# BLAT: Bootstrapping Language-Audio Pre-training Based on AudioSet Tag-guided Synthetic Data

Xuenan Xu
Shanghai Jiao Tong University
wsntxxn@sjtu.edu.cn

Zhiling Zhang
Shanghai Jiao Tong University
blmoistawinde@sjtu.edu.cn

Zelin Zhou
Shanghai Jiao Tong University
ze-lin@sjtu.edu.cn

Pingyue Zhang
Shanghai Jiao Tong University
williamzhangsjtu@sjtu.edu.cn

Zeyu Xie
Shanghai Jiao Tong University
zeyuxie29@gmail.com

Mengyue Wu*
Shanghai Jiao Tong University
mengyuewu@sjtu.edu.cn

Kenny Q. Zhu*
University of Texas at Arlington
kenny.zhu@uta.edu

## ABSTRACT

Compared with ample visual-text pre-training research, few works explore audio-text pre-training, mostly due to the lack of sufficient parallel audio-text data. Most existing methods incorporate the visual modality as a pivot for audio-text pre-training, which inevitably induces data noise. In this paper, we propose to utilize audio captioning to generate text directly from audio, without the aid of the visual modality so that potential noise from modality mismatch is eliminated. Furthermore, we propose caption generation under the guidance of AudioSet tags, leading to more accurate captions. With the above two improvements, we curate high-quality, large-scale parallel audio-text data, based on which we perform audio-text pre-training. We comprehensively demonstrate the performance of the pre-trained audio-text model on a series of downstream audio-related tasks, including single-modality tasks like audio classification and tagging, as well as cross-modal tasks consisting of audio-text retrieval and audio-based text generation. Experimental results indicate that our approach achieves state-of-the-art zero-shot classification performance on most datasets, suggesting the effectiveness of our synthetic data. The audio encoder also serves as an efficient pattern recognition model by fine-tuning it on audio-related tasks. Synthetic data and pre-trained models are available online[1].

## CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence*; **Natural language processing**.

## KEYWORDS

multi-modal learning, contrastive learning, audio captioning, audio classification, zero-shot inference

**ACM Reference Format:**
Xuenan Xu, Zhiling Zhang, Zelin Zhou, Pingyue Zhang, Zeyu Xie, Mengyue Wu, and Kenny Q. Zhu. 2023. BLAT: Bootstrapping Language-Audio Pre-training Based on AudioSet Tag-guided Synthetic Data. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23): October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

## 1 INTRODUCTION

Multi-modal machine learning has become increasingly popular since it mimics our learning experience: we accept and handle information from different modalities. With the success of deep neural networks and large-scale datasets, we have witnessed the rapid development of multi-modal learning in recent years. Vision-language pre-training [10, 31, 34, 43] using Transformer has pushed the state of the art (SOTA) on a wide range of cross-modal tasks, such as visual question answering (VQA) [4], Image-Text Retrieval [32], visual commonsense reasoning (VCR) [50], etc. In these works, a joint representation of vision and language modalities is learned through pre-training on large-scale image-text datasets and then fine-tuned on specific downstream vision-language tasks.

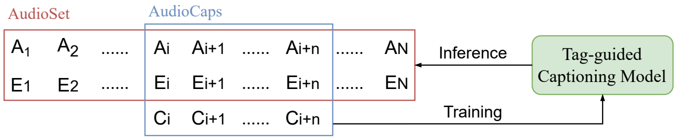

**Figure 1: The illustration of the data expansion approach. "A", "E" and "C" denote audio, event tags and caption respectively.**

In contrast with the rapidlying growing amount of work in vision-language pre-training, audio-related multi-modal learning, however, is still at a preliminary stage. Although audio is an important modality, few works explore pre-training involving audio

*Corresponding authors.

[1] The code, checkpoints and data are available at https://github.com/wsntxxn/BLAT and https://zenodo.org/record/8192397/

*MM '23, October 29–November 3, 2023, Ottawa, ON, Canada*

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/XXXXXX.XXXXXX

and language. The bottleneck of audio-language cross-modal learning lies in the scarcity of audio-text data. Compared with large-scale image-text datasets such as COCO [32] (~1.64M pairs), Visual Genome [30] (~5.06M pairs), and Conceptual Captions [42] (~ 12M pairs), current audio-text datasets contain only about 100K pairs (see Section 3.2). The lack of large-scale audio-text datasets may be attributed to the fact that not only the audio annotation cost is much higher than image description annotation [51], but audio-text co-occurrences are also scarcely available on the web [52].

To alleviate the above problem of data scarcity, prevailing works on audio-text cross-modal learning mostly incorporate CLIP [40], a powerful model enabling image-text alignment, to facilitate audio-language representation learning. The visual modality works as a pivot to connect audio and text since video-audio co-occurrences are abundant from massive video data. However, mismatching audio and visual modalities are commonly observed when detecting objects and events via sound and images. For example, visible objects in videos do not necessarily make sounds while sounds may be produced by objects off the screen. Such a mismatch leads to noise in audio-visual and audio-text alignment based on visual pivoting, indicated by the limited improvement achieved by these studies [22, 47, 52].

To better circumvent the noise when expanding data, we propose an audio-captioning-based approach to expand audio-text data using AudioSet [19], the largest free audio event dataset. We generate captions for audio directly without the aid of the visual modality so that potential noise from modality mismatch is eliminated. Compared with previous audio captioning works [8, 48], we incorporate AudioSet tags into caption generation to improve the generated caption quality. AudioSet contains audio clips and corresponding audio event tags in the original dataset. Its subset AudioCaps [26] provides captions on top of tags. Based on the provided event tags and captions, we bootstrap a tag-guided audio captioning model on AudioCaps and use it to generate large-scale audio-text data on AudioSet. The approach is shown in Figure 1. The bootstrapped data contains 1.22M pairs. To this end, we propose BLAT: **B**ootstrapping **L**anguage-**A**udio pre-training based on **T**ag-guided synthetic data, where contrastive learning is used to pre-train an audio-text bi-encoder just like CLIP.

The pre-training is comprised of two phases: 1) pre-training on the large-scale synthetic data; 2) further pre-training on the real data to adapt to the real distribution. We evaluate the performance of BLAT on a series of downstream tasks, including single-modality classification and cross-modal retrieval and generation. Results reveal that significant achievements are achieved on all tasks by fine-tuning BLAT. BLAT also achieves SOTA zero-shot classification performance on most datasets.

The main contribution of this paper can be summarized as follows:

- We use audio captioning to curate high-quality audio-text data from audio directly, eliminating the noise from other modalities.
- We incorporate AudioSet tags into audio-text data generation to bootstrap large-scale synthetic data for pre-training.
- We validate the effect of pre-training by transferring BLAT to cross-modal and single-modality tasks, achieving significant improvements under zero-shot and fine-tuning settings.

## 2 RELATED WORK

### 2.1 Vision-Language Pre-training

The research on multi-modal pre-trained models initially thrives in the intersection of vision and language modality. Vision-language pre-trained models generally handle three groups of tasks: understanding tasks like Classification, VQA and Visual Entailment, generation tasks like Image Captioning, and Image-Text Retrieval tasks. Researchers have proposed different model structures that are specifically suitable for certain group(s) of tasks. Cross-encoder models process multi-modal inputs in the same encoder to allow full interaction of the two modalities and thus are generally performing well on understanding tasks [10, 31]. Bi-encoder models encode the visual and textual inputs with different encoders to get separate embeddings [24, 40]. Since the embeddings can be pre-computed and stored for query, they are favorable for efficient retrieval. Encoder-decoder models encode single or both modalities in the encoder and use a decoder for generation, which provides the capability for generation tasks [45, 46]. Our model mainly adopts the Bi-Encoder paradigm. We exhibit that it can achieve competitive performance across all three groups of tasks.

For pre-training models, the data size has been shown to be vital for performance. Experimental results from the bi-encoder model CLIP show that its zero-shot image classification performance steadily increases with the number of images involved in pre-training. Another bi-encoder ALIGN [24] further scales up the pre-training data with noisy images from the web and shows that the models pre-trained on noisy data can still outperform those trained on higher-quality data given a larger data size. SimVLM [46], an encoder-decoder model, also achieves great success in both understanding and generation tasks with the large pre-training data ALIGN. Inspired by their findings, we propose synthesizing parallel audio-text data for audio-language pre-training, despite the potential noise in the synthetic data.

### 2.2 Audio-Language Pre-training

With the success of visual-language pre-training, a few recent works have started to incorporate audio into multi-modal pre-training. For instance, an audio encoder is added to CLIP with the contrastive learning paradigm. Large-scale video-text datasets are often utilized since the dataset provides visual-text alignment while audio-visual alignment is naturally available from the video data. VATT [1] and MMV [3] uses HowTo100M [37] and AudioSet for pre-training. The audio-text alignment is learned implicitly through the pivot of visual modality. AudioCLIP [22] performs the tri-modal contrastive learning explicitly by using AudioSet event tags as the corresponding text. Wav2CLIP [47], in contrast, does not incorporate text into pre-training but distills CLIP by audio-visual alignment training on VGGSound [7]. Following these works, we adopt contrastive pre-training to learn audio and text representation.

Compared with either textual AudioSet tags or video descriptions, VIP~A$_{\text{N}}$T [52] is proposed to use CLIP and the prompt "the sound of" to provide audio-focused descriptions for AudioSet audio clips. A frame of the corresponding video is used as the query. In this way, large-scale parallel audio-text pairs are automatically curated using the visual pivot. Audio-language pre-training without explicitly incorporating the visual modality is conducted on the curated audio-text data. Inspired by VIP~$\text{A}_{\text{N}}$T, we generate large-scale parallel audio-text data based on AudioSet and audio captioning. CLAP [15] is concurrent with our work. They adopt a similar contrastive learning framework while only current parallel audio-text data is used for training. We compare our model with these methods on zero-shot audio classification.

## 2.3 Audio Event Recognition

Audio event recognition requires recognizing the rich information in the sounds surrounding us, including the acoustic scenes where we are and what events are present. Audio event recognition contains various tasks like acoustic scene classification [36], audio tagging [19] and sound event detection [6]. In recent years, the release of Detection and Classification of Acoustic Scenes and Events (DCASE) challenges encourages the development of novel datasets, tasks and approaches. The release of AudioSet is also a milestone for audio event recognition. It contains 2.08M 10-second audio clips[2] with 527 annotated sound events. Robust audio representations can be learned by pre-training a deep neural network on AudioSet. Besides AudioSet, datasets like VGGSound and FSD50K [17] are also released recently to facilitate further research.

More recently, audio captioning [14] is proposed. Beyond audio event tags, a caption provides an unconstrained natural language description of an audio clip. Several datasets (see Section 3.1) are proposed to enable audio captioning research. Audio-text retrieval [39] is also proposed recently which requires retrieving audio signals using their textual descriptions and vice versa. It should be noted that though rich textual descriptions are provided in these datasets, they are all relatively small-scale. The audio-language pre-training in this work is conducted based on these small-scale audio-text datasets and the large-scale audio event dataset AudioSet. We evaluate our approach on these single-modality and multi-modal audio event recognition tasks.

## 2.4 Audio Representation Learning

Audio representation learning is an emerging field that has recently attracted increasing attention. It involves learning general-purpose representation which can be transferred to downstream audio-related tasks. Self-supervised speech representation [5, 9, 23] significantly improves performance on speech-related tasks. Audio representation [2, 38, 41] through self-supervised learning achieves competitive results on various tasks involving speech, music and general audio. With the release of AudioSet, many works improve the performance on audio event recognition tasks by pre-training on AudioSet [21, 28, 29]. Our work learns audio representation through audio-text contrastive learning. We validate the effectiveness of our approach by comparing it with a self-supervised COLA [41] and a tag-supervised PANNs [28].

# 3 BOOTSTRAPPING LANGUAGE-AUDIO DATA WITH AUDIOSET TAGS

In this work, we use both currently available audio-text datasets and synthetic parallel audio-text data for pre-training. We describe these datasets and the tag-guided data generation approach in this section.

## 3.1 Current Audio-Text Datasets

| Dataset | # Audio-text pairs | | | Avg # words | Duration /h |
|---|---|---|---|---|---|
| | train | val | test | | |
| AudioCaps | 49501 | 2475 | 4820 | 8.80 | 127 |
| Clotho | 19195 | 5225 | 5225 | 11.33 | 44 |
| MACS | 17275 | | | 9.25 | 11 |
| Total | 85971 | 7700 | 10045 | 9.60 | 182 |

**Table 1: Statistics of current English audio-text datasets.**

Current parallel audio-text datasets are from audio captioning, including AudioCaps [26], Clotho [13] and MACS [35]. AudioCaps is a subset of AudioSet, containing about 50K audio clips. Each audio clip in the training set has one caption annotation while five annotations are provided for audio clips in the validation and test set. Clotho contains 5,929 audio clips with five caption annotations provided for each clip. The audio data are collected from Freesound [18] platform. MACS is a recently released dataset built on TAU Urban Acoustic Scenes 2019 dataset containing 3,930 audio clips. Each audio clip is accompanied by several captions, ranging from two to five. MACS does not provide splits of training, validation or test. A summary of these datasets is in Table 1.

## 3.2 Audio-Text Data Generation from AudioSet

Although about 100K audio-text pairs are available in current audio-text datasets, the dataset size is much smaller than image-text datasets (e.g., ~1.64M pairs in COCO, see Section 1). However, large-scale audio event data are available from AudioSet. To leverage the large-scale audio-only data without caption description, we aim to generate captions for audio clips in AudioSet. Since AudioCaps is a subset of AudioSet, we first train a captioning model on AudioCaps and then use it to generate parallel audio-text data from AudioSet. Recent works tend to make use of supplementary information to guide captioning such as keyword [16] and similar captions [27]). However, these systems often suffer from poor prediction accuracy of supplementary information since the guidance can only be inferred from the input audio during inference. In AudioSet, the label, consisting of audio event tags presented in the audio clip, serves as effective guidance since it is available for all clips. Therefore, to enhance the quality of generated captions, we incorporate the event tags into caption generation. The model generates a caption conditioned on both the input audio and the hint from AudioSet tags. The architecture is shown in Figure 2. It contains an audio encoder and a text decoder. A sequence of audio features $\mathbf{x}$ is fed to the encoder and transformed into a sequence of high-level representations $\mathbf{e}^a$.

$$\mathbf{e}^a = \text{Encoder}(\mathbf{x})$$

[2] Only 1.95M clips are available in this work since some videos are removed.

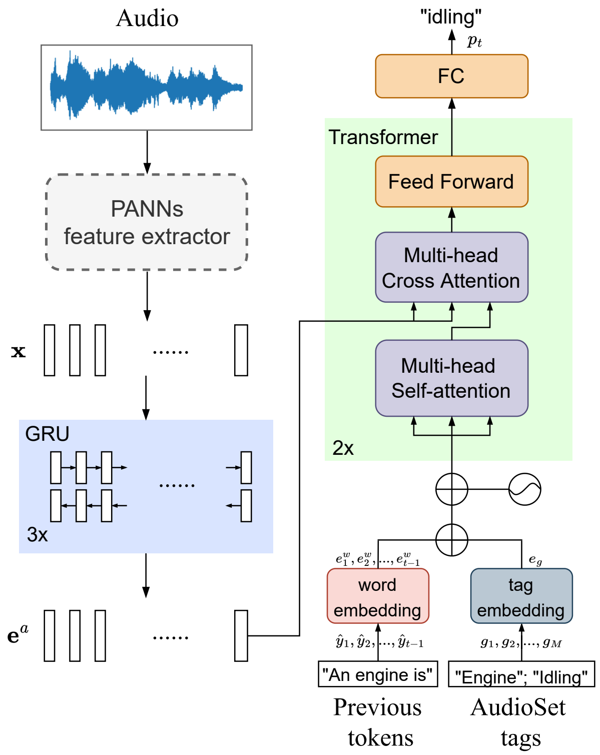

**Figure 2: The proposed audio captioning system with AudioSet tag guidance. The system generates caption based on both the input audio clip and the provided AudioSet tags.**

The decoder predicts the probability of each token at the time-step $t$ conditioned on partly decoded tokens $\{\hat{y}_n\}_{n=1}^{t-1}$, the provided AudioSet tags $\{g_m\}_{m=1}^{M}$ ($M$ is the number of tags) and $\mathbf{e}^a$:

$$p_t = \text{Decoder}(\mathbf{e}^a, \{\hat{e}_n\}_{n=1}^{t-1})$$

$$\hat{e}_n = e_n^w + e^g$$

$$e_n^w = \text{WE}(\hat{y}_n), \quad e^g = \frac{1}{M}\sum_{m=1}^{M}\text{TE}(g_m)$$

where WE and TE denote word embedding and tag embedding layers, transforming $\hat{y}_n$ and $g_m$ into fixed-dimensional vectors. Starting from the special "<BOS>" token, the decoder auto-regressively predicts the next token until "<EOS>" is reached.

In this work, we utilize deep embeddings from PANNs, specifically the *CNN14* variant, as the input audio feature $\mathbf{x}$. The encoder is a three-layer bidirectional gated recurrent unit (GRU) following [16] while the decoder is a two-layer Transformer with the final fully connected (FC) layer. The captioning system is trained by word-level cross entropy (CE) loss:

$$\mathcal{L} = \sum_{t=1}^{T} -\log\left(p_t(y_t)\right)$$

where $y_t$ is the ground truth token at the time-step $t$.

After training the AudioSet tag-guided captioning model, we use it to generate captions for large-scale AudioSet audio clips. However, the data distribution of AudioCaps is different from AudioSet since audio clips with specific event tags are excluded during the construction process of AudioCaps [26]. To circumvent the distribution bias problem, we exclude audio clips with tags that never appear in AudioCaps, with about 1.22M audio clips left. One caption is generated for each audio clip using the enhanced captioning model, resulting in about 1.22M audio-text pairs. We use this large-scale synthetic parallel audio-text data for pre-training.

## 4 AUDIO-TEXT PRE-TRAINING

In this section, we describe the proposed framework. The framework consists of an audio encoder and a text encoder for the two modalities. As Figure 3 shows, the model is first pre-trained by contrastive learning. The pre-training consists of two steps: 1) pre-training on synthetic parallel audio-text data; 2) further pre-training on the real data. Since there is a gap between the quality of real and synthetic data, the second pre-training step is adopted to alleviate the bias caused by synthetic data. We use the combination of all training sets of real audio-text data introduced in Section 3.1 for training.

After pre-training, the pre-trained model is transferred to several kinds of downstream tasks. Take audio classification as an example, the pre-trained model can be used for both zero-shot inference and fine-tuning. For zero-shot inference, the similarity scores between the audio clip and all textual labels are calculated as the estimated probabilities. For fine-tuning, a fully-connected (FC) layer is appended after the audio encoder for further classification fine-tuning to boost performance.

We first illustrate the contrastive pre-training approach. Then the architectures of the two encoders are introduced respectively.

### 4.1 Contrastive Pre-training

Similar to CLIP, the proposed contrastive learning approach learns the correspondence between the text content and the audio events in an arbitrary audio-text pair. For an audio clip $\mathcal{A}$ and a sentence $\mathcal{T}$, the audio and text encoders $\text{Enc}_A$ and $\text{Enc}_T$ transform them into two embeddings $\mathbf{a}$ and $\mathbf{t}$ respectively. A multi-modal embedding space is learned by maximizing the similarity between $\mathbf{a}$ and $\mathbf{t}$ of matched audio-text pairs and minimizing that of mismatched pairs. Following CLIP, the training objective is to minimize the InfoNCE loss [11]. Given a minibatch of $N$ audio-text pairs $(\mathcal{A}_1, \mathcal{T}_1), (\mathcal{A}_2, \mathcal{T}_2), \ldots, (\mathcal{A}_N, \mathcal{T}_N)$, their embeddings are calculated:

$$\mathbf{a}_i = \text{Enc}_A(\mathcal{A}_i)$$

$$\mathbf{t}_i = \text{Enc}_T(\mathcal{T}_i)$$

The training loss is a symmetric cross entropy loss between the predicted cosine similarity scores and the ground truth pairing

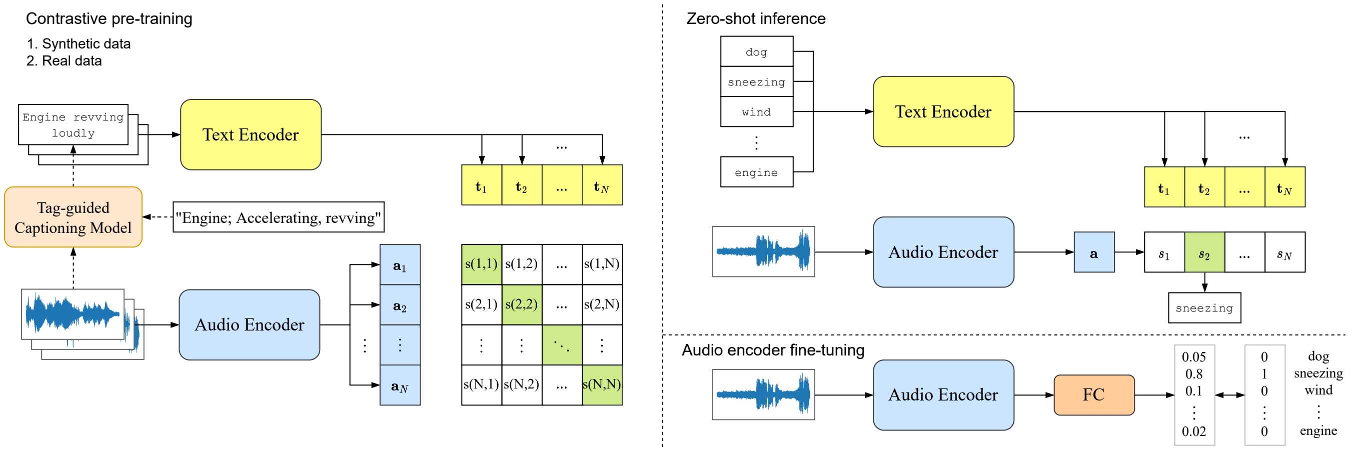

**Figure 3: An overview of our proposed language-audio pre-training approach. We use a tag-guided audio captioning model to generate audio-text data. Then we conduct contrastive learning similar to CLIP (dashed lines indicate the captioning model is not involved in the pre-training) in two stages. The pre-trained model can be transferred by zero-shot inference or fine-tuning.**

labels:

$$s(i,j) = \frac{\mathbf{a}_i \cdot \mathbf{t}_j^{\mathrm{T}}}{\|\mathbf{a}_i\| \cdot \|\mathbf{t}_j\|}$$

$$\mathcal{L}_i^{A\rightarrow T} = -\log \frac{exp\left(s(i,i)/\tau\right)}{\sum_{j=1}^{N} exp(s\left(i,j\right)/\tau)}$$

$$\mathcal{L}_i^{T\rightarrow A} = -\log \frac{exp\left(s(i,i)/\tau\right)}{\sum_{j=1}^{N} exp(s\left(j,i\right)/\tau)}$$

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_i^{A\rightarrow T} + \mathcal{L}_i^{T\rightarrow A})$$

where $\tau$ is the temperature optimized jointly with $\mathrm{Enc}_A$ and $\mathrm{Enc}_T$.

### 4.2 Audio Encoder

Similar to the feature extractor in Section 3.2, we use the pre-trained CNN14 from PANNs [28] as $\mathrm{Enc}_A$ instead of training the model from scratch. Time-frequency representation Log Mel Spectrogram (LMS) is extracted from the input audio and fed to 12 convolution blocks. $2 \times 2$ max pooling is done between every two blocks. After the convolution blocks, the audio embedding $\mathbf{a}$ is obtained by a global pooling on the feature map and transformation through a fully-connected layer. Although Transformer-based models are applied for audio classification recently [21] and achieve better performance than convolutional neural networks (CNN), it works on a sequence of patch embeddings without sub-sampling, resulting in high memory demand. Therefore, we adopt the pre-trained CNN14 to enable a larger minibatch size for training.

### 4.3 Text Encoder

For the text encoding part, we utilize BERT to transform $\mathcal{T}$ into $\mathbf{t}$. It is a deep Transformer pre-trained on large-scale corpora, including BooksCorpus and English Wikipedia, by self-supervised learning. Due to its powerful capability to extract representations with contextual semantics, BERT has exhibited superior performance on a series of language understanding tasks [12]. In this work, we employ $\mathrm{BERT}_{\mathrm{MEDIUM}}$ [44] as $\mathrm{Enc}_T$ for better computation efficiency and lower memory requirements. It consists of eight Transformer layers with a hidden embedding size of 512.

## 5 EXPERIMENTAL SETUP

In this section, we present our experimental setup for expanding audio-text data, pre-training, fine-tuning and downstream evaluation.

### 5.1 Synthetic Audio-Text Data Generation

The AudioSet tag-guided captioning model takes the feature extracted by PANNs CNN14 as the input. The encoder is a 3-layer bidirectional GRU and the decoder is a 2-layer Transformer. Details can be referred to [49]. The model is trained on AudioCaps for 25 epochs with a batch size of 64. The learning rate warms up to $5 \times 10^{-4}$ and then exponentially decays to $5 \times 10^{-7}$ until the end. We use beam search with a size of 3 when expanding audio-text pairs.

### 5.2 Pre-training

In the first pre-training step, we use a batch size of 128 and train the model for 200K iterations. About 1,200 audio-text pairs are randomly selected from the synthetic data to form a separate validation set. The model is validated every 500 iterations on the validation set. We use the Adam optimizer with the maximum learning rate of $1 \times 10^{-4}$. The learning rate is decayed by a cosine scheduler [33] with linear warm up in the first 10k iterations.

The model with the best performance on the synthetic validation set is used to initialize parameters for the second pre-training step. The setup is similar to the first step with several modifications on hyper-parameters. The total training iterations and warm up iterations are 15000 and 750 while the model is validated every 750 iterations.

## 5.3 Downstream Evaluation

| Task | Dataset | # Audio clips | Metric |
|---|---|---|---|
| Audio-text Retrieval | AudioCaps | 50K | R@K |
| | Clotho | 6K | |
| Audio Captioning | AudioCaps | 50K | COCO & FENSE |
| | Clotho | 6K | |
| Classification | ESC50 | 2K | Accuracy |
| | US8K | 8K | |
| | VGGSound | 192K | |
| Tagging | FSD50K | 50K | mAP |
| | AudioSet | 1.93M | |

**Table 2: A summary of downstream cross-modal and single-modality tasks. US8K is the abbreviation of UrbanSound8K.**

The pre-trained BLAT can be transferred to a series of downstream tasks, which is summarized in Table 2, including both cross-modal tasks and single-modality tasks. **Cross-modal audio-text tasks** include ***audio-text retrieval*** and ***audio captioning***. For audio-text retrieval, we use recall at K (R@K) as the evaluation metric. Standard COCO evaluation metrics from image captioning are used to evaluate audio captioning performance. Besides, we also incorporate FENSE [53] into evaluation for its higher correlation with human judgments.

**Single-modality tasks** include ***single-label (classification)*** and ***multi-label (tagging) audio classification***. Accuracy and mean average precision (mAP) are used for evaluation. We include several datasets with the size ranging from 2K to 1.93M for comparison with previous works.

## 5.4 Zero-shot Classification

With the pre-trained BLAT, we can perform zero-shot classification. If a textual label contains "_", we replace "_" with a blank. BLAT calculates the similarity scores between a given audio clip and all these textual labels. These scores are treated as the predicted probability of each audio event for evaluation.

## 5.5 Fine-tuning

Fine-tuning is commonly adopted to transfer the general-purpose pre-trained model to downstream tasks that may focus on specific domains. We illustrate the fine-tuning procedures for two cross-modal tasks and single-modality audio classification respectively.

*5.5.1 Audio-text Retrieval.* The fine-tuning on audio-text retrieval tasks uses almost the same configuration as the pre-training step. For both AudioCaps and Clotho, we fine-tune the pre-trained bi-encoder model for 20 epochs using the InfoNCE loss with a batch size of 128. The learning rate linearly warms up to the maximum value in the first epoch. The maximum learning rate for AudioCaps and Clotho is $5 \times 10^{-5}$ and $2 \times 10^{-6}$, respectively.

*5.5.2 Audio Captioning.* The audio captioning system is similar to the model in Section 3.2 except 1) the audio feature is extracted by BLAT instead of PANNs; 2) the system does not receive guidance from AudioSet tags. For both AudioCaps and Clotho, the training and inference configuration follows Section 5.1.

*5.5.3 Audio Classification and Tagging.* For single-modality tasks, we further fine-tune the pre-trained audio encoder $\text{Enc}_A$. An extra FC layer is added to $\text{Enc}_A$ for classification. We perform two types of fine-tuning: linear probing and fine-tuning the whole $\text{Enc}_A$. For linear probing, $\text{Enc}_A$ is used as a feature extractor and only the final FC layer is trained while no parameters are frozen in the second setting. Cross entropy loss and binary cross entropy loss are used for classification and tagging training respectively.

# 6 RESULTS

In this section, we present the comprehensive performance of BLAT. We first evaluate the quality of bootstrapped synthetic audio-text data. Then we reveal the influence of pre-training on downstream tasks. In experiments where only the audio encoder is used, we take PANNs [28] for comparison since both models share the same CNN14 architecture and use AudioSet for pre-training. We also incorporate self-supervised audio representation COLA [41]. For all experiments except pre-training, we report results based on three randomly seeded runs.

## 6.1 Benefits of Bootstrapped Audio-Text Data

| | $B_4$ | R | M | C | S | F |
|---|---|---|---|---|---|---|
| Synthetic w/o tag | 24.1 | 47.0 | 23.1 | 71.2 | 19.2 | 60.1 |
| Synthetic | 26.4 | 49.0 | 24.5 | 80.4 | 21.0 | 62.5 |
| Human | 29.0 | 49.5 | 28.8 | 90.8 | 28.8 | 68.0 |

**Table 3: The comparison of synthetic parallel audio-text data and real data in terms of audio captioning performance. Metrics include $\text{BLEU}_4$ ($B_4$), $\text{ROUGE}_\text{L}$ (R), METEOR (M), CIDEr (C), SPICE (S) and FENSE (F).**

*6.1.1 Data Quality Comparison on Captioning.* The quality of bootstrapped data is first evaluated in terms of captioning performance. We compare the performance of synthetic captions and human-annotated captions on AudioCaps test set. Since human annotations are used both as the candidate to be evaluated and the reference, we use a round-robin evaluation schedule. Specifically, we exclude one reference annotation in each round and evaluate the caption based on the left four annotations. The five scores are averaged as the performance indicator. We compare the performance of synthetic and human-annotated captions on AudioCaps in Table 3. We also include the captioning system without AudioSet tag guidance to show the effect of importing audio event tags. Metrics reveal that the tag guidance brings significant improvement. For $\text{ROUGE}_\text{L}$, the synthetic data performance is surprisingly comparable with human annotation. In terms of metrics evaluating the semantic similarity like SPICE and FENSE, human annotation is still much better. The model is capable of generating high-quality captions with the tag guidance though there is still a quality gap between the synthetic and real data.

| Training On Target | Configuration | AudioCaps | | | | Clotho | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Audio ⇒ Text | | Text ⇒ Audio | | Audio ⇒ Text | | Text ⇒ Audio | |
| | | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 |
| No | VIP~$A_N$T [52] | 15.2 | 52.9 | 9.9 | 45.6 | 7.1 | 30.7 | **6.7** | **29.1** |
| | template tags | 12.7 | 49.8 | 8.9 | 43.2 | 6.0 | 24.9 | 4.3 | 23.0 |
| | BLAT | **32.6** | **76.7** | **23.5** | **68.4** | **7.6** | **31.5** | 5.6 | 23.8 |
| Yes | from scratch | $40.4_{\pm1.5}$ | $85.7_{\pm0.7}$ | $33.3_{\pm0.3}$ | $82.4_{\pm0.3}$ | $13.9_{\pm0.2}$ | $48.2_{\pm1.4}$ | $12.3_{\pm0.6}$ | $46.1_{\pm0.9}$ |
| | BLAT fine-tuning | $47.5_{\pm0.4}$ | $87.6_{\pm0.2}$ | $38.2_{\pm0.1}$ | $85.1_{\pm0.1}$ | $17.9_{\pm0.4}$ | $50.9_{\pm1.9}$ | $13.7_{\pm0.4}$ | $48.9_{\pm0.5}$ |

**Table 4: Audio-text retrieval performance. The upper half denotes pre-training on different synthetic data and evaluating the pre-trained model without fine-tuning on the target dataset. The lower half shows the performance of training our model on the target dataset. R@K denotes recall at K.**

| Dataset | Audio Feature | $B_4$ | R | M | C | S | F |
|---|---|---|---|---|---|---|---|
| AudioCaps | COLA | $14.1_{\pm0.2}$ | $36.6_{\pm0.4}$ | $15.7_{\pm0.2}$ | $30.7_{\pm0.8}$ | $10.0_{\pm0.0}$ | $38.6_{\pm0.3}$ |
| | PANNs | $\mathbf{27.3_{\pm0.4}}$ | $\mathbf{49.7}_{\pm0.2}$ | $24.4_{\pm0.1}$ | $72.3_{\pm0.8}$ | $18.1_{\pm0.2}$ | $60.6_{\pm0.4}$ |
| | BLAT | $27.2_{\pm0.2}$ | $49.5_{\pm0.0}$ | $\mathbf{24.7_{\pm0.1}}$ | $\mathbf{73.3_{\pm0.4}}$ | $\mathbf{18.4_{\pm0.2}}$ | $\mathbf{61.5_{\pm0.3}}$ |
| Clotho | COLA | $10.0_{\pm0.6}$ | $31.0_{\pm0.2}$ | $13.0_{\pm0.2}$ | $18.7_{\pm1.9}$ | $7.5_{\pm0.2}$ | $29.9_{\pm0.7}$ |
| | PANNs | $15.8_{\pm0.3}$ | $\mathbf{37.6_{\pm0.2}}$ | $17.5_{\pm0.1}$ | $39.3_{\pm0.8}$ | $12.1_{\pm0.1}$ | $43.8_{\pm0.4}$ |
| | BLAT | $\mathbf{16.0_{\pm0.4}}$ | $\mathbf{37.6_{\pm0.0}}$ | $\mathbf{17.8_{\pm0.0}}$ | $\mathbf{41.5_{\pm0.6}}$ | $\mathbf{12.6_{\pm0.1}}$ | $\mathbf{45.8_{\pm0.6}}$ |

**Table 5: A comparison of audio captioning performance using different audio features.**

*6.1.2 Data Quality Comparison on Retrieval.* We also conduct the bootstrapped data in terms of zero-shot audio-text retrieval performance, shown in the upper half of Table 4. We compare our data with VIP~$A_N$T [52], which uses CLIP and the prompt "the sound of" to retrieve captions from AudioCaps and Clotho training corpus. The two synthetic datasets share a similar size (1.22M and 1.08M). For comparison, we also use a simple template "The sound of <tag 1 >, <tag 2 >, ..., and <tag n >" to convert AudioSet tags into captions, denoted as "template tags" in the table. BLAT significantly outperforms template tags on two datasets, indicating that our tag-guided captioning model can generate text data of higher quality. This is likely attributed to missing annotations in AudioSet (as elaborated in Section 6.3.1): missing tags make the template-based text less specific (e.g., the sound of speech) and comprehensive than that generated by our captioning model (e.g., a woman is speaking while something is being fried) bootstrapped from AudioCaps. The comparison between BLAT and VIP~$A_N$T shows that the model trained on our synthetic data significantly outperforms VIP~$A_N$T except for text-to-audio retrieval on Clotho. It indicates that using the visual modality as a pivot between audio and text leads to noisy data. The noise may come from the asynchronous audio and visual modalities. Note that Clotho captions are used to curate audio-text data in VIP~$A_N$T while in our work only AudioCaps is used. The distribution difference between AudioCaps and Clotho captions [35] leads to our model's unsatisfactory performance on Clotho.

## 6.2 Cross-modal Audio-and-Language Tasks

The lower half of Table 4 shows the performance of transferring BLAT to audio-text retrieval. We compare the model fine-tuning from BLAT with one trained from scratch. As the size of Clotho is small, the model trained from scratch performs poorly. With the initialization from BLAT, significant improvement can be witnessed on both AudioCaps and Clotho.

The performance of BLAT transferred to audio captioning is shown in Table 5. Without the supervision of event labels or textual descriptions, self-supervised COLA performs much worse than PANNs and BLAT. BLAT feature outperforms PANNs mainly on metrics regarding the semantic content like CIDEr and FENSE. This indicates that BLAT feature is more representative and helps the model generate more relevant descriptions.

## 6.3 Single-modality Audio Classification

*6.3.1 Zero-shot Transfer.* Under the zero-shot setting, the transferring ability of BLAT is evaluated. Previous works enabling zero-shot inference, including AudioCLIP [22], Wav2CLIP [47], VIP~$A_N$T [52], and CLAP [15] are incorporated for comparison. Except for CLAP, CLIP is utilized for synthetic data generation or pre-training. We also include current SOTA results as a topline for reference. Results are shown in the upper half of Table 6. The parameter numbers of these models are listed. Compared with works relying on CLIP, BLAT achieves SOTA zero-shot performance with a moderate model size, validating the benefit of eliminating noise from the visual modality. On VGGSound, BLAT outperforms Wav2CLIP even though the latter is pre-trained on VGGSound, indicating the transferring ability of BLAT. However, BLAT achieves a low mAP on AudioSet. Apart from the data distribution bias caused by the creation of AudioCaps, we observe that the noise in AudioSet labels exacerbates the problem. Previous works reveal that AudioSet annotations often contain only part of all events presented in a clip [20].

| | Model | # params / M | ESC50 | US8K | VGGSound (mAP) | FSD50K | AudioSet |
|---|---|---|---|---|---|---|---|
| | SOTA | - | 97.2 [22] | 90.1 [22] | 52.5 [25] | 65.3 [29] | 47.1 [29] |
| Zero-shot | AudioCLIP | 95.9 | 69.4 | 68.8 | - | - | - |
| | Wav2CLIP | **74.8** | 41.4 | 40.4 | - (10.0) | 3.0 | - |
| | VIP~A$_N$T | 151.3 | 69.2 | 71.7 | - | - | **13.3** |
| | CLAP | 192.1 | **82.6** | 73.2 | - | 30.2 | 5.8 |
| | BLAT | 123.7 | 80.6 | **77.3** | 14.9 (**13.5**) | **31.3** | 10.5 |
| Linear probing | COLA | 79.7 | 38.2$_{\pm 0.3}$ | 53.8$_{\pm 0.3}$ | 13.9$_{\pm 0.1}$ | 10.7$_{\pm 0.0}$ | 2.1$_{\pm 0.1}$ |
| | PANNs | 79.7 | 89.9$_{\pm 0.1}$ | 82.6$_{\pm 0.3}$ | 41.4$_{\pm 0.8}$ | 29.7$_{\pm 0.2}$ | - |
| | BLAT | 79.7 | **94.8**$_{\pm 0.3}$ | **85.7**$_{\pm 0.3}$ | **42.9**$_{\pm 0.5}$ | **32.4**$_{\pm 0.7}$ | **38.7**$_{\pm 0.0}$ |
| Fine-tuning | COLA | 79.7 | 78.8$_{\pm 0.7}$ | 75.3$_{\pm 0.4}$ | 48.7$_{\pm 0.8}$ | 47.9$_{\pm 0.9}$ | 43.6$_{\pm 0.2}$ |
| | PANNs | 79.7 | 95.4$_{\pm 0.1}$ | 87.4$_{\pm 0.2}$ | **55.3**$_{\pm 0.8}$ | 57.6$_{\pm 0.2}$ | - |
| | BLAT | 79.7 | **95.8**$_{\pm 0.2}$ | **89.0**$_{\pm 0.1}$ | 54.8$_{\pm 0.1}$ | **60.3**$_{\pm 0.5}$ | **44.0**$_{\pm 0.2}$ |

**Table 6: Audio classification and tagging performance in different settings: 1) zero-shot transfer 2) linear probing 3) fine-tuning. On VGGSound, we list mAP in parentheses to compare with Wav2CLIP. We only include parameters necessary for zero-shot inference when counting parameter numbers (i.e., the visual encoding part is excluded for AudioCLIP and VIP~A$_N$T).**

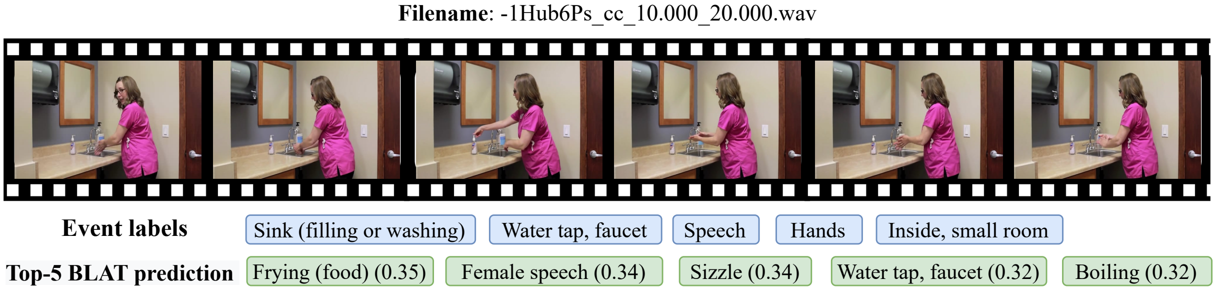

**Figure 4: An example of annotation errors in AudioSet. A woman is speaking in the audio clip while the corresponding event "Female speech, woman speaking" is not annotated.**

An example is shown in Figure 4. Speech from a woman can be clearly heard in the audio clip and CLAP assigns a high probability to the event "Female speech, woman speaking". However, the event does not occur in the AudioSet annotation. We assume such annotation noise makes the results not reliable. On FSD50K where annotations are more reliable, BLAT achieves a much higher mAP. Using a similar audio-text pre-training paradigm on real datasets, CLAP achieves similar results on ESC50 and FSD50K while BLAT outperforms CLAP on US8K and AudioSet with fewer parameters. This validates the benefit of incorporating our bootstrapped data into pre-training.

*6.3.2 Linear probing and Fine-tuning.* The lower half of Table 6 shows the results of transferring BLAT to audio classification by linear probing and fine-tuning. Since PANNs are trained on AudioSet with event labels, we do not further fine-tune PANNs on AudioSet. Like audio captioning, COLA performs much worse than PANNs and BLAT, especially under the linear probing setting. Although COLA can be applied to any audio data, its representation does not generalize well to audio classification tasks without the supervision of labels. In both linear probing and fine-tuning settings, BLAT outperforms PANNs on most datasets[3]. With only one FC classifier, the performance of linear probing BLAT on ESC50 and US8K is even close to current SOTA results, indicating that BLAT serves as a powerful feature extractor. Especially for small datasets, BLAT is able to extract highly discriminative features for classification. By fine-tuning BLAT, we achieve results close to SOTA, which validates its transferring ability to other tasks. The supervision of natural language exhibits a better ability to be transferred to a variety of audio classification tasks than event labels. Note that we do not adopt training techniques like data augmentation or task-specific loss functions.

## 7 CONCLUSION

In this work, we propose an AudioSet tag-guided audio captioning model to bootstrap large-scale audio-text data. Different from previous methods, the data generation approach does not incorporate video to eliminate the noise induced by the visual modality. Based on the bootstrapped data, we pre-train an audio-text bi-encoder using contrastive learning. After pre-training the model on the synthetic data and the real data successively, we obtain BLAT which

[3] Except VGGSound, results are significant at a level of 0.05.

can be transferred to a series of downstream tasks. Experimental results on both cross-modal and single-modality tasks, including retrieval, generation and classification, validate the effectiveness of BLAT. Under the stringent zero-shot condition where no training data is available, BLAT exhibits SOTA performance on most datasets.

## REFERENCES

[1] 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. *Proc. NIPS* 34 (2021).

[2] Haider Al-Tahan and Yalda Mohsenzadeh. 2021. Clar: Contrastive learning of auditory representations. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2530–2538.

[3] 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. *Proc. NIPS* 33 (2020), 25–37.

[4] Stanislaw Antol et al. 2015. Vqa: Visual question answering. In *Proc. ICCV*. 2425–2433.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proc. NIPS* 33 (2020), 12449–12460.

[6] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. 2015. Polyphonic sound event detection using multi label deep neural networks. In *International Joint Conference on Neural Networks (IJCNN)*. 1–7.

[7] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *Proc. IEEE ICASSP*. 721–725.

[8] Kun Chen et al. 2020. Audio Captioning Based on Transformer and Pre-Trained CNN.. In *Proc. DCASE*. 21–25.

[9] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* (2022).

[10] Yen-Chun Chen et al. 2020. Uniter: Universal image-text representation learning. In *Proc. ECCV*. Springer, 104–120.

[11] Van den Oord et al. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* 2, 3 (2018), 4.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristin Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL*. 4171–4186.

[13] Konstantinos Drossos et al. 2020. Clotho: An audio captioning dataset. In *Proc. IEEE ICASSP*. 736–740.

[14] Konstantinos Drossos, Sharath Adavanne, and Tuomas Virtanen. 2017. Automated audio captioning with recurrent neural networks. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 374–378.

[15] Benjamin Elizalde et al. 2022. CLAP: Learning Audio Concepts From Natural Language Supervision. *arXiv preprint arXiv:2206.04769* (2022).

[16] Ayşegül Özkaya Eren et al. 2020. Audio Captioning Based on Combined Audio and Semantic Embeddings. In *Proc. ISM*.

[17] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 30 (2022), 829–852.

[18] Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of ACM International Conference on Multimedia*. 411–412.

[19] Jort F Gemmeke et al. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP*. 776–780.

[20] Yuan Gong et al. 2021. PSLA: Improving Audio Tagging With Pretraining, Sampling, Labeling, and Aggregation. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 29 (2021), 3292–3306.

[21] Yuan Gong, Yu An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proceedings of Conference of the International Speech Communication Association*. ISCA, 56–60.

[22] Andrey Guzhov et al. 2022. Audioclip: Extending clip to image, text and audio. In *Proc. IEEE ICASSP*.

[23] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 29 (2021), 3451–3460.

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. PMLR, 4904–4916.

[25] Evangelos Kazakos et al. 2021. Slow-fast auditory streams for audio recognition. In *Proc. IEEE ICASSP*. 855–859.

[26] Chris Dongjoo Kim et al. 2019. AudioCaps: Generating Captions for Audios in The Wild. In *Proc. NAACL*. 119–132.

[27] Yuma Koizumi et al. 2020. Audio Captioning using Pre-Trained Large-Scale Language Model Guided by Audio-based Similar Caption Retrieval. *arXiv preprint arXiv:2012.07331* (2020).

[28] Qiuqiang Kong et al. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 28 (2020), 2880–2894.

[29] Khaled Koutini et al. 2021. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069* (2021).

[30] Ranjay Krishna et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73.

[31] Xiujun Li et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proc. ECCV*. Springer, 121–137.

[32] Tsung-Yi Lin et al. 2014. Microsoft coco: Common objects in context. In *Proc. ECCV*. Springer, 740–755.

[33] Ilya Loshchilov et al. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[34] Jiasen Lu et al. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proc. NIPS*.

[35] Irene Martin et al. 2021. Diversity and Bias in Audio Captioning Datasets. In *Proc. DCASE*. 90–94.

[36] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2018. A multi-device dataset for urban acoustic scene classification. In *Proc. DCASE*. 9–13.

[37] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proc. ICCV*.

[38] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. 2021. BYOL for audio: Self-supervised learning for general-purpose audio representation. In *International Joint Conference on Neural Networks (IJCNN)*. 1–8.

[39] Andreea-Maria Oncescu, A Koepke, João F Henriques, Zeynep Akata, and Samuel Albanie. 2021. Audio Retrieval with Natural Language Queries. In *Proceedings of Conference of the International Speech Communication Association*.

[40] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICML*. 8748–8763.

[41] Aaqib Saeed et al. 2021. Contrastive learning of general-purpose audio representations. In *Proc. IEEE ICASSP*. 3875–3879.

[42] Piyush Sharma et al. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proc. ACL*. 2556–2565.

[43] Weijie Su et al. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *Proc. ICLR*.

[44] Iulia Turc et al. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962* (2019).

[45] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. *arXiv preprint arXiv:2202.03052* (2022).

[46] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904* (2021).

[47] Ho-Hsiang Wu et al. 2022. Wav2CLIP: Learning Robust Audio Representations From CLIP. In *Proc. IEEE ICASSP*.

[48] Xuenan Xu et al. 2021. Investigating local and global information for automated audio captioning with transfer learning. In *Proc. IEEE ICASSP*. 905–909.

[49] Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kai Yu. 2022. *The SJTU System for DCASE2022 Challenge Task 6: Audio Captioning with Audio-Text Retrieval Pre-training*. Technical Report. DCASE2022 Challenge.

[50] Rowan Zellers et al. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proc. CVPR*. 6720–6731.

[51] Zhiling Zhang et al. 2021. Enriching Ontology with Temporal Commonsense for Low-Resource Audio Tagging. In *Proc. CIKM*. 3652–3656.

[52] Yanpeng Zhao et al. 2022. Connecting the Dots between Audio and Text without Parallel Data through Visual Knowledge Transfer. In *Proc. NAACL*. 4492–4507.

[53] Zelin Zhou et al. 2022. Can Audio Captions Be Evaluated with Image Caption Metrics?. In *Proc. IEEE ICASSP*.