Separate Anything You Describe

Xubo Liu, *Member, IEEE*, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, *Student Member, IEEE*, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D. Plumbley, *Fellow, IEEE*, and Wenwu Wang, *Senior Member, IEEE*

***Abstract*—Language-queried audio source separation (LASS) is a new paradigm for computational auditory scene analysis (CASA). LASS aims to separate a target sound from an audio mixture given a natural language query, which provides a natural and scalable interface for digital audio applications. Recent works on LASS, despite attaining promising separation performance on specific sources (e.g., musical instruments, limited classes of audio events), are unable to separate audio concepts in the open domain. In this work, we introduce AudioSep, a foundation model for open-domain audio source separation with natural language queries. We train AudioSep on large-scale multimodal datasets and extensively evaluate its capabilities on numerous tasks including audio event separation, musical instrument separation, and speech enhancement. AudioSep demonstrates strong separation performance and impressive zero-shot generalization ability using audio captions or text labels as queries, substantially outperforming previous audio-queried and language-queried sound separation models. Specifically, AudioSep achieved strong results including a signal-to-distortion ratio improvement (SDRi) of 7.74 dB across 527 sound classes of the AudioSet; 9.14 dB on the VGGSound dataset; 8.22 dB on the AudioCaps dataset; 6.85 dB on the Clotho dataset; 10.51 dB on the MUSIC dataset; 10.04 dB on the ESC-50 dataset; 8.16 dB on the DCASE 2024 Task 9 dataset; and a segmental signal to noise ratio (SSNR) of 9.21 dB on the Voicebank-Demand dataset.**

***Index Terms*—Language-queried audio source separation (LASS), natural language processing, sound separation.**

I. Introduction

COMPUTATIONAL auditory scene analysis (CASA) [1] aims to design machine listening systems that perceive complex sound environments in a similar way to the human auditory system. As a fundamental research task for CASA, sound separation aims to separate real-world sound recordings into individual source tracks, also known as the "cocktail party problem" [2]. Sound separation has a wide range of applications, including audio event separation [3], [4], music source separation [5], and speech enhancement [6], [7].

Many previous works on sound separation mainly focus on separating one or a few sources, such as in speech enhancement [6], [7], speech separation [8], [9], and music source separation [5]. Recently, universal sound separation (USS) [4] has attracted many research interests. USS aims to separate arbitrary sounds in real-world sound recordings. Separating every sound source from a mixture is challenging due to the wide variety of sound sources existing in the world. As an alternative, query-based sound separation (QSS) has been proposed which aims to separate specific sound sources conditioned on a piece of query information. QSS allows users to extract desired audio sources which could be useful in many applications such as automatic audio editing [10] and multimedia content retrieval [11]. Using the query of different modalities such as vision [12], [13], [14], audio [15], [16], [17], [18] or event labels [19], [20], [21], [22] for sound separation has been investigated in the literature.

Recently, a new paradigm of QSS has been proposed, known as language-queried audio source separation (LASS) [3]. LASS is the task of separating arbitrary sound sources using natural language descriptions of the desired source. LASS provides a potentially useful tool for future digital audio applications, allowing users to extract desired audio sources via natural language instructions. The use of natural language queries offers significant advantages, compared to previous audio-visual [12], [13], [14] or audio-queried [15], [16], [17], [18] methods, such as flexible and convenient acquisition of query information. Compared to label-queried [19], [20], [21], [22] methods that usually pre-define a fixed set of label categories, LASS does not limit the scope of input queries and can be seamlessly generalized to open domain.

The challenge of learning a LASS system is associated with the complexity and variability of natural language expressions. The text query description could range from sophisticated descriptions of multiple sound sources, such as "*people speaking followed by music playing with a rhythmic beat*" to simple and compact phrases such as "*speech, music*". In addition, the same audio source can be delivered with diverse language expressions, such as "*music is being played with a rhythmic beat*" or "*an upbeat music melody is playing over and over again*". LASS not only requires these phrases and their relationships to be captured in the language description but also that one or more sound sources that match the language query should be separated from the audio mixture.

Received 29 October 2023; revised 30 May 2024 and 3 December 2024; accepted 3 December 2024. Date of publication 3 January 2025; date of current version 27 January 2025. This work was supported in part by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/T019751/1 "AI for Sound", British Broadcasting Corporation Research and Development (BBC R&D), in part by the PhD scholarship from the Centre for Vision, Speech and Signal Processing, Faculty of Engineering and Physical Science, University of Surrey and a Research Gift from Google. The associate editor coordinating the review of this article and approving it for publication was Dr. Hakan Erdogan. *(Corresponding author: Xubo Liu.)*

Xubo Liu, Haohe Liu, Yi Yuan, Mark D. Plumbley, and Wenwu Wang are with the Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH Guilford, U.K. (e-mail: xubo.liu@surrey.ac.uk; haohe.liu@surrey.ac.uk; yi.yuan@surrey.ac.uk; m.plumbley@surrey.ac.uk; w.wang@surrey.ac.uk).

Qiuqiang Kong and Yuzhuo Liu are with the Speech, Audio and Music Intelligence Group, ByteDance Inc., China (e-mail: kongqiuqiang@bytedance.com; liuyuzhuo.999@bytedance.com).

Yan Zhao, Rui Xia, and Yuxuan Wang are with the Speech, Audio and Music Intelligence Group, ByteDance Inc., San Jose, CA 95110 USA (e-mail: zhao.yan@bytedance.com; rui.xia@bytedance.com; wangyuxuan.11@bytedance.com).

For reproducibility of this work, we released the source code, evaluation benchmark and pre-trained model at: https://github.com/Audio-AGI/AudioSep.

Digital Object Identifier 10.1109/TASLP.2024.3520017

Our original approach [3] for LASS relies on supervised learning with labeled audio-text paired data. However, such annotated audio-text data is limited in size. To overcome the data scarcity issue, recent advancements have investigated training LASS with multimodal supervision [23], [24], [25]. The key idea behind this approach is to leverage multimodal contrastive pre-training models such as the contrastive language-image pretraining (CLIP) model, as the query encoder. As contrastive learning is capable of aligning text embedding with other modalities (e.g., vision), it enables the training of the LASS system using data-rich modalities and facilitates inference with text in a zero-shot mode. However, existing LASS methods leverage small-scale data for training and focus on separating restricted source types, such as musical instruments and a limited set of sound events. The potential to generalize LASS to open-domain scenarios, such as hundreds of real-world sound sources, has not yet been fully explored. Furthermore, previous works on LASS evaluate model performance on each domain-specific test set, which makes future comparison and reproduction inconvenient and inconsistent.

In this work, our goal is to establish a foundation model for sound separation with natural language descriptions. Our focus is on the development of a sound separation model, leveraging large-scale datasets, to enable robust generalization in open-domain scenarios. With this model, we aim to holistically address the separation of a diverse range of sound sources. This work constitutes an extension of our previous research, as presented in the Interspeech proceeding [3]. The contribution of this work includes:

- We introduce AudioSep, a foundation model for open-domain, universal sound separation with natural language queries. AudioSep is trained with large-scale audio datasets and has shown strong separation performance and impressive zero-shot generalization capabilities. LASS-Net model presented in the Interspeech work did not perform well in open-domain scenarios.
- We extensively evaluate AudioSep. We construct a comprehensive evaluation benchmark for LASS research, involving numerous sound separation tasks such as audio event separation, musical instrument separation, and speech enhancement. We show that AudioSep substantially outperforms off-the-shelf audio-queried sound separation and state-of-the-art LASS models. AudioSep delivered strong results, such as an SDRi of 7.74 dB over 527 sound classes of the AudioSet; 9.14 dB on the VGGSound dataset; 8.22 dB on the AudioCaps dataset; 6.85 dB on the Clotho dataset; 10.51 dB on the MUSIC dataset; 10.04 dB on the ESC-50 dataset; 8.16 dB on the DCASE 2024 Task 9 dataset; and an SSNR of 9.21 dB on the Voicebank-Demand dataset.
- We conduct in-depth ablation studies to investigate the impact of scaling up AudioSep using large-scale multimodal supervision [23], [24], [25]. Our findings provide valuable insights for future research.
- We released the source code, evaluation benchmark, and pre-trained model at: https://github.com/Audio-AGI/AudioSep to promote research in this area.

II. Related Work

A. Audio Source Separation

Audio source separation [26] is a fundamental technique in signal processing, aimed at extracting independent source signals from their mixtures without prior knowledge of the mixing process. In recent years, audio source separation has shown significant advancements through integrating deep learning models [8]. Most deep learning-based source separation systems adopt the supervised learning approach, which usually requires the creation of simulated target and mixture sources and the developing of audio-to-audio mapping models. These mapping models generally fall into two categories: time domain-based and frequency domain-based approaches. Time domain-based methods focus on mapping directly onto the audio waveform, such as WaveUNet [27], ConvTasNet [9], and Demucs [28]. Frequency domain-based approaches conduct source separation within the spectral domain. While direct mapping on the audio spectrogram for audio source separation has demonstrated success [29], the mask-estimation-based methods, such as the estimation of ideal ratio masks [30], is still the most widely used approach in the frequency domain source separation. Recent work has also explored the integration of both time and frequency supervisions [31] to further enhance the performance and robustness of separation systems. Besides the mapping-based approach, deep clustering-based methods [32], [33] have also shown promising results on speech separation based on the learning of a representation space with contrastive objectives.

B. Universal Sound Separation

A substantial amount of source separation research has been concentrated on domain-specific sound separation, focusing primarily on areas such as speech [8], [9] or music [5]. Universal sound separation (USS) [4] aims to separate a mixture of arbitrary sound sources in terms of their classes. The challenge inherent in USS is the diversity of sound classes in real-world scenarios, which increases the difficulty of separating all of these sound sources with a single sound separation system. The work in [4] reported promising results on separating arbitrary sounds using permutation invariant training (PIT) [34], a supervised method initially designed for speech separation. The PIT method uses synthetic training mixtures simulated from single-source ground truth, performing sub-optimally due to a mismatch in the distribution between these synthetic mixtures and real-world sound recordings. Furthermore, it is not feasible to record a large database of single sources for PIT, as such sound recordings are often tainted by cross-talk. An unsupervised method called mixture invariant training (MixIT) [35] was proposed for sound separation using noisy audio mixtures. MixIT has achieved competitive performance compared to supervised methods (e.g., PIT), showing substantial improvements in reverberant sound separation performance. Both PIT and MixIT methods need a post-selection process to classify separated sources into specific sound classes.

C. Query-Based Sound Separation

Query-based sound separation (QSS), also known as target source extraction, aims to separate a specific source from an audio mixture given some query information. Existing QSS approaches could be divided into three categories: audio-visual, audio-queried, and label-queried.

1) Audio-Visual Sound Separation: In the computer vision community, there has been active research focusing on utilizing visual information to extract target sounds in speech [36], [37], music [12], and acoustic events [13]. AudioScope [14] has been recently proposed to perform on-screen sound separation based on the MixIT method. Such vision-queried approaches are beneficial for automatically decomposing sound sources from audio-visual video data. However, their performance is affected by the dynamic visibility conditions of visual objects, and can be degraded in environments with visual occlusion or low-light conditions. In addition, video data often contains off-screen sounds. Learning from noisy video data is a key challenge for audio-visual sound separation systems [14], [23].

2) Audio-Queried Sound Separation: Another line of research leverages the audio modality as a query to separate acoustically similar sounds. Recent studies [17], [18], [38], [39], [40] have addressed the problem of using one or a few examples of a target source as the query to separate a particular sound source: this is known as one-shot or few-shot sound separation. These methods separate the target sound conditioned on the average audio embedding of a few audio examples of the target source, which requires labeled single sources for calculating query embedding during training. The work in [15], [16] proposes to train the audio-queried sound separation system with large-scale weakly labeled data (e.g., AudioSet [41]), by first using a sound event detection model [42] to detect the anchor segment of sound events which are further used to constitute the artificial mixtures for the training of audio-queried sound separation models. These audio-queried sound separation approaches have shown great potential in the separation of unseen sound sources. However, during test time, the preparation of reference audio samples for the desired sound is often a time-consuming process.

3) Label-Queried Sound Separation: An intuitive way to query a specific sound source is to use the label of its sound class [19], [20], [21], [22]. Although acquiring a label query involves less effort than acquiring an audio or visual query, the label set is often pre-defined and is limited to a finite set of source categories. This imposes a challenge when attempting to generalize the separation system into an open-domain scenario, which may require re-training the sound separation model, or the use of continual learning methods [20], [43]. Label lacks the capability to describe the relationship between multiple sound events, such as their spatial relation and temporal order. This poses a challenge when the user intends to separate multiple sources rather than a single sound event.

D. Language-Queried Audio Source Separation

Language-queried audio source separation (LASS) is a recently proposed new paradigm of QSS. LASS uses a natural language description of an arbitrary target source to separate the source from an audio mixture. Such natural language descriptions can include auxiliary information for describing the target source, such as spatial and temporal relationships of sound events, for example "*a dog barks in the background*" or "*people applaud followed by a woman speaking*".

LASS-Net [3] was our first attempt to perform end-to-end language-queried sound separation. LASS-Net consists of a language query encoder and a separation model. The separation model performs target source separation in the frequency domain and the target waveform is reconstructed using the noisy phase and inverse short-time Fourier transform (iSTFT). LASS-Net is trained on a subset ($\sim$17.3 hours, 33 sound categories) of the AudioCaps [44] dataset and has shown great success in separating a wide range of sounds using audio caption queries. Kilgour et al. [24] proposed a similar model that accepts audio or text queries in a hybrid manner. Tzinis et al. [45] propose an optimal condition training (OCT) strategy for LASS. OCT performs greedy optimization toward the highest-performing condition among multiple conditions (e.g., signal energy, harmonicity) associated with a given target source, to improve the separation performance.

These LASS methods [3], [24], [45] require labeled text-audio paired data for supervised training, while such labeled data is often limited in practice. Recent work [23], [25] has investigated the potential of a self-supervised approach for LASS, including leveraging the visual modality as a bridge to learn the desired audio-textual correspondence. Instead of using a pre-trained language model (e.g., BERT [46]) [3], Dong et al. [23] use the contrastive language-image pretraining (CLIP) [47] model as the query encoder and train a LASS model conditioned on the visual context of unlabeled noisy videos. Thanks to the aligned embedding space learned by the CLIP model, at the inference time, the separation model can be queried with text inputs in a zero-shot setting. Experimental results show that combining text and image conditions for hybrid training leads to better text-queried sound separation performance on musical instrument separation. Although preliminary studies have been undertaken into LASS, existing approaches work under the constraints of limited-source scenarios, such as musical instruments and a restricted set of universal sound classes. As such, these approaches have not met the expectation of LASS for zero-shot, open-domain sound separation.

E. Multimodal Audio-Language Learning

Recently, the field of multi-modal audio-language has emerged as an important research area in audio signal processing and natural language processing. Audio-language tasks hold potential in various application scenarios. For instance, automatic audio captioning [48], [49] aims to provide meaningful language descriptions of audio content, benefiting the hearing-impaired in comprehending environmental sounds. Language-based audio retrieval [50], [51] facilitates efficient multimedia content retrieval and sound analysis for security surveillance. Text-to-audio generation [52], [53], [54], [55], [56], [57] aims to synthesize audio content based on language descriptions, serving

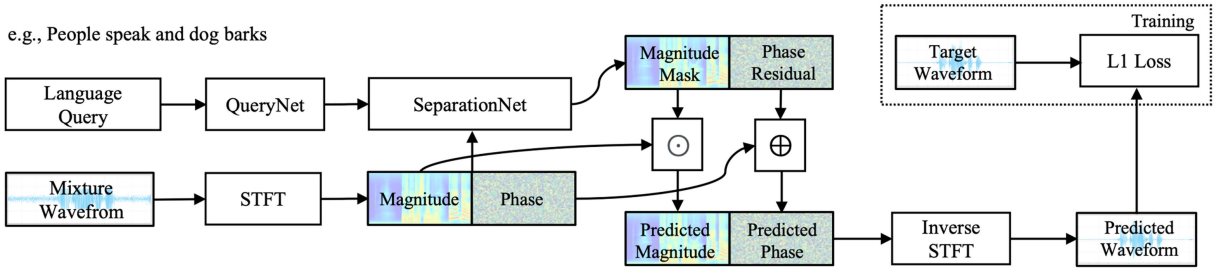


Fig. 1. Framework of AudioSep. AudioSep has two key components: a QueryNet and a SeparationNet. The QueryNet is the text encoder of CLIP [47] or CLAP [58] model. The SeparationNet is a frequency-domain ResUNet [5], [15] model.

as sound synthesis tools for film-making, game design, virtual reality, and digital media, and aiding text understanding for the visually impaired. Contrastive language-audio pre-training (CLAP) [58] aims to learn an aligned audio-text embedding space via contrastive learning. CLAP facilitates downstream audio-text multimodal tasks (e.g., zero-shot audio classification) [52], [59]. In this work, we focus on the intersection between audio source separation and natural language processing, which is an important field for CASA research but less explored.

III. Method

We introduce AudioSep, a foundation model for open-domain sound separation with natural language queries. AudioSep has two key components: a QueryNet and a SeparationNet, as illustrated in Fig. 1. We will introduce the details of each component below.

A. QueryNet

For QueryNet, we use the text encoder of the contrastive language-audio pre-training model (CLAP) [58]. The QueryNet is used to extract the text embedding of the natural language query. The input text query is denoted as $q = \{q_n\}_{n=1}^{N}$, consisting of a sequence of N word tokens. The sequence is processed by a text encoder to obtain the text embedding for the input language query. The text encoder encodes the input text tokens via a stack of Transformer blocks. After passing through the transformer layers, the output representations are aggregated, resulting in a fixed-length D-dimensional vector representation, where D corresponds to the latent dimension of the CLAP model. The text encoder is frozen during training.

B. SeparationNet

For SeparationNet, we apply the frequency-domain ResUNet model [5], [15] as the separation backbone. The input to the ResUNet model is a mixture of audio clips. First, we apply a short-time Fourier transform (STFT) on the waveform to extract the complex spectrogram $X \in \mathbb{C}^{T\times F}$, the magnitude spectrogram and phase of X are denoted as $|X|$ and $e^{j\angle X}$, where $X = |X|e^{j\angle X}$. Then, we follow the same setting of [15], and we construct an encoder-decoder network to process the magnitude spectrogram. The ResUNet encoder-decoder comprises 6 encoder blocks, 4 bottleneck blocks, and 6 decoder blocks. In each encoder block, the spectrogram is downsampled into a bottleneck feature using 4 residual convolutional blocks, while each decoder block utilizes 4 residual deconvolutional blocks to upsample the feature and obtain the separation components. A skip connection is established between each encoder block and the corresponding decoder block, operating at the same downsampling/upsampling rate. The residual block consists of 2 CNN layers, 2 batch normalization layers, and 2 Leaky-ReLU activation layers. Furthermore, we introduce an additional residual shortcut connecting the input and output of each residual block. The ResUNet model inputs the complex spectrogram X and outputs the magnitude mask $|M|$ and the phase residual $\angle M$ conditioned on the text embedding e_q. $|M|$ controls how much the magnitude of $|X|$ should be scaled, and the angle $\angle M$ controls how much the angle of $\angle X$ should be rotated. The separated complex spectrogram can be obtained by multiplying the STFT of the mixture and the predicted magnitude mask $|M|$ and phase residual $\angle X$:

$$\hat{S} = |M| \odot |X| e^{j(\angle X + \angle M)}, \tag{1}$$

where $\odot$ is the Hadamard product. The SeparationNet performs time-frequency masking in the magnitude domain, supplemented by phase correction relative to the noisy phase. An alternative approach is complex spectral mapping (CSM) [29], which predicts the real and imaginary spectrograms jointly and has been shown to perform better than time-frequency masking in speech enhancement task. However, in this work, we focus on the time-frequency masking-based approach.

To bridge the text encoder and the separation model, we use a Feature-wise Linearly modulated (FiLm) layer [60] after each ConvBlock in the ResUNet. Specifically, let $H^{(l)} \in \mathbb{R}^{m\times h\times w}$ denote the output feature map produced by ConvBlock l, where m represents the number of channels, and h and w are the spatial dimensions (height and width) of the feature map $H^{(l)}$, respectively. The modulation parameters are applied to each channel $H_i^{(l)} \in \mathbb{R}^{h\times w}$, where i indexes the i-th channel of the feature map, using the FiLm layer as follows:

$$\mathrm{FiLm}(H_i^{(l)}|\gamma_i^{(l)}, \beta_i^{(l)}) = \gamma_i^{(l)} H_i^{(l)} + \beta_i^{(l)} \tag{2}$$

Here, $\gamma_i^{(l)} \in \mathbb{R}$ and $\beta_i^{(l)} \in \mathbb{R}$ are the per-channel modulation parameters (scalars) for the i-th channel, generated from the text embedding e_q through a neural network $g(\cdot)$, i.e., $(\gamma^{(l)}, \beta^{(l)}) = g(e_q)$. The modulation parameters $\gamma_i^{(l)}$ and $\beta_i^{(l)}$ are broadcast

across the $h \times w$ spatial dimensions of the corresponding channel $H_i^{(l)}$, enabling feature-wise modulation, where i ranges over all channels, i.e., $i \in \{1, \ldots, m\}$. In this work, we model $g(\cdot)$ with two fully connected layers followed by ReLU activation, which is jointly trained with the ResUNet separation model.

C. Loss and Training

During training, we use the loudness augmentation method proposed in [15]. When constituting the mixture x with s_1 and s_2, we first calculate the energy of s_1 and s_2 as E_1 and E_2 by $E = \|s\|_2^2$. We then determine a desired Signal-to-Noise Ratio (SNR) in decibels (dB), denoted as SNR_{dB}, which is randomly chosen from the range $[-15, 15]$ dB. The scaling factor α is then calculated using the formula:

$$\alpha = \sqrt{\frac{E_1}{E_2} \cdot 10^{\text{SNR}_{\text{dB}}/10}}$$

This scaling factor ensures that the mixture x has the specified SNR. The mixture x is then formed as follows:

$$x = s_1 + \alpha s_2. \tag{3}$$

We train AudioSep end-to-end using an L1 loss function between the predicted and target waveforms. Since waveform-based L1 loss is simple to implement and has shown good performance on universal sound separation tasks [15].

$$\text{Loss}_{L1} = \|s - \hat{s}\|_1 \tag{4}$$

The lower L1 loss value indicates that the separated signal $\hat{s}$ is closer to the ground truth signal s.

D. Connections to the Prior Works

The design of the AudioSep model is extended from our previous works, LASS-Net [3] and USS-ResUNet [15]. LASS-Net is a model for language-queried sound separation, while USS-ResNet is a model for audio-queried sound separation. The separation network, ResUNet, is the same across AudioSep, LASS-Net, and USS-ResNet. For the text encoder in AudioSep, we use CLAP [58], which connects language and audio through an audio encoder and a text encoder via a contrastive learning objective. The CLAP model brings audio and text descriptions into a joint audio-text latent space, which facilitates effective audio-text representation learning.

IV. Datasets and Evaluation Benchmark

In this section, we provide a detailed description of the training dataset used for AudioSep, along with the established evaluation benchmark. The statistics of the dataset are presented in Tables I and II.

A. Training Datasets

1) AudioSet: AudioSet [41] is a large-scale, weakly-labelled audio dataset with 2 million 10-second audio snippets sourced from YouTube. Each clip in the collection is categorised by the presented sound classes, without the timing information of sound events. AudioSet includes an ontology[1] of 527 distinct sound classes, such as "Human sounds", "Music", "Natural Sounds", among others. The training set comprises 2063839 clips, including a balanced subset of 22160 audio clips. For each sound class in the balanced subset, there are at least 50 audio clips. After accounting for unavailable YouTube links, we were able to download 1934187 audio clips with their corresponding video streams, equating to 94% of the complete training set. All the clips are either padded with silence or cut short to a duration of 10 seconds. Since a large amount of YouTube audio recordings have a sampling rate below 32 kHz, we have converted all audio recordings to a mono format and resampled them at 32 kHz. All audio clips within the training set are used as training data.

2) VGGSound: VGGSound [61] is a large-scale audio-visual dataset sourced from YouTube. VGGSound contains nearly 200000 video clips each of length 10 seconds, annotated across 309 sound classes consisting of human actions, sound-emitting objects, and human-object interactions. The creation process of VGGSound has ensured that the object producing each sound is also discernible in the corresponding video clip. We utilize the original version of the VGGSound dataset, which includes 183727 audio-visual clips for training and 15449 for testing. We resample all audio clips at 32 kHz. All data within the training split are used for training.

3) AudioCaps: AudioCaps [44] is the largest publicly available human-annotated audio captioning dataset, with 50725 10-second audio clips sourced from AudioSet. AudioCaps is divided into three splits: training, validation, and testing sets. The audio clips are annotated by humans with natural language descriptions through the Amazon Mechanical Turk crowd-sourced platform. Each audio clip in the training sets has a single human-annotated caption, while each clip in the validation and test set has five human-annotated captions. We retrieved AudioCaps based on the AudioSet we downloaded. The AudioCaps dataset

TABLE I
AudioSep Training Datasets

	Caption	Label	Video	Num. clips	Hours
AudioSet	×	✓	✓	2 063 839	5800
VGGSound	×	✓	✓	183 727	550
AudioCaps	✓	✓	✓	49 768	145
Clotho v2	✓	×	×	4884	37
WavCaps	✓	×	×	403 050	7568

TABLE II
The Evaluation Benchmark for Open-Domain Sound Separation With Natural Language Queries

	Num. mixtures	Sr (Hz)	Query Type
AudioSet	5270	32 k	text label
VGGSound	1000	32 k	text label
AudioCaps	4785	32 k	caption
Clotho v2	5225	32 k	caption
ESC-50	2000	32 k	text label
MUSIC	5004	32 k	text label
DCASE 2024 T9	3000	16 k	caption
Voicebank-DEMAND	824	16 k	text label

[1][Online]. Available: https://research.google.com/audioset/ontology/index.html

we obtained consists of 49274 out of 49837 audio clips (98%) in the training set, 494 out of 495 clips (99%) in the validation set, and 957 out of 975 clips (98%) in the test set. All audio clips within the training and validation sets are used for our training process.

4) Clotho v2: Clotho v2 [62] is an audio captioning dataset that comprises sound clips obtained from the FreeSound platform[2]. Each audio clip in Clotho has been human-annotated via the Amazon Mechanical Turk crowd-sourced platform. Particular attention was paid to fostering diversity in the captions during the annotation process. In this work, we use Clotho v2 which was released for Task 6 of the DCASE 2021 Challenge[3]. Clotho v2 contains 3839, 1045, and 1045 audio clips for the development, validation, and evaluation split respectively. The sampling rate of all audio clips in the Clotho dataset is 44100 Hz, each with five captions. Audio clips are of 15 to 30 seconds in duration and captions are 8 to 20 words long. We merge the development and validation split, forming a new training set with 4884 audio clips. All audio clips within the new training set and evaluation set are resampled at 32 kHz.

5) WavCaps: WavCaps [63] is a recently released large-scale weakly-labeled audio captioning dataset, comprising 403050 audio clips with paired captions, totaling approximately 7568 hours. The audio clips constituting WavCaps originate from diverse sources, including FreeSound, BBC Sound Effects[4], SoundBible[5], and AudioSet. The audio captions are filtered and generated using the assistance of ChatGPT[6] based on the online-harvested raw audio descriptions. The average duration of audio clips is 67.59 seconds and the average text length of captions is 7.8 words. We resampled all audio clips within WavCaps at 32 kHz for training.

B. Evaluation Benchmark

1) AudioSet: The evaluation set of AudioSet [41] contains 20317 audio clips with 527 sound classes. We downloaded 18887 audio clips from the evaluation set (93%) out of 20317 audio clips. Source separation on AudioSet presents a considerable challenge, given that AudioSet contains a diverse range of sounds within a hierarchical ontology. To create evaluation data, we adopt the pipeline proposed in [15], which uses a sound event detection system [42] to analyze each 10-second audio clip to pinpoint anchor segments. Subsequently, two anchor segments from different sound classes are selected and combined to form a mixture with a SNR of 0 dB. We generate 10 mixtures for each sound class, leading to 5270 mixtures for all 527 sound classes in total.

2) VGGSound: In a similar way to [23], we manually selected 100 clean samples that each contain a distinct target sound event from the VGGSound test set. We refer to this set of 100 samples as VGGSound-Clean. For each audio sample from VGGSound-Clean, we randomly selected 10 audio samples from the remaining VGGSound test set to generate mixtures. Specifically, following [64], we first uniformly sampled the loudness of the two audio samples between -35 dB and -25 dB LUFS (Loudness Units Full Scale); then we mixed the signals together. The mixtures were scaled to 0.9 if clipping occurred. Finally, we constructed an evaluation set with 1000 samples. The average SNR of the evaluation set is around 0 dB.

3) AudioCaps: Our downloaded test set of the AudioCaps dataset [44] includes 957 audio clips, each annotated with five captions. To generate audio mixtures, we initially select an audio clip from the test set to serve as the target source, followed by a random selection of another audio clip as the background source, considering that the sound event tag[7] of the background source does not coincide with that of the target source. For the test mixtures, each test audio is mixed with five randomly chosen background sources with an SNR at 0 dB. Each mixture is assigned one of the five audio captions of the target source. Consequently, 4785 test mixtures are created.

4) Clotho v2: The Clotho v2 [62] evaluation set includes 1045 audio clips, each provided with five human-annotated captions. The duration of audio clips varies between 15 and 30 seconds. For the creation of test mixtures, we designate each audio clip in the evaluation set as a target source. Subsequently, we select two audio clips at random from the evaluation set, concatenate them, and then truncate it to match the length of the target source, thereby producing the interference source. Applying this pipeline, each audio clip in the evaluation set is mixed with five audio clips at an SNR of 0 dB. Each created mixture is then assigned one of the five audio captions of the target source. This procedure culminates in a total of 5225 mixtures for evaluation.

5) ESC-50: The ESC-50 dataset [65] contains 2000 environmental audio recordings evenly arranged into 50 semantic classes including natural sounds, non-speech human sounds, domestic sounds, and urban noise. Each class contains 40 examples, with each audio clip having a duration of 5 seconds and with a sampling rate of 44.1 kHz. To make a consistent evaluation, we first downsample all audio clips at 32 kHz. Then, we randomly mix two audio clips from different sound classes with an SNR at 0 dB to form a pair. We constitute 40 mixtures for each sound class. This leads to a total of 2000 evaluation pairs, which are used to evaluate the zero-shot performance of our model on environmental sound separation with text label queries.

6) MUSIC: The MUSIC dataset [12] is a collection of 536 video recordings of people playing a musical instrument from 11 instrument classes such as accordion, acoustic guitar, and cello. These video clips are crawled from YouTube and are relatively clean. Following the previous work [23], we downloaded the 46 video recordings in the test split. We further segmented all test videos into non-overlapping 10-second clips and resampled them at 32 kHz. For each video segment, we randomly select one segment from each of the other instrument classes to create a

[2][Online]. Available: https://freesound.org/
[3][Online]. Available: https://dcase.community/challenge2021
[4][Online]. Available: https://sound-effects.bbcrewind.co.uk/
[5][Online]. Available: https://soundbible.com/
[6][Online]. Available: https://openai.com/blog/chatgpt

[7]The sound event tags of each audio clip in AudioCaps can be retrieved from its corresponding AudioSet annotations.

mixture with an SNR at 0 dB, resulting in a total of 5004 evaluation pairs, which are used to evaluate the zero-shot performance of our model on musical instrument separation with text label queries.

7) DCASE 2024 T9: This is a new dataset[8] collected for evaluating the performance of LASS systems for DCASE 2024 Task 9: Language-Queried Audio Source Separation.[9] Specifically, 1000 audio files were sourced from the Freesound platform, uploaded between April and October 2023. Each audio file has been manually annotated with three captions. In the annotation guidance, annotators were instructed to describe the content of audio clips using five to twenty words. The tags of each audio file were verified and revised according to the FSD50K [66] sound event categories. Each audio file has been chunked into a 10-second clip and downsampled to 16 kHz. 3000 synthetic audio mixtures with signal-to-noise ratios (SNR) ranging from -15 dB to 15 dB are generated for evaluation. The revised tag information is used to ensure that the two audio clips used in each mix do not share overlapping sound source classes. We use this dataset to evaluate the zero-shot performance of our model on universal sound separation with natural language queries. For models trained with 32 kHz audio, we first upsample all audio clips to 32 kHz for separation and then downsample the separated audio clips to 16 kHz for evaluation.

8) Voicebank-DEMAND: The Voicebank-DEMAND dataset [67] integrates the Voicebank dataset [67], which includes clean speech, and the DEMAND [68] dataset, which encompasses a variety of background sounds that are used to create noisy speech. The noisy utterances are created by mixing the Voicebank dataset and the DEMAND dataset under signal-to-noise ratios of 15, 10, 5, and 0 dB. The test set of the Voicebank-DEMAND dataset includes a total of 824 utterances, which is used to evaluate the zero-shot performance of our model on speech enhancement. To make a fair comparison with previous speech enhancement systems [6], [7], [69], [70], we resample all audio clips to 16 kHz. We use "Speech" as the input text query to perform speech enhancement.

C. Evaluation Metrics

We utilize signal-to-distortion ratio improvement (SDRi) [15], [20] and scale-invariant SDR (SI-SDR) [71] to evaluate the performance of sound separation systems. For the speech enhancement task, following previous works [6], [7], [69], [70], we apply the perceptual evaluation of speech quality (PESQ) [72], mean opinion score (MOS) predictor of signal distortion (CSIG), MOS predictor of background-noise intrusiveness (CBAK), MOS predictor of overall signal quality (COVL) [73] and segmental signal-to-noise ratio (SSNR) [74] for evaluation. For each evaluation metric, higher values indicate better performance.

[8][Online]. Available: https://zenodo.org/records/10886481

[9][Online]. Available: https://dcase.community/challenge2024/task-language-queried-audio-source-separation

TABLE III
DIFFERENCES BETWEEN THE AUDIOSEP AND OTHERS IN TERMS OF THE MODEL SIZE, ARCHITECTURE, TRAINING DATA

	Training Data (hrs)	Parameters	Architecture
LASSNet [3]	17	63.4 M	BERT + ResUNet
CLIPSep [23]	550	181.6 M	CLIP + UNet
USS-ResNet30 [15]	5800	121 M	CNN + ResUNet
AudioSep	14 100	238.6 M	CLAP + ResUNet

V. EXPERIMENTS AND RESULTS

A. Training Details

We randomly sample two audio segments from two audio clips from the training set and mix them together to constitute a training mixture. The length of the audio segment is 5 seconds. We extract the complex spectrogram from the waveform signal with a Hann window size of 1024 and a hop size of 320. For the CLAP model, we use the publicly-available state-of-the-art checkpoint 'music_speech_audioset_epoch_15_esc_89.98.pt', which is trained on music, and speech datasets in addition to the original LAION-Audio-630 k dataset [58]. For the separation model, we use a 30-layer ResUNet consisting of 6 encoder and 6 decoder blocks, which is the same as the previous work [15] on universal sound separation. Each encoder block consists of two convolutional layers with kernel sizes of 3×3. The number of output feature maps of the encoder blocks is 32, 64, 128, 256, 512, and 1024, respectively. The decoder blocks are symmetric to the encoder blocks. We apply an Adam optimizer with a learning rate of 1×10^{-3} to train the AudioSep with the batch size of 96. We train the AudioSep model for 4M steps on 8 Tesla V100 GPU cards.

B. Comparison Systems

1) LASS Models: We employ two state-of-the-art publicly available LASS models as the comparison systems. The first one is LASS-Net [3], which uses a pre-trained BERT and ResUNet as the text query encoder and the separation model, respectively. LASS-Net is trained on a subset ($\sim$17 hours) of AudioCaps including universal sounds of categories such as human sounds, animal, sounds of things, natural sounds, and environmental sounds. The second one is CLIPSep, which uses CLIP [47] as the query encoder and a model based on Sound-of-Pixels (SOP) [12] for separation. CLIPSep [23] is trained with noisy audio-visual videos ($\sim$500 hours) from VGGSound [61] dataset using hybrid vision and text supervision signal. In addition, CLIPSep employs a training strategy called noise invariant training (NIT), designed to enable robust learning from noisy video data. As the CLIPSep model is trained with 16 kHz sampling rate, for 32 kHz audio clips in the evaluation sets, we downsample the audio clips to 16 kHz for evaluation. Both LASS-Net and CLIPSep are frequency domain separation models, the noisy phase is used to reconstruct the waveform in the test time. Differences between the AudioSep and others in terms of the model size, architecture, training data are described in Table III. AudioSep benefits from being trained on a significantly larger dataset, totaling 14100 hours, compared to the smaller datasets used by LASS-Net (17.3

TABLE IV
BENCHMARK EVALUATION RESULTS OF AUDIOSEP AND COMPARISON WITH THE STATE-OF-THE-ART LASS SYSTEMS

	VGGSound		AudioCaps		Clotho		MUSIC		ESC-50		DCASE 2024 T9	
	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi
LASSNet [3]	−4.50	1.17	−0.96	3.32	−3.42	2.24	−13.55	0.13	−2.11	3.69	−4.01	1.93
CLIPSep [23]	1.22	3.18	−0.09	2.95	−1.48	2.36	−0.37	2.50	−0.68	2.64	−1.09	1.90
AudioSep	**9.04**	**9.14**	**7.19**	**8.22**	**5.24**	**6.85**	**9.43**	**10.51**	**8.81**	**10.04**	**6.71**	**8.16**

TABLE V
EVALUATION RESULTS OF AUDIOSEP AND COMPARISON WITH AUDIO-QUERIED SOUND SEPARATION SYSTEMS ON AUDIOSET

	SI-SDR	SDRi	Query Modality
USS-ResUNet30 [15]	-	5.57	Audio
USS-ResUNet60 [15]	-	5.70	Audio
AudioSep	6.90	7.74	Text

hours) and CLIPSep (550 hours). We further show that the data scaling enables AudioSep to generalize better across various sound separation tasks.

2) Audio-Queried Sound Separation Models: We use audio-queried separation models as comparison systems. Kong et al. [15] proposed a universal sound separation system trained on AudioSet (∼5800 hours). This system performs query-based sound separation by using the average audio embedding calculated from query examples from the training set as the condition and can separate hundreds of sound classes using a single model. Empirical studies [15] conducted by Kong et al. have assessed the effectiveness of various sound separation systems such as ConvTasNet [9], UNet [75], and ResUNet [5], along with different audio embedding extractors such as PANNs [42] and HTSAT [76]. The results indicate that a system combining PANNs and ResUNet achieved the best performance. As a result, we adopt the PANN-ResUNet separation model with 30 and 60 layers of the ResUNet as the comparison systems, denoted as USS-ResUNet30 and USS-ResUNet60, respectively.

3) Speech Enhancement Models: Following [7], we adopt four off-the-shelf speech enhancement models as the comparison systems including Wiener filter [69], SEGAN [6], AudioSet-UNet [7], Wave-U-Net [70], CCMGAN [77] and MP-SENet [78]. Wiener filter [69] is a method based on signal processing methods. SEGAN is designed with the generative adversarial network [79]. AudioSet-UNet is a frequency-domain UNet-based model trained with weakly labeled AudioSet [41] data. Wave-U-Net is a time-domain UNet-based speech enhancement model. CMGAN [77] and MP-SENet [78] are two recently developed state-of-the-art models for speech enhancement. CMGAN [77] is a Conformer-based speech enhancement model optimized with MetricGAN [80]. MP-SENet [78] is an encoder-decoder architecture combining convolution-augmented transformers, designed to denoise magnitude and phase spectra simultaneously.

C. Evaluation Results on Seen Datasets

We first assess the performance of AudioSep on datasets seen during training including AudioSet, VGGSound, AudioCaps, and Clotho, as shown in Tables IV and V. AudioSep shows strong sound separation performance using text labels or audio captions as input queries. On the AudioSet, AudioSep achieved an SI-SDR of 6.90 dB and an SDRi of 7.74 dB. For the VGGSound dataset, the AudioSep model demonstrated an SI-SDR of 9.04 dB and an SDRi of 9.14 dB. On the AudioCaps and Clotho datasets, the AudioSep achieved an SI-SDR of 7.19 dB and 5.24 dB, respectively, along with SDRi values of 8.22 dB and 6.85 dB, respectively.

Two state-of-the-art models, LASS-Net and CLIPSep, perform poorly in separating target sounds on our benchmarks. Audio-queried sound separation baseline systems USS-ResUNet30 and USS-ResUNet60 have achieved an SDRi of 5.57 dB and 5.7 dB, respectively, which underperformed AudioSep by around 2 dB. In the inference stage, AudioSep only requires the text description of the target source, which is more convenient to obtain, as compared with anchor audio required in the audio-queried methods. Evaluation results on seen datasets indicate the strong performance of AudioSep compared with previous state-of-the-art LASS models and off-the-shelf audio-queried sound separation models.

D. Zero-Shot Evaluation Results

We conducted further evaluations to assess the zero-shot separation performance of AudioSep on unseen datasets, including MUSIC, ESC-50, DCASE 2024 T9, and Voicebank-DEMAND. AudioSep achieved promising zero-shot separation results as shown in Table IV. Specifically, for the MUSIC dataset, AudioSep achieved an SI-SDR of 9.43 dB and an SDRi of 10.51 dB. For the ESC-50 dataset, AudioSep obtained an SI-SDR of 8.81 dB and an SDRi of 10.04 dB. For the DCASE 2024 Task 9 dataset, AudioSep reached an SI-SDR of 6.71 dB and an SDRi of 8.16 dB. Although CLIPSep achieved good performance in separating musical instruments from background noise [23], its performance degrades when faced with musical instrument mixtures. Both CLIPSep and LASS-Net perform poorly in these evaluation datasets.

Table VI shows the results of Voicebank-DEMAND speech enhancement. Noisy speech without enhancement has PESQ, CSIG, CBAK, COVL, and SSNR of 1.97 dB, 3.35 dB, 2.44 dB, 2.63 dB and 1.68 dB respectively. AudioSep achieves PESQ, CSIG, CBAK, COVL, and SSNR of 2.43 dB, 3.37 dB, 3.17 dB, 2.88 dB and 9.21 dB respectively. AudioSep surpasses all the speech enhancement baselines in the PESQ metric and performs on par with traditional speech enhancement models including SEGAN [6] and Wave-U-Net [70]. While the results achieved by AudioSep in speech enhancement are far from the state-of-the-art methods such as CMGAN [77] and MP-SENet [78], this discrepancy may be attributed to the presence of large-scale

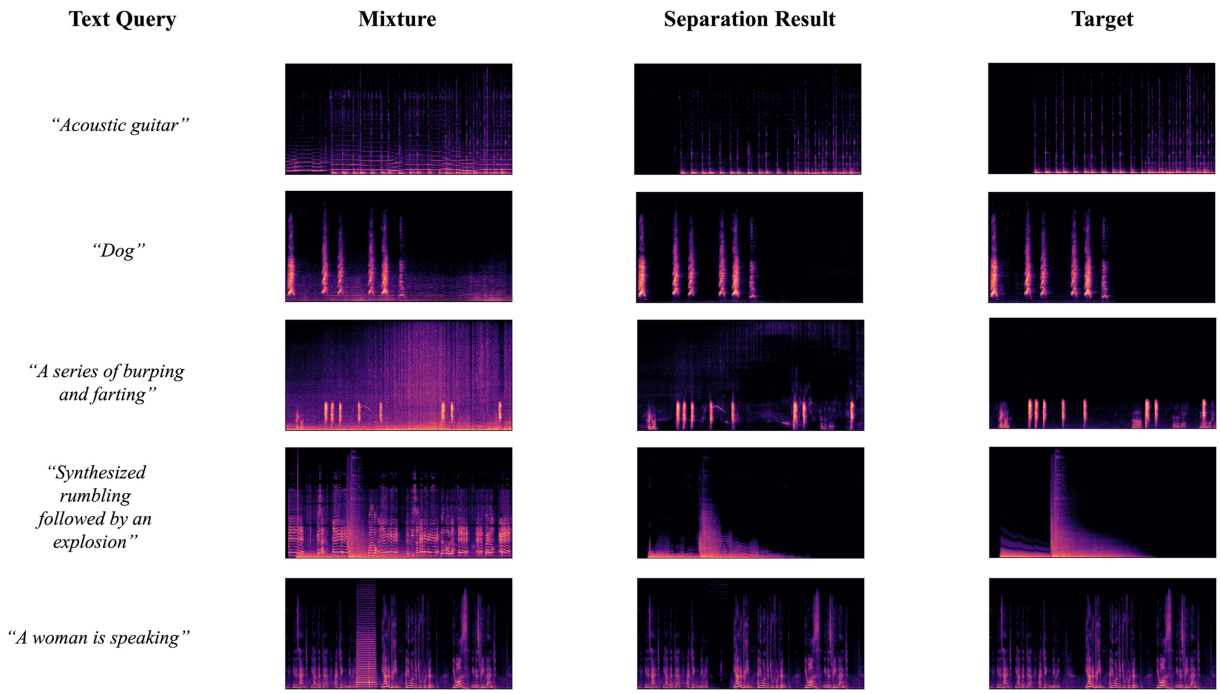


Fig. 2. Visualization of separation results obtained by AudioSep.

TABLE VI
SPEECH ENHANCEMENT RESULTS ON VOICEBANK-DEMAND DATASET

	PESQ	CSIG	CBAK	COVL	SSNR
Noisy	1.97	3.35	2.44	2.63	1.68
Wiener [69]	2.22	3.23	2.68	2.67	5.07
SEGAN [6]	2.16	3.48	2.94	2.80	7.73
AudioSet-UNet [7]	2.28	2.43	2.96	2.30	8.75
Wave-U-Net [70]	2.40	3.52	3.24	2.96	9.97
CMGAN [77]	3.41	4.63	3.94	4.12	11.10
MP-SENet [78]	3.50	4.73	3.95	4.22	10.64
AudioSep	2.43	3.37	3.17	2.88	9.21

noisy speech signals (e.g., AudioSet) in the training data. This noise can hinder the model's ability to effectively learn and separate clean speech signals, particularly in scenarios with an SNR range of 0 to 15 dB.

E. Visualization of Separation Results

We visualized spectrograms for audio mixtures, target audio sources, and separated sources using text queries of diverse synthetic mixtures (e.g., musical instruments, audio events, speech) with the AudioSep model, as shown in Fig. 2. The spectrogram patterns of the separated sources closely resemble those of the target sources, aligning with our objective experimental results. Additionally, we conducted several case studies using real audio clips sourcing from Freesound. The results demonstrated clear isolation of target audio components, highlighting the model's robustness and effectiveness. Audio samples are available on our project page[10].

[10][Online]. Available: https://audio-agi.github.io/Separate-Anything-You-Describe

VI. ABLATION STUDIES

A. Learning With Multimodal Supervision

Recent research has explored the potential of using multimodal supervision [23], [24], [25] to enhance the scalability of training LASS models. For example, Contrastive Language-Image Pre-training (CLIP) model is pre-trained on large-scale image-text paired data using contrastive learning, where its text encoder learns to map textual descriptions into the same semantic space as visual representations. The key advantage of using the CLIP text encoder for LASS is that it enables us to train or scale up the LASS model using large-scale unlabeled audio-visual data [41], [61], leveraging visual embeddings as a substitute for annotated audio-text paired data. However, these works mainly focused on small-scale training sets (e.g., VGGSound [61]). In this section, we present ablation studies to investigate the efficacy of using large-scale multimodal supervision with CLIP and CLAP models to scale up AudioSep. We aim to gain insights into the applicability and performance of leveraging large-scale multimodal supervision for LASS.

Following [23], we trained the proposed model using either CLIP or CLAP text encoders with a hyperparameter to control the training text rate (TR). The TR controls the percentage of training examples conditioned with text instead of audio or visual modality. For example, at 0% TR, the AudioSep model is trained exclusively with audio conditions from CLAP or visual conditions from CLIP. Conversely, at 100% TR, the model is trained solely with text conditions. Following [23], we used the 'ViT-B-32' checkpoint for the CLIP model. For video data processing, frames were uniformly extracted at one-second intervals, and their averaged CLIP embeddings were computed to serve as the query embeddings. Each model in this study was

TABLE VII
ABLATION STUDY OF SCALING UP AUDIOSEP WITH MULTIMODAL SUPERVISION

	AudioSet		VGGSound		AudioCaps		Clotho		MUSIC		ESC-50		DCASE 2024 T9	
	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi
AudioSep-CLIP-TR0.0	0.45	2.91	−3.84	0.78	−0.02	3.29	−1.50	2.42	−0.21	2.97	−0.24	3.67	−2.49	0.95
AudioSep-CLIP-TR0.5	6.48	7.28	7.01	7.27	5.84	7.25	4.32	6.10	7.40	9.05	8.74	9.97	3.68	5.18
AudioSep-CLIP-TR0.75	6.52	7.31	**7.46**	**7.67**	5.98	7.41	4.38	6.16	8.00	9.35	8.92	10.10	3.94	5.52
AudioSep-CLIP-TR1.0	**6.60**	**7.37**	7.24	7.50	5.95	7.45	4.54	6.28	**9.14**	**10.45**	8.90	10.03	3.96	5.46
AudioSep-CLAP-TR0.0	−0.34	3.32	−1.64	2.96	2.12	5.09	0.63	4.22	4.42	6.93	1.24	5.68	0.65	3.77
AudioSep-CLAP-TR0.5	5.94	6.88	7.04	7.24	6.31	7.62	4.19	6.13	8.16	9.65	8.36	9.63	3.92	5.38
AudioSep-CLAP-TR0.75	5.94	6.90	7.20	7.39	6.28	7.60	4.29	6.16	8.42	9.80	8.83	10.03	3.65	5.27
AudioSep-CLAP-TR1.0	**6.58**	**7.30**	**7.38**	**7.55**	**6.45**	**7.68**	**4.84**	**6.51**	**8.45**	**9.75**	**9.16**	**10.24**	**4.47**	**5.86**

trained for 1M steps on 8 Tesla V100 GPU cards. The model training applied a data augmentation method from [15], which first augments the audio clips with the same energy level before mixing them together to create the mixture.

For the CLIP-based models, we experiment with TRs of 0%, 50%, 75%, and 100%. The resulting models are referred to as AudioSep-CLIP-TR0.0, AudioSep-CLIP-TR0.5, AudioSep-CLIP-TR0.75, and AudioSep-CLIP-TR1.0, respectively. The visual condition is exclusively adopted for the training using AudioSet [41] and VGGSound [61] datasets that have video stream information. For datasets that solely consist of audio and text, such as WavCaps [63], we use text condition. This training configuration allows us to leverage the available modalities of each dataset. Experimental results are shown in the upper part of Table VII. When training AudioSep-CLIP-TR0.0 without text supervision from the AudioSet and VGGsound datasets, we observed clearly inadequate performance across all evaluation datasets. As large-scale video data may contain irrelevant audio contexts, this experimental result highlights the significance of text supervision from audio event labels provided by AudioSet and VGGSound in effectively training the LASS model. When training using text ratios of 50% and 75%, the overall performance is comparable to that of AudioSep-CLIP-TR1.0, which is trained exclusively with text supervision. This finding suggests that additional supervision from the visual modality does not improve separation performance in large-scale training settings, likely due to the inherent noise present in large-scale video data.

For the CLAP-based models, we use TRs of 0%, 50%, 75%, and 100%. The resulting models are denoted as AudioSep-CLAP-TR0.0, AudioSep-CLAP-TR0.5, AudioSep-CLAP-TR0.75, and AudioSep-CLAP-TR1.0. Experimental results are shown in the bottom part of Table VII. At 0% TR, where only audio modalities are used without text conditioning, the model underperforms, showing negative SI-SDR values on datasets like AudioSet and VGGSound, with only slightly better but still inadequate performance on others. For example, it achieves an SDRi of 4.22 dB on Clotho and 6.93 dB on MUSIC. This limited performance indicates that audio-only conditioning struggles in separating sources when text conditions are used during inference. However, as text supervision increases to 50% and 75% TR, the model's performance improves substantially, indicating the important role of text conditioning during training. The separation performance continues to improve up to the TR1.0 setting, where the model is trained exclusively with text, achieving optimal results across all CLAP variants and demonstrating that text conditioning is crucial for effective representation learning in audio separation tasks. Overall, we observed that using additional audio supervision to scale AudioSep did not lead to improvements in our experiments, likely due to the *modality gap* phenomenon [81]: the embedding spaces in the CLAP models may not be well aligned. A recent study [82] demonstrated that it is possible to further improve the performance of AudioSep-CLAP model by including audio modality as supervision. However, this approach requires embedding manipulation techniques, such as applying dropout to the audio embeddings or adding Gaussian noise. While these techniques provide some improvement, the gains are not significant, we leave further exploration of this approach for future work.

Comparing the optimal results of CLIP-based and CLAP-based models, we observed that the CLAP-based model has better performance on audio datasets with natural language queries, such as AudioCaps, Clotho, and DCASE 2024 T9. This suggests that CLAP models may be more adept at learning representations from natural language queries. Both CLIP-based and CLAP-based models show comparable results on datasets where simple text labels serve as queries, including AudioSet, VGGSound, and ESC-50. This indicates that both models are equally effective when dealing with straightforward textual information (e.g., text labels). The CLIP model performs better on the MUSIC dataset, likely because its text embeddings capture distinctions between musical instruments more effectively than those of the CLAP model. Although CLIP was not trained on instrumental audio data, it was trained on a potentially vast amount of music-related image-text data, enabling it to form clear distinctions in its text embedding space and achieve good performance in text-queried music instrument separation tasks. This observation is consistent with findings from the CLIPSep [23] experiments.

B. Studies of Using Various Text Queries

In practice, human descriptions of an audio source are generally personalized, which poses a challenge to develop LASS models that can handle a variety of natural language queries well. In this section, we conduct an ablation study to investigate how does our proposed model perform when using various text queries. Specifically, we randomly selected 50 audio clips from the test set of the AudioCaps dataset and engaged four English native speakers from the University of Surrey to individually re-annotate the selected clips. Each annotator provided a single description per audio clip without any specific hints or restrictions. Consequently, we obtained a set of six descriptions for each audio clip. This set included one caption sourced from

TABLE VIII
EXPERIMENTAL RESULTS OF COMPARISON OF DIFFERENT TEXT QUERIES ON AUDIOCAPS-MINI DATASET

	SI-SDR	SDRi
AudioSep-CLIP (text label)	6.39	7.70
AudioSep-CLIP (original caption)	7.27	8.25
AudioSep-CLIP (re-annotated caption)	6.44	7.75
AudioSep-CLAP (text label)	6.32	7.65
AudioSep-CLAP (original caption)	7.73	8.47
AudioSep-CLAP (re-annotated caption)	6.59	7.80

The suffix of TR1.0 is ignored.

TABLE IX
AN EXAMPLE FROM AUDIOCAPS-MINI

	Text Description
Text label	"Vibration"
Original caption	"Clicking followed by vibrations"
Re-annotated caption1	"Gear change, moving vehicle"
Re-annotated caption2	"The distant engine sound"
Re-annotated caption3	"The engine is running in distant"
Re-annotated caption4	"The engine is starting"

AudioCaps, four captions provided by the annotators, and the AudioSet audio event text labels. To create test mixtures, each audio clip was mixed with ten background sources randomly selected with an SNR at 0 dB, considering that the sound event labels of the background source do not overlap with that of the target source. We denote this test dataset as AudioCaps-Mini.

We evaluate the performance of AudioSep-CLIP-TR1.0 and AudioSep-CLAP-TR1.0 on AudioCaps-Mini. To investigate the effects of various text queries, we utilize the retrieved AudioSet event labels, the original AudioCaps audio captions, and our re-annotated natural language descriptions, which are referred to as the "text label", "original caption" and "re-annotated caption", respectively. An example of AudioCaps-Mini can be found in Table IX. Experimental results are shown in Table VIII. For both the CLIP-based and CLAP-based models, we have observed a considerable performance improvement when using the "original caption" as text queries instead of using the "text label". This may be attributed to the fact that human-annotated captions provide more comprehensive and accurate descriptions of the source of interest compared to audio event labels. Despite the personalized nature and different word distribution of our re-annotated captions, the results obtained using the "re-annotated caption" are slightly worse than those using the "original caption", while still marginally outperforming the results obtained with the "text label". These experimental findings demonstrate the promising generalization performance and robustness of AudioSep in real-world cases with diverse text queries.

We further investigated how the model's performance degrades when using reannotated queries compared to the original queries. We observed several instances where the model's performance degrades slightly as distortion increases. An example could be found in Fig. 3 .

C. Studies of Using Invalid Text Queries

In this section, we perform an ablation study to investigate the effect of using invalid queries, i.e., queries that do not match any sources presented in the audio mixtures, on the performance of audio source separation. We used three different types of invalid queries: one randomly sampled caption from the AudioCaps test set, another from the ESC-50 event class labels, and a DIY text query labeled "Separate Anything You Describe." Examples of these different invalid queries are presented in Table X. We conducted this ablation study on DCASE 2024 Task 9 evaluation set and using the pre-trained AudioSep model.

The separation results are presented in Table XI. For all three types of invalid queries, the SI-SDR values are negative. Through manual inspection, we observed that the AudioSep model tends to suppress some signals randomly when presented

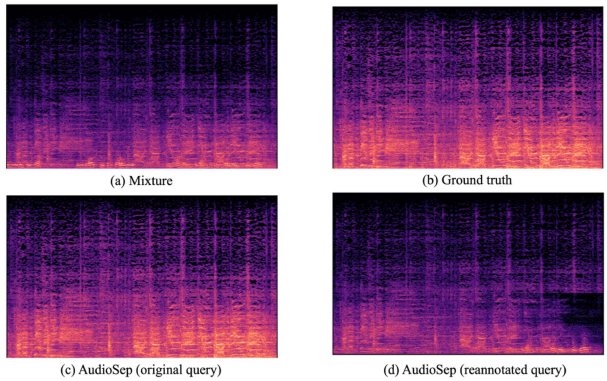


Fig. 3. Case studies of (a) an audio mixture; (b) the ground truth target source; (c) the separated source queried by AudioCaps's original caption: *"People laugh followed by people singing while music plays"*; (d) the separated source queried by our reannotated caption: *"A music show is presenting to the public"*. Results are obtained using the AudioSep-CLAP-TR1.0 model.

TABLE X
EXAMPLES OF DIFFERENT INVALID TEXT QUERIES

	Text Description
DIY	"Separate Anything You Describe"
ESC-50	"Church bells"
AudioCaps	"Clicking followed by vibrations"
Ground truth	"A man is laughing and then he says something."

TABLE XI
EXPERIMENTAL RESULTS OF USING DIFFERENT INVALID TEXT QUERIES ON DCASE 2024 T9 DATASET

Negative Caption	SI-SDR	SDRi
DIY	−6.56	1.79
ESC-50	−7.68	1.48
AudioCaps	−8.51	1.35

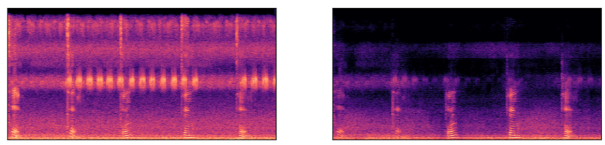


Fig. 4. An example of performing separation using AudioSep with an invalid query: "Separate Anything You Describe".

with invalid queries. An example can be found in Fig. 4. As potential optimizations, we could consider supervising the model to generate silence in response to invalid queries or leveraging novel methods proposed in the OCT [45]. We leave these for our future work.

VII. Conclusion and Future Work

We have introduced AudioSep, a foundation model for open-domain universal sound separation with natural language descriptions. AudioSep can perform zero-shot separation using text labels or audio captions as queries. We have presented a comprehensive evaluation benchmark including numerous sound separation tasks such as audio event separation, musical instrument separation, and speech enhancement. AudioSep outperforms state-of-the-art text-queried separation systems and off-the-shelf audio-queried sound separation models. We show that AudioSep is a promising approach to flexibly address the CASA problem with strong sound separation performance. In future work, we will improve the separation performance of AudioSep via unsupervised learning techniques [14], [35] and extend AudioSep to support audio-visual sound separation, audio-queried sound separation, and text-guided speaker separation [83] tasks.

Acknowledgment

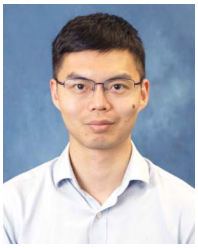
The authors would like to thank the reviewers and associate editor for their constructive comments to further improve the paper. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC-BY) license to any Author Accepted Manuscript version arising.

References

[1] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.
[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comput.*, vol. 17, no. 9, pp. 1875–1902, 2005.
[3] X. Liu et al., "Separate what you describe: Language-queried audio source separation," in *Proc. Interspeech*, 2022, pp. 1801–1805.
[4] I. Kavalerov et al., "Universal sound separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 175–179.
[5] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep ResUNet for music source separation," in *Proc. ISMIR*, 2021, pp. 342–349.
[6] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.
[7] Q. Kong, H. Liu, X. Du, L. Chen, R. Xia, and Y. Wang, "Speech enhancement with weakly labelled data from AudioSet," in *Proc. Interspeech*, 2021, pp. 191–195.
[8] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
[10] S. Rubin, F. Berthouzoz, G. J. Mysore, W. Li, and M. Agrawala, "Content-based tools for editing audio stories," in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, 2013, pp. 113–122.
[11] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.
[12] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 570–586.
[13] M. Chatterjee, J. Le Roux, N. Ahuja, and A. Cherian, "Visual scene graphs for audio source separation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1204–1213.
[14] E. Tzinis et al., "Into the wild with AudioScope: Unsupervised audio-visual separation of on-screen sounds," in *Proc. Int. Conf. Learn. Representations*, 2021.
[15] Q. Kong et al., "Universal source separation with weakly labelled data," 2023, *arXiv:2305.07447*.
[16] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Zero-shot audio source separation through query-based learning from weakly-labeled data," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 4, pp. 4441–4449.
[17] B. Gfeller, D. Roblek, and M. Tagliasacchi, "One-shot conditional audio filtering of arbitrary sounds," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 501–505.
[18] Y. Wang, D. Stoller, R. M. Bittner, and J. P. Bello, "Few-shot musical source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 121–125.
[19] B. Veluri, J. Chan, M. Itani, T. Chen, T. Yoshioka, and S. Gollakota, "Real-time target sound extraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
[20] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, Y. Ohishi, and S. Araki, "SoundBeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 121–136, 2023.
[21] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to what you want: Neural network-based universal sound selector," in *Proc. Interspeech*, 2020, pp. 1441–1445.
[22] H. Wang, D. Yang, C. Weng, J. Yu, and Y. Zou, "Improving target sound extraction with timestamp information," in *Proc. Interspeech*, 2022, pp. 1526–1530.
[23] H.-W. Dong, N. Takahashi, Y. Mitsufuji, J. McAuley, and T. Berg-Kirkpatrick, "CLIPSep: Learning text-queried sound separation with noisy unlabeled videos," in *Proc. Int. Conf. Learn. Representations*, 2023.
[24] K. Kilgour, B. Gfeller, Q. Huang, A. Jansen, S. Wisdom, and M. Tagliasacchi, "Text-driven separation of arbitrary sounds," in *Proc. Interspeeh*, 2022, pp. 5403–5407.
[25] R. Tan et al., "Language-guided audio-visual source separation via trimodal consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10575–10584.
[26] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.
[27] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proc. 19th Int. Soc. Music Inf. Retriev. Conf.*, 2019, pp. 334–340.
[28] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "DEMUCS: Deep extractor for music sources with extra unlabeled data remixed," 2019, *arXiv:1909.01174*.
[29] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.
[30] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7092–7096.
[31] L. Wan, H. Liu, Y. Zhou, and J. Jia, "Multi-loss convolutional network with time-frequency attention for speech enhancement," in *Proc. Int. Conf. Inf. Commun. Signal Process.*, 2023, pp. 629–633.
[32] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 31–35.
[33] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 1–5.
[34] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 241–245.
[35] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 3846–3857.
[36] R. Gao and K. Grauman, "VisualVoice: Audio-visual speech separation with cross-modal consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15490–15500.
[37] J. Wu et al., "Time domain audio visual speech separation," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 667–673.

[38] J. Zhao et al., "Universal sound separation with self-supervised audio masked autoencoder," in *Proc. 32nd Eur. Signal Process. Conf.*, 2024, pp. 1–5.
[39] Y. Liu et al., "Audio prompt tuning for universal sound separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 1446–1450.
[40] J. H. Lee, H.-S. Choi, and K. Lee, "Audio query-based music source separation," in *Proc. 20th Int. Soc. Music Inf. Retriev. Conf.*, 2019, pp. 878–885.
[41] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 776–780.
[42] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.
[43] Y. Xiao et al., "Continual learning for on-device environmental sound classification," in *Proc. 7th Detection Classification Acoust. Scenes Events Workshop*, 2022.
[44] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, vol. 1, pp. 119–132.
[45] E. Tzinis, G. Wichern, P. Smaragdis, and J. Le Roux, "Optimal condition training for target source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2019, pp. 4171–4186.
[47] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[48] X. Mei, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Audio captioning transformer," in *Proc. 6th Detection Classification Acoust. Scenes Events Workshop*, 2021, pp. 211–215.
[49] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust.*, 2017, pp. 374–378.
[50] X. Mei, X. Liu, J. Sun, M. Plumbley, and W. Wang, "On metric learning for audio-text cross-modal retrieval," in *Proc. Interspeech*, 2022, pp. 4142–4146.
[51] A.-M. Oncescu, A. Koepke, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries," in *Proc. Interspeech*, 2021, pp. 2411–2415.
[52] H. Liu et al., "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023.
[53] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction-tuned LLM and latent diffusion model," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023.
[54] J. Huang et al., "Make-An-Audio 2: Temporal-enhanced text-to-audio generation," 2023, *arXiv:2305.18474*.
[55] D. Yang et al., "DiffSound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1720–1733, 2023.
[56] F. Kreuk et al., "AudioGen: Textually guided audio generation," in *Proc. Int. Conf. Learn. Representations*, 2023.
[57] X. Liu, T. Iqbal, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Conditional sound generation using neural discrete time-frequency representation learning," in *Proc. IEEE 31st Int. Workshop Mach. Learn. Signal Process.*, 2021, pp. 1–6.
[58] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
[59] J. Liang et al., "Adapting language-audio models as few-shot audio learners," in *Proc. Interspeeh*, 2023, pp. 276–280.
[60] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLm: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 3942–3951.
[61] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A large-scale audio-visual dataset," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 721–725.
[62] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 736–740.
[63] X. Mei et al., "WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," in *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2023, vol. 32, pp. 3339–3354.
[64] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," 2020, *arXiv:2005.11262*.
[65] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1015–1018.
[66] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 829–852, 2022.
[67] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA Conf. Asian Spoken Lang. Res. Eval.*, 2013, pp. 1–4.
[68] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: A collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. Meetings Acoust.*, 2013, pp. 1–6.
[69] P. Scalart et al., "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1996, vol. 2, pp. 629–632.
[70] C. Macartney and T. Weyde, "Improved speech enhancement with the Wave-U-Net," 2018, *arXiv:1811.11307*.
[71] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.
[72] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2001, pp. 749–752.
[73] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[74] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.
[75] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Munich, Germany, Oct. 2015, pp. 234–241.
[76] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 646–650.
[77] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric GAN for speech enhancement," in *Proc. Interspeech*, 2022, pp. 936–940.
[78] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra," in *Proc. Interspeech*, 2023, pp. 3834–3838.
[79] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[80] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGaN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2031–2041.
[81] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 17612–17625.
[82] K. Saijo, J. Ebbers, F. G. Germain, S. Khurana, G. Wichern, and J. L. Roux, "Leveraging audio-only data for text-queried target sound extraction," 2024, *arXiv:2409.13152*.
[83] X. Hao, J. Wu, J. Yu, C. Xu, and K. C. Tan, "Typing to listen at the cocktail party: Text-guided target speaker extraction," 2023, *arXiv:2310.07284*.

Xubo Liu (Member, IEEE) received the B.Eng. degree from Queen Mary University of London, London, U.K., in 2020. He is currently working toward the Ph.D. degree with the Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford, U.K. His Ph.D. degree is on multimodal learning for audio and language, focusing on the understanding, separation, and generation of audio signals in tandem with natural language. He has coauthored over 40 papers in top conferences such as CVPR, ICML, AAAI, EMNLP, ICASSP, Interspeech. He organized the "Multimodal Learning for Audio and Language" special session at EUSIPCO 2023. He organized the "Generative AI for Media Generation" special session with MLSP 2024. He led the organization of the "Language-Queried Audio Source Separation" challenge on DCASE 2024. He is a frequent reviewer for TASLP, CVPR, ECCV, ICLR, ICML, ACL, EMNLP, ICASSP, Interspeech, and MLSP.



Qiuqiang Kong received the Ph.D. degree from the University of Surrey, Guildford, U.K., in 2019. He is currently an Assistant Professor with the Electronic Engineering Department of the Chinese University of Hong Kong, Hong Kong. Following the Ph.D., he joined ByteDance as a research scientist. He has coauthored over 50 papers in journals and conferences, including IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, ICASSP, INTERSPEECH, IJCAI, DCASE, EUSIPCO, LVA-ICA.His research topic includes the classification, detection, separation, and generation of general sounds and music. He was the top 2% scientist in 2021 in "Updated science-wide author databases of standardized citation indicators. He was known for developing pretrained audio neural networks for audio tagging and was awarded the IEEE SPS Young Author Best Paper in 2023. He won the detection and classification of acoustic scenes and events challenge in 2017. He was known for transcribing the largest piano MIDI dataset GiantMIDI-Piano in the world.

Yan Zhao received the M.S. degree in electrical and computer engineering from The Ohio State University, Columbus, OH, USA, in 2014, and the M.S. and a Ph.D. degrees in computer science and engineering from The Ohio State University, Columbus, OH, USA, in 2018 and 2020, respectively. He is currently a Research Scientist with ByteDance Inc., San Jose, CA, USA. His research interests include speech and audio processing, with specific interests in separation and enhancement, and he is also with speech/audio LLMs.

Haohe Liu received the B.Eng. degree from Northwestern Polytechnical University, Xi'an, China, in 2020. He is currently working toward the Ph.D. degree with the Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, U.K. His past research has contributed to the fields of audio quality enhancement, audio generation, source separation, and audio recognition. Haohe is mostly known for developing AudioLDM for textto-audio generation, generating wide attention in the open-source community. His work as the primary author has been accepted in top journals and conferences such as TPAMI, TASLP, JSTSP, ICML, AAAI, ICASSP, and INTERSPEECH. Notable recent projects include AudioLDM, VoiceFixer, AudioSR, and NaturalSpeech.

Yi Yuan (Student Member, IEEE) received the B.Eng. degree from the University of Sydney, Sydney, Australia, and the Harbin Engineering University, Harbin, China, in 2021, and the M.S. degree in Artificial Intelligence from the University of Surrey, Guildford, U.K., in 2022. He is currently working toward the Ph.D. degree in Vision, Speech and Signal Processing with the University of Surrey. His research interests include on the area of deep-learning-based audio generation. In 2023, he achieved the top-1 ranking with DCASE data challenge task 7.

Yuzhuo Liu received the bachelor's degree in information security from Harbin Engineering University, Harbin, China, and the Ph.D. degree in signal and information processing from The Institute of Acoustics, Chinese Academy of Sciences, Beijing, China. She is currently working with ByteDance. Her research interests are in the areas of paralinguistic understanding, audio captioning, multimodal large language models, and audio source separation.

Rui Xia, biography not available at the time of publication.

Yuxuan Wang, biography not available at the time of publication.

Mark D. Plumbley (Fellow, IEEE) received the B.A.(Hons.) degree in electrical sciences and the Ph.D. degree in neural networks from University of Cambridge, Cambridge, U.K., in 1984 and 1991, respectively. He is currently a Professor of Signal Processing with the Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford, U.K. His research interests include concerns AI, machine learning and signal processing for analysis, recognition and generation of sound. He led the first international data challenge on Detection and Classification of Acoustic Scenes and Events, and currently holds an Engineering and Physical Sciences Research Council Fellowship on "AI for Sound". He is a Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, and a Fellow of the IET .

Wenwu Wang (Senior Member, IEEE) was born in Anhui, China. He received the B.Sc., M.E., and Ph.D. degrees in from Harbin Engineering University, Harbin, China, in 1997, 2000, and 2002, respectively. In 2007 he was then with King's College London, Cardiff University, Tao Group Ltd. (now Antix Labs Ltd.), and Creative Labs, before joining University of Surrey, U.K., where he is currently a Professor in Signal Processing and Machine Learning, and a Co-Director of the Machine Audition Lab within the Centre for Vision Speech and Signal Processing. He is also an AI Fellow within the Surrey Institute for People Centred Artificial Intelligence. He has coauthored over 300 publications in these areas. He is a coauthor or corecipient of over 15 awards including the 2022 IEEE SPS Young Author Best Paper Award, ICAUS 2021 Best Paper Award, DCASE 2020, 2023 and 2024 Judges' Award, DCASE 2019 and 2020 Reproducible System Award, and LVA/ICA 2018 Best Student Paper Award. He is an Associate Editor from 2020 to 2025 for IEEE/ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING, and an Associate Editor from 2024 to 2026 for IEEE TRANSACTIONS ON MULTIMEDIA. He was a Senior Area Editor from 2019 to 2023 and an Associate Editor from 2014 to 2018 for IEEE TRANSACTIONS ON SIGNAL PROCESSING. His research interests include blind signal processing, sparse signal processing, audio-visual signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. He is the elected Chair from 2023 to 2024 of IEEE SPS Machine Learning for Signal Processing Technical Committee, elected Chair from 2025 to 2027 and Vice Chair from 2022 to 2024 of EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, an elected Member from 2021 to 2026 of the IEEE Signal Processing Theory and Methods Technical Committee, and an elected Member from 2019 of the International Steering Committee of Latent Variable Analysis and Signal Separation. He was a Member of the organisation committee of IEEE ICASSP 2019 and 2024, INTERSPEECH 2022, IEEE MLSP 2013, 2024, and 2025, and a Technical/Program Committee Member of over 100 international conferences.