well. Using it greatly improves zero-shot multi-class classification accuracy (see Table 4). Take VIP~ANT, the prompt gives rise to an improvement of 7.2% on ESC50 and 6.9% on US8K, but hurts multi-label classification performance on AS.

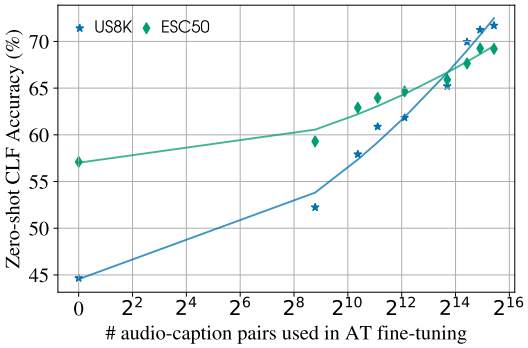
Random curation helps. Even when the audio-text pairs used to train that objective are sampled entirely at random (+AT w/ RC), VIP~ANT improves, e.g., R@1 for Text $\rightarrow$ Audio retrieval increases from 0.8% to 3.8%. We conjecture that RC at least makes audio representations aware of and lean towards the text cluster of the joint VT representation space. While this result also holds for AS classification (+1.5% mAP), performance decreases for ESC50 (-5.5% accuracy) and US8K (-2.4% accuracy).

Unsupervised curation is universally helpful. VIP~ANT fine-tuned with unsupervised audio captions (+AT w/ AC) outperforms both pivoting (VIP~ANT) and random curation (+AT w/ RC) in all cases. Thus, explicitly mining unsupervised AT pairs can be a helpful approach. Performance with automatically generated captions (FC) is similar to captions written by humans (AC).

Supervision is still the most helpful. Fine-tuning VIP~ANT on GC pairs leads to the highest accuracies on ESC50 and US8K. However, we do



(a) R@1 of AT retrieval on AudioCaps test set.



(b) Zero-shot classification (CLF) on ESC50 and US8K.

Figure 7: Audio retrieval and zero-shot classification performance versus level of language supervision.

not see similar improvements on AS, presumably because multi-label classification is more challenging and requires more direct language supervision, such as audio labels. This is further evident when we fine-tune VIP~ANT on GL and obtain the highest accuracy (18.9% mAP) on AS (see Table 4).

For retrieval, GL uses only audio labels as the text, which provide less dense language supervision than GC and is thus slightly worse than GC, but still, it gives better AT alignment than all automatic methods. As captions become semantically further from the audio-caption domain, e.g., GC < AC < FC < VC, the AT alignment becomes weaker, and thus leading to worse retrieval performance. The fine-tuned audio encoder generalizes to the out-of-domain Clotho successfully, displaying a trend similar to AudioCaps.

Supervision improves per-class accuracy in general. We further plot zero-shot classification accuracy for each audio class (see Figure 6 for US8K and Figure 12 in Appendix H for ESC50). Clearly, language supervision improves per-class accuracy in general. The highest improvement is observed on 'siren' because 'siren' rarely appears in image descriptions while GC contains a lot of textual descriptions of 'vehicle' audio.

4.2 Level of language supervision

We have observed that AT fine-tuning on AT pairs mined without any additional supervision (e.g., AC, FC, and VC) can improve the AT alignment, but supervised alignments are still the most effective. But: how much supervised data is really needed? To understand the relationship between supervision and performance, we vary the number of gold AT pairs (i.e., training samples of AudioCaps) used for AT fine-tuning. On the audio-text retrieval task (see Figure 7a), unsurprisingly, fine-tuning on more aligned AT pairs results in higher audio-text retrieval / zero-shot classification performance. Surprisingly, using only 442 (around 1%) AT pairs of AudioCaps gives rise to as strong AT alignment as VT alignment (*cf.* OracleAV-CLIP in Table 3).

Beyond the very few-shot setting, as we increase the number of supervised AT pairs used during fine-tuning, we observe a roughly linear relationship between zero-shot performance and the log of the number of supervised pairs (this observation is similar to (Kaplan et al., 2020)'s observations regarding transformers). While it's not clear how reliable extrapolations from this roughly linear trend are, we roughly estimating the amount of annotated AT pairs required for the zero-shot performance to equal human parity for ESC50 of 81% (Piczak, 2015): our estimate is that $2^{21} \approx 2$M supervised audio caption pairs would be needed. We're hopeful both (1) that larger curated audio-text datasets will become available; and (2) that future work can improve the data efficiency of the pre-training process.

5 Conclusion

We have presented VIP~ANT for unsupervised audio-text alignment induction. Based on the pivoting idea, our model learns image-text alignment and image-audio alignment explicitly and separately via bi-modal contrastive pre-training. The image modality is shared between the two and thus pivots audio and text in the tri-modal embedding space implicitly, without using any paired audio-text data. We empirically find that our model achieves strong performance on zero-shot audio-text tasks. We further strengthen the audio-text alignment by using varying kinds of audio-text supervision. Experimental results show that even un-aligned audio-caption pairs can help.

References

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 25–37. Curran Associates, Inc.

Venkatesh Boddapati, Andrej Petef, Jim Rasmusson, and Lars Lundberg. 2017. Classifying environmental sounds using image recognition networks. *Procedia Computer Science*, 112:2048–2056. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176, Berkeley, California, USA. Association for Computational Linguistics.

Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. 2017. Very deep convolutional neural networks for raw waveforms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 421–425.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740.

Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, page 411–412, New York, NY, USA. Association for Computing Machinery.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Morton Ann Gernsbacher. 2015. Video captions benefit everyone. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):195–202. PMID: 28066803.

Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2021a. Audioclip: Extending CLIP to image, text and audio. *CoRR*, abs/2106.13043.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2021b. Esresnet: Environmental sound classification based on visual domain models. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4933–4940.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2399–2409, Berlin, Germany. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Bongjun Kim and Bryan Pardo. 2019. Sound event detection using point-labeled data. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D. Plumbley. 2018. Audio set classification with attention model: A probabilistic perspective. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 316–320.

Bruno Korbar, Du Tran, and Lorenzo Torresani. 2018. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Anurag Kumar, Maksim Khadkevich, and Christian Fügen. 2018. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 326–330.

Anurag Kumar and Bhiksha Raj. 2017. Audio event and scene recognition: A unified approach using strongly and weakly labeled data. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3475–3482.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7327–7337.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision – ECCV 2020*, pages 121–137, Cham. Springer International Publishing.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision.

Richard F. Lyon. 2010. Machine hearing: An emerging field [exploratory dsp]. *IEEE Signal Processing Magazine*, 27(5):131–139.

Brian McFee, Justin Salamon, and Juan Pablo Bello. 2018. Adaptive pooling operators for weakly labeled sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2180–2193.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*.

Hideki Nakayama and Noriki Nishida. 2017. Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot. *Machine Translation*, 31(1/2):49–64.

Andreea-Maria Oncescu, João F. Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. 2021a. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2265–2269.

Andreea-Maria Oncescu, A. Sophia Koepke, João F. Henriques, Zeynep Akata, and Samuel Albanie. 2021b. Audio Retrieval with Natural Language Queries. In *Proc. Interspeech 2021*, pages 2411–2415.

Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3212.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 1041–1044, New York, NY, USA. Association for Computing Machinery.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York. Association for Computational Linguistics.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Luyu Wang, Pauline Luc, Adrià Recasens, Jean-Baptiste Alayrac, and Aäron van den Oord. 2021. Multimodal self-supervised learning of general audio representations. *CoRR*, abs/2104.12807.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2021. Wav2clip: Learning robust audio representations from CLIP. *CoRR*, abs/2110.11499.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

Yang You, Igor Gitman, and Boris Ginsburg. 2017. Scaling SGD batch size to 32k for imagenet training. *CoRR*, abs/1708.03888.

Abstract

This supplementary material includes (1) data statistics (§ A), (2) hyperparameters of optimizers (§ B), (3) details about bi-bi-modal pre-training (§ C), (4) supervised audio classification (§ D), (5) interpolating pre-trained position embeddings for Clotho audio-caption retrieval (§ E), (6) comparison between VAT fine-tuning and AT fine-tuning (§ F), (7) a qualitative study of the geometry of the tri-modal embedding space (§ G), and (8) additional findings from the audio-text retrieval task.

A Data statistics

Table 5 presents data statistics of all the datasets used in the paper.

B Optimizer hyperparameters

Table 6 presents optimizer hyperparameters used in our learning tasks.

C Bi-bi-modal pre-training

C.1 Audio-Video co-occurences

For training data, we gather VA co-occurrences from AudioSet,which contains temporally aligned audio and video frames from 10-second clips gathered from around 2 million YouTube videos (Gemmeke et al., 2017). To construct aligned image-audio pairs from AS, we adopt a sparse sampling approach (Lei et al., 2021): we first, extract four equal-spaced video frames from each clip. Then, during training, we randomly sample a frame from the four, and treat it as co-occurring with the corresponding audio clip. At test time, we always use the second video frame as the middle frame to construct image-audio pairs. We use the unbalanced training set, which consists of around 2 million video clips, to pre-train the audio encoder. Since AudioSet does not provide an official validation set, we validate the audio encoder and tune model hyperparameters on the balanced training set.

C.2 Audio preprocessing

We use Kaldi (Povey et al., 2011) to create Mel-filter bank features (FBANK) from the raw audio signals. Specifically, we use the Hanning window, 128 triangular Mel-frequency bins, and 10 millisecond frameshift. We always use the first audio channel when an audio clip has more than one channel. We apply two normalizations: (1) before applying Kaldi, we subtract the mean from the raw audio signals; and (2) we compute the mean and standard deviation of FBANK on the unbalanced AS training set, and then normalize the FBANK of each audio clip. For data augmentation, inspired by Gong et al. (2021), we frequency masking and the time masking: we randomly mask out one-fifth FBANK along the time dimension and one-forth FBANK along the frequency dimension during training.

C.3 Training dynamics

The architecture of our audio encoder follows the vision Transformer of CLIP (ViT-B/32, see (Radford et al., 2021) for more details). For the trade-off of efficiency and efficacy, we set the convolution stride to 16×24. This results in around 300 audio tokens for a kernel size of 32×32 and an input size of 1000×128 (all in the form of *time* $\times$ *frequency*). We optimize the model with LARS (You et al., 2017), where the initial learning rates for model weights and model biases are set to 2e-1 and 4.8e-3, respectively (detailed hyperparameters can be found in Table 6 in Appendix B). We pre-train our model on 4 NVIDIA Quadro RTX 8000 GPUs and for 25 epochs. We empirically set the batch size to 432 to fit the GPU memory. The full pre-training can be done within 24 hours.

D Supervised audio classification

To perform supervised audio classification, we add a classification head (a linear layer) on top of the pre-trained audio encoder. For *multi-class* classification, the classification head projects the vector representation of an audio clip onto the class space. We fine-tune the model by minimizing the cross-entropy loss:

$$\sum_i \log p(y^{(i)}|\boldsymbol{a}^{(i)}) \,, \tag{6}$$

where $y^{(i)}$ is the gold label of $\boldsymbol{a}^{(i)}$. For supervised *multi-label* classification, the classification head estimates the likelihood that an audio clip has some textual label. We thus minimize the per-label binary cross-entropy loss:

$$\sum_i \sum_l \log p(l = 1|\boldsymbol{a}^{(i)}) \,, \tag{7}$$

where l enumerates all possible audio labels.

ESC50 and US8K classification. We initialize the audio encoder from random initialization, CLIP, and VIP~ANT, respectively. Among them,

STAT.	AudioSet	ESC50	US8K	AudioCaps	Clotho
# Train	2041789 (unbalanced)	2000 (5-fold)	8732 (10-fold)	44118 ($\times$1 caption)	3839 (dev-train)
# Dev	22160 (balanced)				1045 (dev-val)
# Val				441 ($\times$5 caption)	1045 (dev-test)
# Test	20371 (balanced)			860 ($\times$5 caption)	1043 (withheld)
# Class	527	50	10		5 captions / audio
Duration	10s	5s	0-4s	10s	15-30s
Task	Multi-label CLF	Multi-class CLF	Multi-class CLF	Captioning	Captioning
Source	YouTube	Freesound	Freesound	YouTube (AudioSet)	Freesound

Table 5: Statistics of the data used in this paper. CLF is the abbreivation of "classification". In AudioSet (Gemmeke et al., 2017) audio clips come from distinct videos. Balanced split means that there are at least 59 samples for each of 527 sound classes. We managed to download 18036 out of 22160 videos in the balanced training split, 16416 out of 20371 videos in the test / evaluation split, and 1715367 out of 2041789 videos in the unbalanced split.

Hyperparam.	VA	AT	ESC50	US8K
Optimizer	LARS (You et al., 2017)			
Batch size	432	64	50	
Weight decay	1e-6			
LR of weight	2e-1		1e0	
LR of bias	4.8e-3		2.4e-2	
Warmup epoch	10			
Training epoch	25		50	

Hyperparam.	AS balanced	AS unbalanced
Optimizer	Adam (Kingma and Ba, 2015)	
Batch size	12	128
Weight decay	1e-7	
Learning rate	5e-5	
Warmup step	1000	
Training epoch	25	5
LR scheduler	MultiStepLR ($\gamma = 0.5$)	

Table 6: Hyperparameters of the optimizers used for VA pre-training, AL fine-tuning, ESC50 classification, US8K classification, balanced AS classification, and unbalanced AS classification. The learning rate (LR) in balanced AS classification is scheduled by epoch: 5, 9, 10, 11, 12 epochs. In unbalanced AS classification it is scheduled by optimization step: 7.5, 15, 20, 25, 35, 40, 45, 50 thousand steps.

AS Classification				
Dataset	AST	AST*	AST†	VIP~ANT
Unbalanced			43.4	**44.7**
Balanced	34.7	35.8	31.4	**37.9**

US8K and ESC50 Classification				
Dataset	AST-S	AST-P	CLIP	VIP~ANT
US8K			$82.5_{\pm 6.0}$	$\mathbf{86.0}_{\pm 2.8}$
ESC50	$88.7_{\pm 0.7}$	$95.6_{\pm 0.4}$	$89.7_{\pm 1.5}$	$\mathbf{95.7}_{\pm 1.4}$

Table 7: Multi-label classification mAPs (%) on AS and Supervised audio classification accuracies (%) on ESC50 and US8K. AST, AST-S, and AST-P indicates the results reported by Gong et al. (2021). We follow their suggestions and test the their best model (AST*) on our test set. Note that the best model has been trained on the combination of balanced and unbalanced AS training sets. † indicates that we follow the settings of AST and train it on our data. CLIP and VIP~ANT indicate that the audio encoder is initialized from CLIP and from VIP~ANT, respectively.

VIP~ANT performs best. It surpasses random initialization and CLIP on both datasets (see Table 7).[6] Notably, it outperforms the strong baseline AST-P on ESC50 (+0.1%), though AST-P has used gold audio labels for supervised pre-training.

AS classification. We consider balanced and unbalanced training for AS classification and train an individual model on the balanced set and the unbalanced set, respectively. Since the audio encoder has been pre-trained on the unbalanced AudioSet training set, it could be directly used without further fine-tuning. Nevertheless, we fine-tune the last k layers of the Transformer architecture of VIP~ANT and investigate whether task-specific fine-tuning helps(see Figure 8). When $k = 0$ the model is basically a linear probe. It inspects if contrastive pre-training learns separable audio representations. As we increase k, i.e., fine-tuning more layers, the model exhibits a tendency of over-fitting the training set. We use $k = 4$ as a trade-off between under-fitting and over-fitting. Our model achieves the best mAP of 37.9% for balanced training, which surpasses AST by 6.5% (see Table 7). While for unbalanced training, we find it crucial to fine-tune the whole model. Again, our model outperforms AST (+1.4% mAP).

[6] We find that VIP~ANT initialization leads to fast convergence, so it can bring better classification results than other initialization methods with the same number of training epochs.

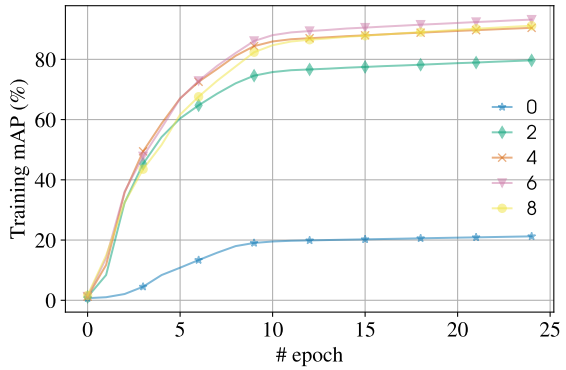


Figure 8: Fine-tuning last $k = 0, 2, 4, 6, 8$ layers of the pre-trained audio encoder for AS classification. mAP is measured on the AS balanced training set per fine-tuning epoch.

E Position embedding interpolation

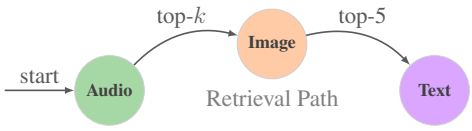
Clotho (Drossos et al., 2020) audio has a duration of 15-30 seconds, which is longer than 10-second audio clips used in pre-training. To apply our pre-trained audio encoder to Clotho audio-caption retrieval, we up-sample the pre-trained positional embeddings to account for the longer audio token sequences. Table 8 shows retrieval performance of 10-second audio and 18-second audio. In general, longer audio gives rise to better audio-caption retrieval performance.

F VAT versus AT fine-tuning

Given caption-augmented AudioCaps audio (Kim et al., 2019), we can improve the pre-trained audio encoder via contrastive vision-audio-text (VAT) fine-tuning and contrastive audio-text (AT) fine-tuning. Figure 9 shows a comparison between the two fine-tuning techniques on zero-shot ESC50 classification and AudioCaps audio retrieval. In general, AT fine-tuning results in better results on the two tasks.

G Analyzing tri-modal representations

To better understand the geometry of tri-modal embeddings of our pivoting, unsupervised curation, and supervised curation, we study how AT fine-tuning influences the tri-modal representation space. Specifically, we analyze VIP~ANT (pivoting), VIP~ANT +AT (w/ RC) (unsupervised curation), and VIP~ANT +AT (w/ GC) (supervised curation) using *pivotability*.



Pivotability measures how likely images can pivot audio and text. We quantify it for each aligned VAT triplet via a two-step retrieval probe. Starting at a given audio clip, we retrieve k nearest image neighbors; for each image neighbor, we retrieve the top-5 nearest captions. Since each audio clip has 5 gold captions, we compute pivotability as the ratio of the number of retrieved gold captions to 5. A gold caption may be retrieved more than one time, but we always count it as 1, so pivotability is always between 0 and 1.

We conduct this experiment on AudioCaps test set. For each k, i.e., how many images will be retrieved for a given audio clip, we average pivotability scores over all test triplets (see Figure 10).

Which pairs are pivotable? To study what kinds of audio are more likely to be pivoted with text by images, we set $k = 5$, i.e., 5 images will be retrieved for each given audio clip. We consider an AT pair as pivotable if at least 3 out of 5 gold captions of the audio clip are retrieved, i.e., pivotability is equal to or larger than 0.6. Figure 11 illustrates the categories of the audio clips in pivotable AT pairs. Unsurprisingly, audio about speech and vehicle is more pivotable because the two categories are among the top three frequent categories in AS.[7] Given that AT fine-tuning improves Audio → Image retrieval, we wonder if it could also help find novel categories of audio that can be pivoted with text. We find that this is indeed the case (see Table 9). For example, VIP~ANT +AT (w/ GC) finds more fine-grained speech categories because most AT pairs in AudioCaps are about speech. In contrast, VIP~ANT +AT (w/ RC) finds two additional novel insect categories, presumably because RC suffers from less data bias than GC.

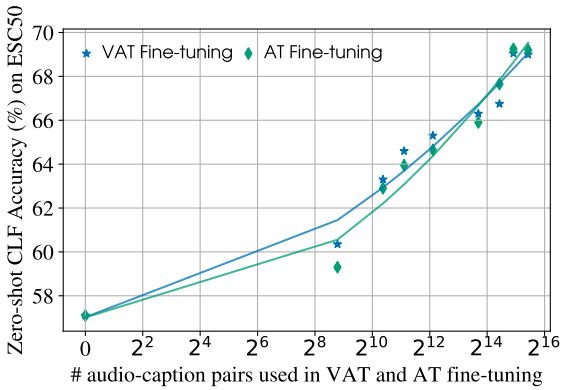
H Additional results

Asymmetric retrieval performance. For Text → Audio retrieval, our unsupervised pivoting model is not as good as on Audio → Text. This could be because audio is intrinsically more difficult to retrieve with specificity than text in our corpus, e.g., because sound events co-occur (a baby

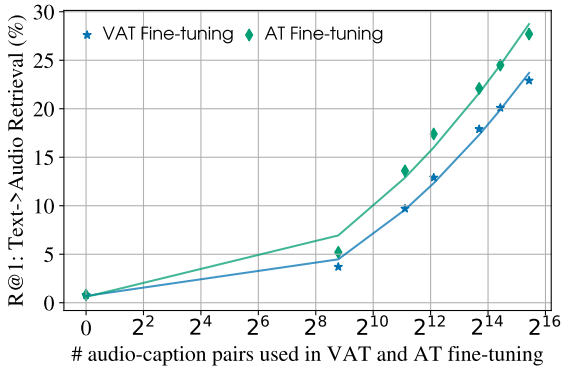
[7] Music is the second most frequent category in AS. It is not shown in the figure because AudioCaps excludes all music audio.

Model	10-second Clotho (eval)				18-second Clotho (eval)			
	Text→Audio		Audio→Text		Text→Audio		Audio→Text	
	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
VA-Rand	1.4	7.4	3.2	13.1	1.3	7.5	3.2	13.5
VIP~ANT	1.9	10.1	6.1	23.7	1.9	9.5	7.0	25.6
+AT w/ GL	6.0	27.1	6.1	25.4	6.7	29.0	6.8	27.0
+AT w/ GC	10.2	39.0	10.3	37.2	11.1	40.5	11.8	41.0
+AT w/ AC	5.9	26.3	8.2	30.3	6.7	29.1	7.1	30.7
+AT w/ FC	5.7	26.6	6.6	28.0	6.5	27.7	7.8	29.7
+AT w/ VC	5.2	25.2	7.0	25.9	5.5	25.6	7.6	28.2
+AT w/ RC	3.5	16.3	5.7	23.6	3.5	16.9	5.5	24.9

Table 8: Interpolating positional embeddings to account for Clotho audios which are longer than 10 seconds.



(a) Zero-shot classification (CLF) accuracy versus level of supervision.



(b) R@1 of audio retrieval with text as the query versus level of supervision.

Figure 9: Comparing VAT and AT fine-tuning on zero-shot ESC50 classification and AudioCaps audio retrieval.

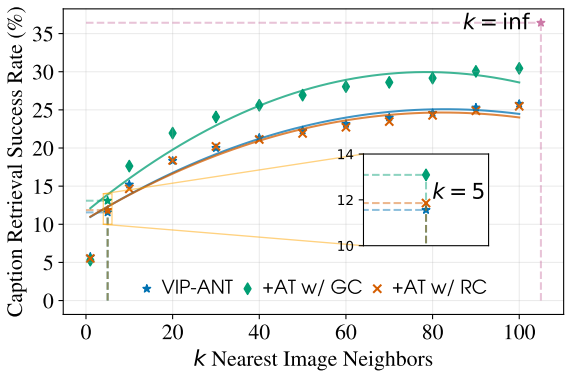


Figure 10: Tri-modal pivotability. +AT (w/ GC) and +AT (w/ RC) indicate that VIP~ANT is further fine-tuned on GC and RC, respectively.

Figure 11: Categories of the audio that can be pivoted with text by images. Larger text indicates that the related audio is more likely to be pivoted with text.

may cry in street with sirens in the background or in a room with dogs barking), there may be a broader range of captions that accurately describe them. However, it could also be the case that AT alignment is bounded by VT alignment because VA pre-training biases audio representations towards image representations. We check this hypothesis by conducting image-text retrieval on AudioCaps. AudioCaps provides aligned image-audio-text triplets, so we simply replace audio with the corresponding image. We find that the Text → Image retrieval performance of CLIP is much better than the Text → Audio retrieval performance of VIP~ANT (see the OracleAV-CLIP row of Table 3). It is also close to the Image → Text retrieval performance of CLIP.

+AT w/ GC	'female speech, woman speaking', 'narration, monologue', 'vibration'
+AT w/ RC	'bee, wasp, etc.', 'female speech, woman speaking', 'insect', 'narration, monologue', 'vibration'

Table 9: Comparing against VIP~ANT, the two fine-tuned versions of VIP~ANT find novel audio categories in pivotable AT pairs.

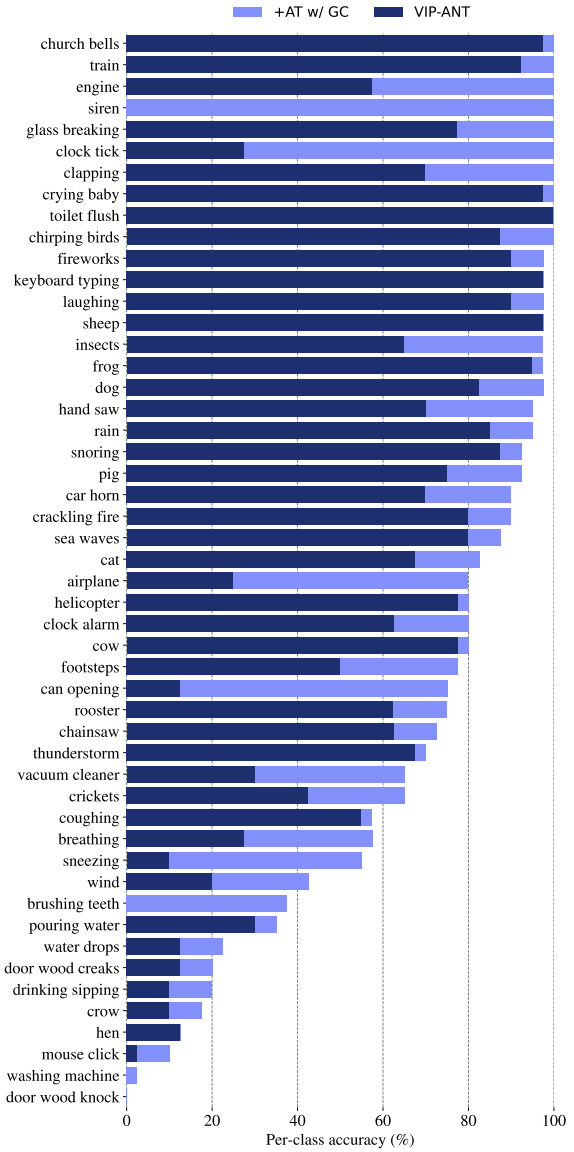


Figure 12: Per-class accuracy on ESC50.

In contrast, VIP~ANT exhibits a large gap between the Text → Audio retrieval performance and the Audio → Text retrieval performance.

Per-class accuracy on ESC50 is illustrated in Figure 12.