OpenSep: Leveraging Large Language Models with Textual Inversion for Open World Audio Separation

Tanvir Mahmud, Diana Marculescu
Chandra Family Department of Electrical and Computer Engineering
The University of Texas at Austin
tanvirmahmud@utexas.edu, dianam@utexas.edu

Abstract

Audio separation in real-world scenarios, where mixtures contain a variable number of sources, presents significant challenges due to limitations of existing models, such as over-separation, under-separation, and dependence on predefined training sources. We propose OpenSep, a novel framework that leverages large language models (LLMs) for automated audio separation, eliminating the need for manual intervention and overcoming source limitations. OpenSep uses textual inversion to generate captions from audio mixtures with *off-the-shelf* audio captioning models, effectively parsing the sound sources present. It then employs few-shot LLM prompting to extract detailed audio properties of each parsed source, facilitating separation in unseen mixtures. Additionally, we introduce a multi-level extension of the *mix-and-separate* training framework to enhance modality alignment by separating single source sounds and mixtures simultaneously. Extensive experiments demonstrate OpenSep's superiority in precisely separating new, unseen, and variable sources in challenging mixtures, outperforming SOTA baseline methods. Code will be released at https://github.com/tanvir-utexas/OpenSep.git.

1 Introduction

Audio and music mostly appear in real-world mixtures containing various audio sources and background noise (Kim et al., 2019). Separating clean sources from noisy mixtures have numerous applications in audio processing (Liu and Wang, 2018; Stöter et al., 2019). Precise separation of clean sound sources require their complete semantic understanding, as well as extensive training on natural mixtures. However, in open world scenarios, audio mixtures may contain a variable number of sources, as well as new, unseen, and possibly noisy sources. Hence, gathering clean sounds from a variety of sources, and modeling exhaustive mixture combinations for training are often impractical, which limits the use of audio separators in practice.

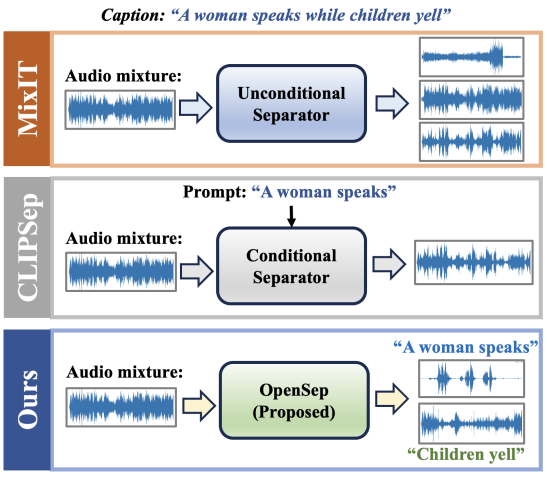

Figure 1: Unconditional audio separators suffer from both over-separation and under-separation in noisy mixtures, and cannot parse audio entities without additional classifiers. Furthermore, conditional separators rely on manual text prompts for source separation, limiting their use in practice. In contrast, OpenSep fully automates the source parsing and separation flow, even with varying number of unseen and noisy sources in open world.

Prior work on audio separation mostly focused on two approaches: unconditional and conditional source separation. Unconditional separators (Yu et al., 2017; Wisdom et al., 2020) mostly attempt to disentangle the mixture into a fixed set of output predictions, and later rely on post-processing to select/process the separated sources. However, this approach is largely limited to training sources and mixture combinations, which, in practice, results in over-/under-separation (Karamatlı and Kırbız, 2022). In addition, unconditional separators cannot provide the corresponding class entities of separated sources. Pishdadian et al. (2020) introduced source classifiers with audio separators, but their method is limited by training sources only.

Conditional source separation simplifies the problem by relying on conditional prompts from

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 13244–13260
November 12-16, 2024

other modalities/signals, such as text, vision, and sometimes, clean audio representing the target sound (Zhao et al., 2018; Mahmud et al., 2024). However, these methods are limited by the availability of target prompts for separation, which are often difficult to gather in practice. Moreover, the conditioning signal usually contains simple class instances as prompts which cannot explicitly describe the target audio source. Hence, these conditional audio separators are mostly limited to seen sources, often under-performing on unseen, noisy source mixtures in open world.

To overcome these limitations, we introduce OpenSep, a novel framework to automatically separate and parse unseen audio from noisy mixtures with a variable number of sources in open world (Fig 1). In particular, we propose textual inversion by using an *off-the-shelf* audio captioning model to extract text representations of noisy mixtures. We also leverage the world knowledge of audio sources in large language models (LLMs) to automatically parse, disentangle, and extract enriched representations of audio properties of each source present in the mixture. Finally, we train a text-conditioned audio separator to extract audio sources from noisy mixtures. For enhancing the modality alignment between conditional prompts and separated sources, we propose a multi-order separation objective by extending the baseline *mix-and-separate* framework. Extensive experiments demonstrates significant performance improvements of OpenSep over prior work, *e.g.*, OpenSep achieves $+64\%$ and $+180\%$ SDR improvements on unseen sources in MUSIC and VGGSound datasets, respectively, over baseline state-of-the-art models.

Our contributions are summarized as follows:

1. Our work is the first to introduce knowledge parsing from large language models for open-world audio separation.

2. OpenSep fully automates the source separation and recognition pipeline from noisy, unseen mixtures without manual intervention.

3. We propose a multi-level extension of *mix-and-separate* training framework to enhance audio and text modality alignment.

4. Extensive experiments on three benchmark datasets show the superiority of OpenSep over existing state-of-the-art methods.

2 Related Work

Unconditional Sound Separation Prior work on unconditional audio separation in speech and music mostly relies on post-processing methods to pick the target sound (Stöter et al., 2019; Sawata et al., 2021; Takahashi and Mitsufuji, 2021; Wang and Chen, 2018; Yu et al., 2017; Zhang et al., 2021; Luo and Mesgarani, 2018). Later, permutation invariant training (PIT) (Yu et al., 2017; Kavalerov et al., 2019), followed by its variant mixture invariant training (MixIT) (Karamatlı and Kırbız, 2022; Wisdom et al., 2020, 2021) relies on permutation alignment on source prediction for performance enhancement. However, these methods suffer from both over- and under-separation, due to training distribution misalignment in open world scenarios. Later, weakly supervised training with classifiers has been explored (Pishdadian et al., 2020; Tzinis et al., 2020), however, such methods are limited in their use on a fixed number of training sources. In contrast, OpenSep attempts to separate a variable number of sources in open world, without being limited to certain training sources.

Conditional Sound Separation Prior work on conditional sound separation used visual guidance (Gao and Grauman, 2019; Zhao et al., 2018; Tian et al., 2021; Chatterjee et al., 2021; Lu et al., 2018), text guidance (Dong et al., 2022; Liu et al., 2022; Kilgour et al., 2022; Tan et al., 2023; Liu et al., 2023a; Mahmud et al., 2024), and clean audio source guidance (Chen et al., 2022; Gfeller et al., 2021) for conditioning on noisy mixtures. Most of these methods mostly rely on a *mix-and-separate* framework (Zhao et al., 2018). However, the requirement for users to explicitly specify which sources to separate is often impractical in dynamic or complex audio scenes. Moreover, in general, these methods struggle with unseen, new sources for learning the conditional guidance with specific class prompts. In contrast, OpenSep attempts to fully automate the separation of all sources present in noisy, unseen source mixtures in open world, without using hand crafted prompts.

3 Methodology

OpenSep addresses two critical challenges of audio separation: handling a variable number of sources without manual intervention and enhancing performance on unseen sources during inference. As illustrated in Fig 2, the OpenSep architecture com-

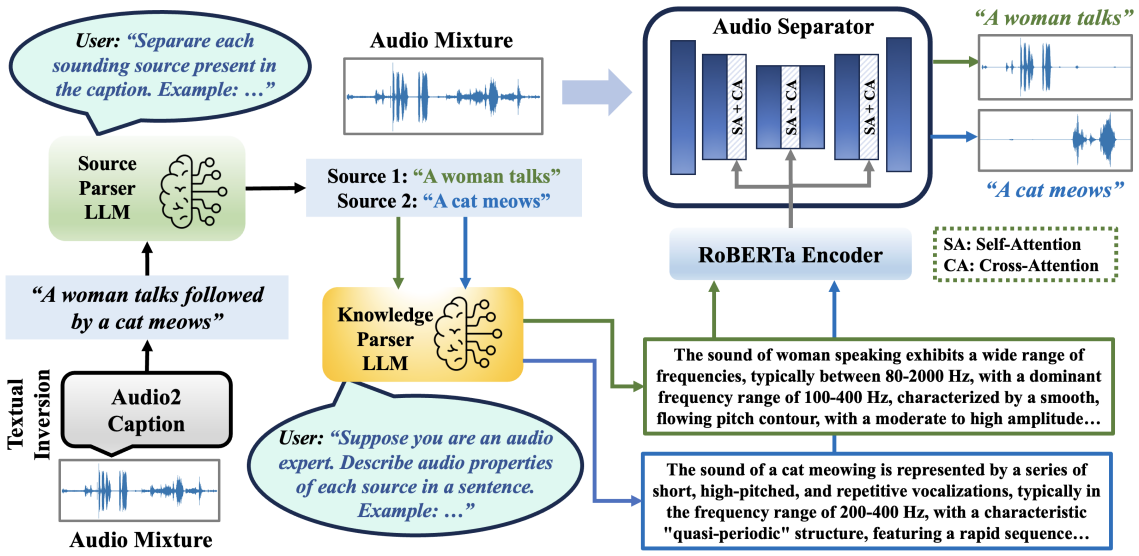

Figure 2: Proposed OpenSep pipeline: We initially apply textual inversion on noisy audio mixtures with an *off-the-self* audio captioning model to extract text descriptions. Afterwards, a pre-trained instruction-tuned LLM is used to parse audio sources from the caption, followed by the extraction of detailed audio properties of each source. Finally, a text-conditional audio-separator is used for separating each audio source from the noisy mixture using the enriched text prompts. Here, the audio separator is trained for leveraging detailed audio properties in textual representation.

prises of three key components. First, to overcome the training and test distribution misalignment of the unconditional separator in real-world mixtures, while also eliminating the need for manual prompts in conditional source separators, OpenSep leverages textual inversion to convert the audio mixture into a text representation. This step addresses the challenge of detecting a variable number of sources, while eliminating the human intervention to apply text conditions. Following that, we introduce knowledge parsing for each sounding source from large language models. We propose few-shot prompting with instruction-tuned LLM for parsing various sound sources from the caption, followed by the extraction of detailed audio properties of each sound source in text representation. This step is important for performing complex audio mixture separation in open-world scenarios on new, unseen sources. Finally, OpenSep uses a text-conditional audio-separator to extract the target sources based on the LLM-parsed text description. Unlike prior methods, OpenSep relies on detailed LLM-parsed knowledge to gather the context of each sound source for generalization, without being overfitted to training sources, which makes it suitable for real-world applications.

The following sections detail each component of the OpenSep framework to achieve robust and scalable audio separation.

3.1 Source Parsing with Textual Inversion

The audio source separation task can be split into two key phases: (1) detecting the sound sources present in the mixture, and (2) separating each source correctly. Prior work on unconditional separators attempts to perform both tasks simultaneously. This approach is challenging since it attempts to align the data distribution on varying number of sources, often resulting in over/under separation. Conditional separators achieve superior performance by eliminating the source recognition phase, but they rely heavily on manual prompts, which limits their applicability to automatically parsing the sources present in the mixture. OpenSep solves this complex mixture disentanglement challenge in open world by using textual inversion and leveraging the world knowledge in large language models.

Textual Inversion: To simplify the source parsing in complex mixtures, we propose converting the audio mixture to a text representation by using an *off-the-shelf* audio captioning model. This model processes the audio input and generates a textual description or caption, which encapsulates the salient features and sources present in the mixture. For example, a caption might describe an audio mixture as "a person speaking with background music and occasional dog barks." This textual representation enables the subsequent use of LLMs to further

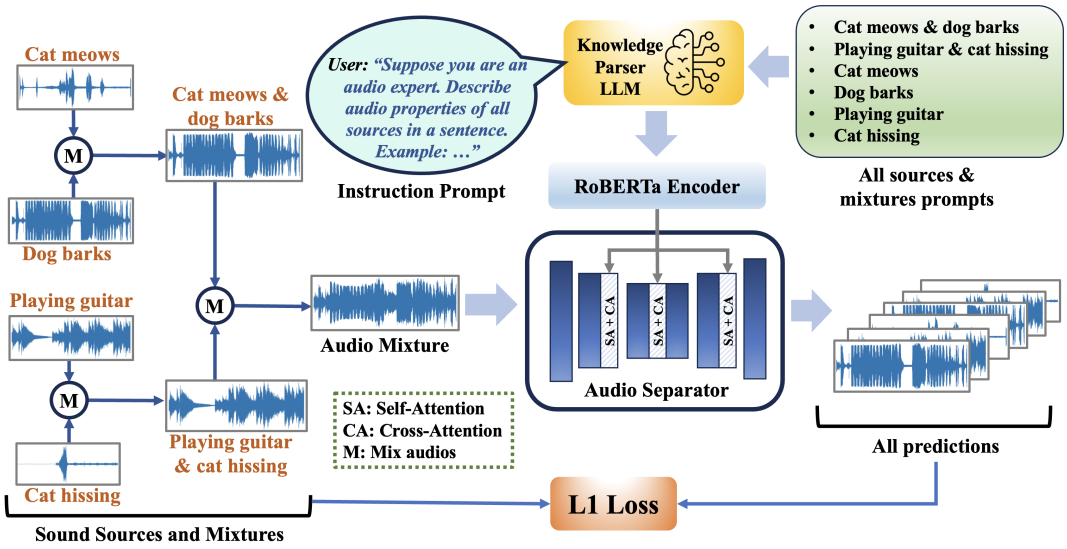

Figure 3: Proposed training pipeline: We extend the baseline *mix-and-separate* framework with multi-order separation objective for enhanced modality-alignment. Initially, we sample four independent single source sounds, and prepare synthetic mixtures of two and four sources. We parse enriched text prompts for mixtures and single-source sounds with a knowledge parser LLM. The audio separator is trained to separate both single-source and lower-order mixtures based on enriched text guidance using an L1 loss objective.

analyze and extract detailed information about the individual sound sources present in the audio.

Source parser LLM: Parsing audio sources requires semantic understanding of each source, which is difficult to train for the dynamic use cases in audio modality, limiting the practical use of audio separators. To parse each source, we propose to use a source parser LLM on the inverted text captions of the raw audio. By leveraging the semantic world knowledge of LLMs and few-shot prompting, OpenSep can precisely extract sound sources from mixtures. We use an instruction-tuned LLM, with the instruction for parsing individual sound sources present in the mixture. For example, with the input caption, *"Children yelling while a dog is barking in the background and a car horn honks afterwards"*, LLM splits each sources as *"Children yelling. Dog barking. Car horn honking."*

3.2 Knowledge Parsing for Each Source

To precisely separate each parsed source from complex mixtures, it is necessary to have more detailed knowledge of audio properties of the target source. Traditional conditional separators only rely on the class representation of each source. Despite their promising performance on seen classes used for training, these models underperform on unseen classes. To overcome this limitation, we propose the use of an instruction-tuned LLM as a knowledge parser to incorporate detailed audio properties of each source prior to the separation. This approach enables denser text alignment via separation of unseen sources in open-world scenarios. To facilitate the knowledge parsing from the LLM, we use specific instruction prompts to utilize the LLM as an audio expert providing the detailed audio properties of each source mentioned. Furthermore, we focus on several key properties of each sounding source for enriched representation, such as frequency range (pitch), amplitude (loudness), timbre (tone quality), usual duration, attack and decay, dynamic envelope characteristics, spectral content detailing the presence of harmonics, overtones, and the overall shape of the sound spectrum. Several hand-crafted high quality prompts are designed for several sources, and these are used for few-shot prompting with instruction guidance to leverage the pre-trained LLMs as an audio expert.

Rather than using a single class name for each sounding instance, we integrate richer details of each sounding source by leveraging knowledge parser LLMs, which in general, enhances audio separation performance by focusing on detailed audio feature guidance. Such denser audio-language alignment plays a pivotal role in separating unknown and noisy sources in real world.

3.3 Text-Conditioned Audio Separator

OpenSep uses a text-conditioned audio separator as a core building block given its superior performance over unconditional separators. The key dif-

ference of OpenSep over prior conditional separators is the automatic extraction of enriched audio features via modality inversion. Therefore, a longer context window is used for encoding the text prompts in the audio separator, to align the separator with enhanced text prompts. For the audio separator, we use the U-Net architecture as in prior work. For denser alignment between separated audio features and given text prompts, we use self-attention followed by cross-attention layers after each building block in the U-Net. The model initially converts the audio mixture into a magnitude spectrogram using short-term Fourier transform (STFT), and predicts the mask of the target sound given the text prompt using simple 2D convolutional kernels followed by self-attention and cross-attention. To reconstruct the sound, we use the filtered spectrogram with the phase residuals extracted from the original mixture, following prior work (Dong et al., 2022).

3.4 Proposed Training Pipeline

Most prior work relies on a *mix-and-separate* framework for conditional separator, which learns to separate single sources from synthetic audio mixtures given a conditional prompt. OpenSep primarily focuses on separating the target source by using richer text conditioning rather than a simple class prompt. The overall performance improvement on unseen and noisy sources mostly comes from the deeper audio and textual feature grounding on diverse audio properties. To achieve this, we propose a two-level separation training objective, extending the baseline *mix-and-separate* framework, such as mixture, and single-source separation.

As illustrated in Fig 3, we initially sample four single source audios (x_1, x_2, x_3, x_4) and prepare two synthetic mixtures (y_1, y_2) by mixing each pair of sources, given by $y_1 = \texttt{Mix}(x_1, x_2)$, and $y_2 = \texttt{Mix}(x_3, x_4)$. For mixing, we use amplitude re-scaling of each source, followed by simple addition on raw audio waveforms. Afterwards, we generate a higher order mixture z of four sources using $z = \texttt{Mix}(y_1, y_2)$. By leveraging the instruction tuned LLM on class entities of each source, we extract single source text prompts (S_1, S_2, S_3, S_4), and two-source mixture prompts (M_1, M_2). The model generates a series of predictions $\mathbf{P}$ of separated audios given the text prompt $\mathcal{T}$ and higher-order mixture z, where $\mathcal{T} \in \{S_1, S_2, S_3, S_4, M_1, M_2\}$. Finally, the L_1 loss between predicted magnitude spectrogram and source spectrogram is used as training objective.

The proposed multi-level extension of *mix-and-separate* training method facilitates the separation of lower-order audio mixtures as well as single source sounds from higher-order mixtures, following the detailed text feature guidance from the LLM. Hence, this method enhances the modality alignment between separated sounds and enriched LLM-generated text prompts, facilitating the separation on unseen and noisy sources.

4 Results

4.1 Evaluation Setup

Dataset: For the experiments on synthetic mixtures, we primarily use MUSIC (Zhao et al., 2018) and VGGSound (Chen et al., 2020) datasets. MUSIC dataset contains 21 musical instruments, separately played and recorded for $1 \sim 5$ minutes duration. VGGSound contains $162,433$ audio samples, mostly containing noisy single source sounds of 10s duration. For analyzing the performance on natural mixtures, we use AudioCaps (Kim et al., 2019) dataset containing $44,309$ audio mixtures having around $1 \sim 6$ sources with audio captions.

Implementation Details: We use LLaMA-3-8b (Touvron et al., 2023) language model for source and knowledge parsing. We use RoBERTa-Base (Liu et al., 2019) text encoder with a context window of 512 for encoding parsed LLM knowledge. For audio captioning, we use the CLAP-based captioning model (Elizalde et al., 2023), which combines an audio encoder with a GPT-2 text decoder. We use a U-Net based audio separator with self and cross-attention conditioning.

Training: All models are trained for 80 epochs with initial learning rate of 0.001. The learning rate is decreased by a factor of 0.1 every 20 epochs. An Adam optimizer is used with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The training was carried out with 8 RTX-A6000 GPUs with 48GB memory. For training, we randomly sample different sources, then, mix and separate using our training method. Also, the class label of each source is used for knowledge parsing during training.

Evaluation: We use signal-to-distortion ratio (SDR) and signal-to-interference ration (SIR) (Vincent et al., 2006) for evaluating different models. In general, SDR estimates overall separation quality of the target sound, which is widely used in prior

Table 1: Performance comparison on seen classes. All classes are used for training in each dataset. Baseline results are reproduced under the same setup for a fair comparison. Our method outperforms both conditional and unconditional baselines, without accessing the source prompts during evaluation.

Methods	MUSIC		VGGSound	
	SDR ↑	SIR ↑	SDR ↑	SIR ↑
PIT (Yu et al., 2017)	7.98 ± 0.27	10.81 ± 0.29	2.01 ± 0.32	4.13 ± 0.27
MixIT (Wisdom et al., 2020)	5.46 ± 0.42	7.39 ± 0.22	1.46 ± 0.19	3.56 ± 0.25
MixPIT (Karamatlı and Kırbız, 2022)	8.07 ± 0.25	11.01 ± 0.28	1.87 ± 0.26	3.95 ± 0.31
MixIT + PIT	8.13 ± 0.28	11.17 ± 0.23	2.14 ± 0.24	4.98 ± 0.41
CLIPSep (Dong et al., 2022)	8.33 ± 0.29	11.65 ± 0.26	2.24 ± 0.21	5.41 ± 0.23
WeakSup (Pishdadian et al., 2020)	6.49 ± 0.35	8.87 ± 0.31	1.73 ± 0.34	4.67 ± 0.29
LASSNet (Liu et al., 2022)	8.35 ± 0.32	11.83 ± 0.29	2.32 ± 0.29	5.95 ± 0.34
AudioSep (Liu et al., 2023b)	8.64 ± 0.31	12.18 ± 0.27	2.45 ± 0.23	6.14 ± 0.31
OpenSep (Ours)	**9.56** ± 0.28	**13.42** ± 0.26	**3.71** ± 0.18	**8.31** ± 0.19

Table 2: Performance comparison on unseen classes. All models are trained with 50% class samples, and the evaluation is carried out on remaining 50% class mixtures. OpenSep demonstrates significantly better performance on new, unseen classes compared to existing baselines.

Methods	MUSIC		VGGSound	
	SDR ↑	SIR ↑	SDR ↑	SIR ↑
PIT (Yu et al., 2017)	3.56 ± 0.29	4.97 ± 0.33	0.45 ± 0.29	1.54 ± 0.33
MixIT (Wisdom et al., 2020)	2.28 ± 0.35	3.45 ± 0.29	0.16 ± 0.33	0.97 ± 0.21
MixPIT (Karamatlı and Kırbız, 2022)	2.87 ± 0.26	4.11 ± 0.35	0.28 ± 0.31	1.13 ± 0.35
MixIT + PIT	3.98 ± 0.26	5.25 ± 0.24	0.66 ± 0.24	2.15 ± 0.34
CLIPSep (Dong et al., 2022)	4.97 ± 0.27	7.13 ± 0.24	1.08 ± 0.27	4.12 ± 0.29
WeakSup (Pishdadian et al., 2020)	-3.56 ± 0.47	-4.36 ± 0.37	-4.54 ± 0.52	-5.86 ± 0.89
LASSNet (Liu et al., 2022)	5.01 ± 0.23	7.38 ± 0.29	0.95 ± 0.26	3.89 ± 0.35
AudioSep (Liu et al., 2023b)	5.14 ± 0.24	7.56 ± 0.29	1.12 ± 0.42	4.45 ± 0.27
OpenSep (Ours)	**8.45** ± 0.32	**11.72** ± 0.35	**3.14** ± 0.31	**7.23** ± 0.39

work (Dong et al., 2022; Mahmud et al., 2024). SIR estimates the amount of interference in separation from other sources present in the mixture.

4.2 Main Comparisons

Baseline Models: We used single-source permutation invariant training, PIT (Yu et al., 2017) and multi-source MixIT (Wisdom et al., 2020, 2021) for unconditional baselines. Note that we use synthetic mixtures for MixIT training. We also combine MixIT and PIT to train on both mixtures and single source sounds. For conditional baselines, we use LASSNet (Liu et al., 2022), AudioSep (Liu et al., 2023b), and CLIPSep (Dong et al., 2022) models with text conditions, which are built using the *mix-and-separate* (Zhao et al., 2018) framework. For a fair comparison, all baseline results are reproduced with similar architecture and data split. For the conditional models, we explicitly provide the class labels as conditions. For the OpenSep and unconditional baselines, no external text conditions are used, and the best match between prediction and ground truth source is used for evaluation.

Comparison on Seen Classes: In Table 1, we provide a comparison on seen classes. For this analysis, we generate a synthetic test set of two-source mixtures with 10% samples from each class. All training was done separately on each dataset on all classes. We note that unconditional models suffer from over- and under-separation causing lower performance. In general, conditional models achieve superior performance over unconditional models by leveraging the user-defined class prompts. In VGGSound dataset, OpenSep improves by $+99\%$ SDR over MixPIT (Karamatlı and Kırbız, 2022) and $+52\%$ SDR over AudioSp models, without accessing text prompts. Moreover, we particularly observe large increase in SIR demonstrating significantly less interference in OpenSep, due to the better disentanglement of corresponding sources.

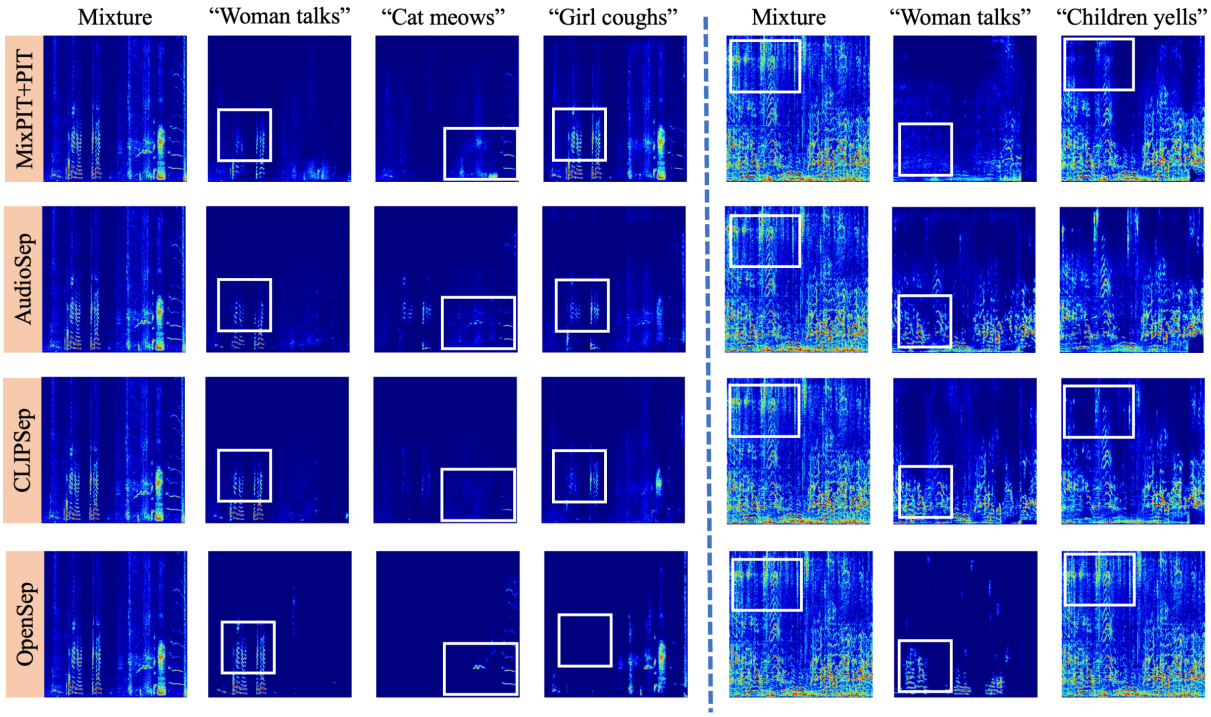

Figure 4: Qualitative results on natural mixtures from AudioCaps. (Left) All baselines show large spectral overlap of *woman talking* sound in other two source predictions. OpenSep precisely disentangles all three sources minimizing the spectral overlap across sources, while preserving spectral details. (Right) For the dominant noisy sound of *children yelling*, all baselines can hardly separate the *woman talking* sound. OpenSep significantly reduces noise in *woman talking*, while preserving spectral details of noisy *children yelling* sound.

Table 3: **User preference study on natural mixture separation**. OpenSep demonstrates superior separation quality without accessing conditional prompts.

Model	OpenSep wins (mean ± std)	Draws (mean ± std)	OpenSep loses (mean ± std)
vs. PIT	81.2 ±4.3	17.3 ±6.5	1.5 ±3.8
vs. MixIT	95.6 ±3.2	3.2 ±4.9	1.2 ±4.4
vs. CLIPSep	75.8 ±8.5	20.4 ±7.8	3.8 ±5.6
vs. AudioSep	65.8 ±7.3	30.9 ±8.5	3.3 ±6.6
vs. LASSNet	69.7 ±8.1	27.4 ±6.8	2.9 ±4.8

Generalization to Unseen Classes: For this analysis in Table 2, only 50% classes from each datasets are used for training, while the test set is formed with the remaining 50% classes in each set. Unconditional separators largely struggle in separating unseen sounds. Conditional separators comparatively achieve superior performance, however, we observe significant performance drop from seen classes. For example, CLIPSep experiences 40% and 54% SDR drops in MUSIC and VGGSound, respectively, while OpenSep achieves significantly higher performance over all baselines, with minimal performance SDR drops of 8% and 15%.

Comparison on higher order natural mixtures: We use natural mixtures from AudioCaps dataset containing 2 $\sim$ 3 sources. We train all models on VGGSound dataset and evaluate their performance on AudioCaps to mimic real-world use. For conditional models, we manually extract the condition prompts of all sources using the ground truth caption. Since we don't have access to ground truth sources for evaluation on natural mixtures, we perform human evaluations on the separated sounds given in Tab. 3. OpenSep wins over conditional models in separating all sources, even without using any conditions (71% *vs.* 3.8% excluding draws). When compared with unconditional models such as MixIT, OpenSep wins over 95% (*vs.* 1.2% excluding draws) cases denoting its robustness. We also provide some qualitative results in Fig. 4 (see more in appendix). Of note, OpenSep precisely detects, disentangles, and preserves the details in source separation from complex mixtures.

4.3 Ablation Study

We use VGGSound dataset in both seen and unseen cases for the ablation study.

Effect of knowledge parsing: We study the performance of OpenSep with and without knowledge parsing (Tab. 4). We observe that, by injecting

Table 4: Ablation study of proposed building blocks. The combination of LLM-knowledge parsing with few-shot prompting, and multi-level extension of mix-and-separate contribute to the best performance in OpenSep.

Knowledge Parsing	Few-shot Prompting	Multi-level Training	Seen Classes		Unseen Classes	
			SDR ↑	SIR ↑	SDR ↑	SIR ↑
✗	✗	✗	2.19 ± 0.27	5.25 ± 0.29	1.01 ± 0.31	4.13 ± 0.24
✗	✗	✓	2.78 ± 0.23	6.45 ± 0.23	1.95 ± 0.34	5.59 ± 0.32
✓	✗	✗	2.92 ± 0.31	6.84 ± 0.27	2.06 ± 0.29	6.01 ± 0.29
✓	✓	✗	3.26 ± 0.29	7.38 ± 0.22	2.56 ± 0.27	6.69 ± 0.33
✓	✓	✓	**3.71** ± 0.22	**8.31** ± 0.19	**3.14** ± 0.31	**7.23** ± 0.39

Table 5: Ablation of few-shot prompts in instruction-guided knowledge parsing from LLM.

Sample Shots	Seen Classes		Unseen Classes	
	SDR ↑	SIR ↑	SDR ↑	SIR ↑
1	3.31 ± 0.29	7.63 ± 0.27	2.67 ± 0.32	6.52 ± 0.26
2	3.52 ± 0.31	7.94 ± 0.31	2.93 ± 0.34	6.85 ± 0.29
3	3.63 ± 0.26	8.15 ± 0.28	3.06 ± 0.33	7.11 ± 0.31
5	**3.71** ± 0.22	**8.31** ± 0.25	**3.14** ± 0.31	**7.23** ± 0.39

Table 6: Ablation of different large language models in knowledge parsing for audio sources.

LLM Arch.	Seen Classes		Unseen Classes	
	SDR ↑	SIR ↑	SDR ↑	SIR ↑
LLaMA-3-8b	**3.71** ± 0.22	**8.31** ± 0.19	**3.14** ± 0.31	**7.23** ± 0.39
LLaMA-2-7b	3.58 ± 0.24	8.12 ± 0.24	3.06 ± 0.33	7.11 ± 0.25
Mistral-7b	3.63 ± 0.23	8.17 ± 0.21	3.09 ± 0.25	7.15 ± 0.29
Phi-3-medium	3.55 ± 0.29	8.01 ± 0.24	3.05 ± 0.29	7.12 ± 0.28
Gemma-7b	3.45 ± 0.31	7.83 ± 0.29	2.85 ± 0.29	6.93 ± 0.24

details of audio properties through knowledge parsing, we achieve SDR improvements of $+33\%$ and $+98\%$, on seen and unseen classes, respectively, highlighting the effectiveness of our approach.

Effect of proposed training method: We ablate the proposed multi-level extension of *mix-and-separate* framework (Tab. 4). We observe large improvements of 12% and 22% on seen and unseen classes, respectively, by using the proposed extension of *mix-and-separate* training. This shows its effectiveness in enhancing audio-language alignment for better performance.

Effect of few shot prompting: To guide LLM model parsing the salient audio properties, some high quality manually designed prompts are used. We ablate the effect of few shot prompting (see Tab. 4, 5). We observe notable performance gain with few shot prompting by guiding the LLM to extract required details.

Ablation of LLMs: We ablate the context from various open-source LLMs, such as LLaMA-2-7b, LLaMA-3-8b (Touvron et al., 2023), Mistral-7b (Jiang et al., 2023), Phi-3-medium (Abdin et al., 2024), and Gemma-7b (Team et al., 2024), while keeping the context length of 512 (See Tab. 6). Though we found competitive results across LLMs, LLaMA-3-8b generates best results.

Table 7: Analyzing source parsing accuracy with different audio captioning models in complex mixtures.

Caption Model	2-Source	3-Source	4-Source
ms-CLAP	96.89	91.57	87.93
Whisper-Large	93.47	88.09	84.68
ACT-Large	90.58	83.67	78.43
AudioLLM	**97.59**	**93.43**	**90.34**

Table 8: Effect of different audio captioning models in OpenSep performance on VGGSound dataset.

Combination	Seen Classes		Unseen Classes	
	SDR	SIR	SDR	SIR
ms-CLAP	3.71	8.31	3.14	7.23
AudioLLM	**3.94**	**8.68**	**3.56**	**7.95**

Accuracy of LLM Parser and captioning: To estimate the accuracy of the LLM parser with several captioning models, such as ms-CLAP (Elizalde et al., 2023), fine-tuned Whisper-Large (Kadlčík et al., 2023), ACT-Large (Mei et al., 2021), and AudioLLM (Gong et al., 2023). We use synthetic mixtures from VGGSound dataset for this analysis. Since each sample contains significant background noise which is not present in class labels, we only consider the top predictions with highest similarity with source labels in mixtures. We use a CLAP-text encoder to estimate the similarity between parsed source texts and ground truth labels. We observe considerably higher accuracy with AudioLLM captioning model compared to other models during source parsing for its large-scale pre-training.

Ablation with SOTA audio captioning model: The OpenSep framework is designed to be flexible and can incorporate any off-the-shelf model

for generating captions or parsing sound sources. AudioLLM (Gong et al., 2023), for example, can be used as a superior alternative to enhance caption generation over ms-CLAP (Elizalde et al., 2023). This model can be easily integrated into the current OpenSep framework. In Table 8, we present the comparative results using AudioLLM and ms-CLAP for audio captioning within the OpenSep framework in the VGGSound dataset. By introducing the AudioLLM in OpenSep, we observe significant performance improvements on both seen and unseen classes.

5 Conclusion

In this paper, we introduced OpenSep, a novel framework for audio source separation in open-world scenarios. In particular, OpenSep leverages an *off-the-shelf* audio captioning model and the world knowledge embedded in large language models (LLMs) to automatically parse and disentangle audio sources from noisy mixtures. By employing a text-conditioned audio separator and a multi-level mixture separation training objective, our method effectively enhances the alignment between conditional prompts and separated sources. Extensive experiments on three benchmark datasets demonstrate the superior performance of OpenSep, which achieves significant SDR improvements over state-of-the-art methods. Our work paves the way for practical, automated audio separation, addressing key limitations of existing methods and enabling future research in open-world audio processing.

6 Limitations

OpenSep performance is limited by the precise detection of sound sources in noisy mixtures, which mostly stems from the challenge in having a suitable audio captioning model. Nevertheless, OpenSep framework can potentially integrate any superior audio captioning approach to scale-up on real world cases. Finally, given its use of multiple building blocks, OpenSep is computationally expensive in general. However, by further optimizing the architecture, such as using mobile LLMs (*e.g.*, Phi-3-mini, Gemma-2b), OpenSep computational cost can be largely reduced, which we leave for future study.

7 Ethics Statement

We only use publily available dataset for this study.

Acknowledgement

This research was supported by ONR Minerva program, iMAGiNE - the Intelligent Machine Engineering Consortium at UT Austin, and a UT Cockrell School of Engineering Doctoral Fellowship.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. 2021. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1204–1213.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. Zero-shot audio source separation through query-based learning from weakly-labeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4441–4449.

Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian McAuley, and Taylor Berg-Kirkpatrick. 2022. Clipsep: Learning text-queried sound separation with noisy unlabeled videos. *arXiv preprint arXiv:2212.07065*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ruohan Gao and Kristen Grauman. 2019. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3879–3888.

Beat Gfeller, Dominik Roblek, and Marco Tagliasacchi. 2021. One-shot conditional audio filtering of arbitrary sounds. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 501–505. IEEE.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Marek Kadlčík, Adam Hájek, Jürgen Kieslich, and Radosław Winiecki. 2023. A whisper transformer for audio captioning trained with synthetic captions and transfer learning. *arXiv preprint arXiv:2305.09690*.

Ertuğ Karamatlı and Serap Kırbız. 2022. Mixcycle: Unsupervised speech separation via cyclic mixture permutation invariant training. *IEEE Signal Processing Letters*, 29:2637–2641.

Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R Hershey. 2019. Universal sound separation. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 175–179. IEEE.

Kevin Kilgour, Beat Gfeller, Qingqing Huang, Aren Jansen, Scott Wisdom, and Marco Tagliasacchi. 2022. Text-driven separation of arbitrary sounds. *arXiv preprint arXiv:2204.05738*.

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

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. 2023a. Separate anything you describe. *arXiv preprint arXiv:2308.05037*.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. 2023b. Separate anything you describe. *arXiv preprint arXiv:2308.05037*.

Xubo Liu, Haohe Liu, Qiuqiang Kong, Xinhao Mei, Jinzheng Zhao, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. 2022. Separate what you describe: Language-queried audio source separation. *arXiv preprint arXiv:2203.15147*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yuzhou Liu and DeLiang Wang. 2018. A casa approach to deep learning based speaker-independent co-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5399–5403. IEEE.

Rui Lu, Zhiyao Duan, and Changshui Zhang. 2018. Listen and look: Audio–visual matching assisted speech source separation. *IEEE Signal Processing Letters*, 25(9):1315–1319.

Yi Luo and Nima Mesgarani. 2018. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE.

Tanvir Mahmud, Saeed Amizadeh, Kazuhito Koishida, and Diana Marculescu. 2024. Weakly-supervised audio separation via bi-modal semantic similarity. *arXiv preprint arXiv:2404.01740*.

Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. 2021. Audio captioning transformer. In *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, pages 211–215, Barcelona, Spain.

Juan F. Montesinos. 2021. Torch-mir-eval: Pytorch implementation of mir-eval.

Fatemeh Pishdadian, Gordon Wichern, and Jonathan Le Roux. 2020. Finding strength in weakness: Learning to separate sounds with weak supervision. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2386–2399.

Ryosuke Sawata, Stefan Uhlich, Shusuke Takahashi, and Yuki Mitsufuji. 2021. All for one and one for all: Improving music separation by bridging networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 51–55. IEEE.

Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. 2019. Open-unmix-a reference implementation for music source separation. *Journal of Open Source Software*, 4(41):1667.

Naoya Takahashi and Yuki Mitsufuji. 2021. Densely connected multi-dilated convolutional networks for dense prediction tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 993–1002.

Reuben Tan, Arijit Ray, Andrea Burns, Bryan A Plummer, Justin Salamon, Oriol Nieto, Bryan Russell, and Kate Saenko. 2023. Language-guided audio-visual source separation via trimodal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10575–10584.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Yapeng Tian, Di Hu, and Chenliang Xu. 2021. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2745–2754.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. 2020. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. 2006. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469.

DeLiang Wang and Jitong Chen. 2018. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726.

Scott Wisdom, Aren Jansen, Ron J Weiss, Hakan Erdogan, and John R Hershey. 2021. Sparse, efficient, and semantic mixture invariant training: Taming in-the-wild unsupervised sound separation. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 51–55. IEEE.

Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron Weiss, Kevin Wilson, and John Hershey. 2020. Unsupervised sound separation using mixture invariant training. *Advances in Neural Information Processing Systems*, 33:3846–3857.

Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. 2017. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE.

Zhuohuang Zhang, Yong Xu, Meng Yu, Shi-Xiong Zhang, Lianwu Chen, and Dong Yu. 2021. Adl-mvdr: All deep learning mvdr beamformer for target speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6089–6093. IEEE.

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. 2018. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586.

A Appendix

A.1 Implementation Details

We use audio segments of 10s duration with a sampling rate of 16000 Hz for all experiments. Each audio sample is processed with short-term Fourier transform (STFT) using the frame window of 1022, and the hop length of 256. Following prior work (Dong et al., 2022; Zhao et al., 2018), the supervision is provided on the mask prediction for each source, instead of final reconstruction. We use the similar conditional U-Net based encoder decoder architecture following prior work (Mahmud et al., 2024). The U-Net model contains a total of seven successive encoding and decoding stages with convolutional kernels having 43.2M parameters. We apply self-attention followed by cross-attention in the skip connection of four bottom feature levels. A multi-head attention (Vaswani et al., 2017) is used with 8 heads for each attention operation. We use LLaMA-3-8b (Touvron et al., 2023) language model for parsing both sound sources and their corresponding details of audio properties, with 5-shot prompting for both parsing phases. These examples are manually curated and refined for guiding LLMs in diverse scenarios. We use RoBERTa-base language encoder to encode the extracted knowledge from LLM. A single sentence knowledge of audio properties is extracted for each parsed source targeting the maximum context length of 512. For the evaluation, we use *torch-mir-eval* (Montesinos, 2021) package for estimating both SDR and SIR in source separation from mixtures.

A.2 Sample of Source Parsing with LLM

We present several samples of textual inversion with audio captioning and source parsing from real-world audio mixtures collected from AudioCaps (Kim et al., 2019) in Table 9. We use ms-CLAP (Elizalde et al., 2023) model for audio captioning, followed by instruction-tuned LLaMA-3-8b model for source parsing. In most cases, the generated captions contain all sound sources presented in the audio mixtures. In general, the generated captions simplify the source descriptions compared to the ground truth captions. Nevertheless, by leveraging the detailed knowledge parsing from the LLM, we enrich details of each source. In a few cases of higher order mixtures, we observe the captioning model cannot detect muffled, short duration sounds. Nonetheless, we observe accurate source parsing with the LLM from the generated captions. Though source parsing is a simple objective, however, it can be complicated in some scenarios with complex captions. For example, the "*cat meowing*" sound is detected twice with the generated captions for repeated sounds, though it represents a single source. However, LLM based source parsing effectively solves such challenges.

A.3 Sample of Knowledge Parsing with LLM

We provide samples of knowledge parsing from LLM in Table 10 for various sound sources. We use instruction-tuned LLaMA-3-8b for such parsing. Several key properties of audios, such as frequency range, amplitude, dynamic envelope characteristics, usual duration, and spectral contents are focused by guiding the LLM with manually curated 5-shot prompts. Such details of audio properties largely help the audio separator model to ground diverse audio features with text description for enhanced separation, particularly in noisy, unseen mixtures.

A.4 Additional Qualitative Results

We provide additional qualitative comparisons for natural audio mixture separation in Fig. 5 and Fig. 6. In general, OpenSep largely reduces the spectral overlap across multiple sources while preserving details of each source in separation from challenging mixtures, compared to state-of-the-art baseline methods, without accessing manual source prompts. We also provide audio samples from different models in the supplementary for better understanding of the separation performance.

A.5 Details of User Study

We conduct a user study to analyze the audio separation performance from real-world mixtures, where we don't have access to any ground truth sources. Each user is provided with 20 mixture samples, and their corresponding separated audios with competitive models. For each sample, user rates the superior separation quality between two choices. We collect the data of wins, loses, and ties across different models using the user study. The results of this study are reported in Table 3. We attach the screenshot of the human evaluation form with detailed guidelines in Fig. 7.

Table 9: Sample of source parsing and generated captions from real-world audio mixtures of AudioCaps dataset. We use ms-CLAP model for audio captioning followed by the source parsing from LLaMA-3-8b with five shot prompts.

Ground Truth Captions	**Generated Captions**	**Parsed Sources**
People talk while an audience laughs and water runs from a faucet	A man and woman are talking and laughing while water runs in the background.	A man talks. A woman talks. People laugh. Water runs.
A child screams, an adult male is talking and vehicles are revving.	A man speaks while a vehicle engine revs.	A man speaks. A vehicle engine revs.
A child cries and yells and an adult speaks in response	A child cries and a woman speaks	A child cries. A woman speaks.
A whistle and another man speaks and the whistling stops.	A man speaks followed by whistling.	A man speaks. A whistling sound.
A crowd is cheering and shouting, thumping occurs, an adult female speaks.	A crowd is cheering and a woman is speaking.	A crowd is cheering. A woman is speaking.
A large engine roars and someone is shouts a loud short rout.	A motor is running while people are shouting.	A motor is running. People are shouting.
A cat is meowing over and over	A cat meows and a cat meows.	A cat meows.
People are talking and bees are quietly buzzing in the background.	A bee buzzes and a man speaks.	A bee buzzes. A man speaks.
International music plays as water pours into a pot and finally some splashes	A stream of water is running while music plays in the background.	A stream of water is running. Music plays in the background.
Several small shuffles of hard objects knocking, followed by loud explosions interspersed by a grown man chuckling.	A loud bang followed by laughter.	A loud bang. Someone laughs.
Birds are flapping their wings and some bird call	Birds chirping and flapping their wings.	Birds chirping. Birds flapping their wings.
A toilet flushes and then woman speak in the background	A toilet flushes followed by a woman speaking.	A toilet flushes. A woman is speaking.

Table 10: Samples of knowledge parsing for various audio sources from the LLM. We use instruction-tune LLaMA-3-8b model for extracting the detailed audio properties of individual sources. We use manually curated 5-shot prompts to guide the LLM for focusing on specific audio properties for assisting the source spearation.

Source	**Parsed Knowledge**
Alarm clock ringing	The sound of an alarm clock ringing is represented by a sharp, piercing, and repetitive tone, typically in the frequency range of 1-4 kHz, with a characteristic "square-wave" or "sawtooth" shape, featuring a sudden onset, a steady amplitude, often accompanied by a slight decay and a gentle reverberation tail, and a relatively high overall energy level due to the sudden and attention-grabbing nature of the sound.
Baby laughter	The sound of baby laughter is represented by a series of short, high-pitched, and joyful vocalizations, typically in the frequency range of 200-400 Hz, characterized by a rapid sequence of rising and falling frequencies, with a relatively constant amplitude, often accompanied by subtle variations in pitch and timbre, featuring a "giggly" or "chirpy" quality, and a relatively short duration of around 0.5-1.5 seconds, with a gentle, warm, and intimate quality, and a moderate overall energy level.
Cat hissing	The sound of a cat hissing is represented by a high-pitched, sharp, and intense vocalization, typically in the frequency range of 2-4 kHz, with a distinctive "s" or "sh" shape in the spectrogram, featuring a rapid series of short, staccato bursts of energy, often accompanied by a subtle tremolo effect, and a relatively short duration of around 0.1-0.3 seconds, with a moderate to high overall energy level, and a characteristic "attack"and "decay"pattern, with a sudden onset and a gradual release.
Duck quacking	The sound of a duck quacking is represented by a series of loud, sharp, and explosive vocalizations, typically in the frequency range of 200-400 Hz, with a characteristic "honking" quality, featuring a sudden onset and decay, and a relatively constant amplitude, often accompanied by subtle variations in pitch and timbre, and a moderate to high overall energy level, with a distinctive spectral shape and a duration of around 0.5-1.5 seconds, and a possible presence of echoes or reverberations in the environment.
Fox barking	The sound of a fox barking is represented by a series of short, sharp, and high-pitched vocalizations, typically in the frequency range of 1-5 kHz, with a characteristic "yipping" or "yelping" quality, featuring a rapid sequence of rising and falling frequencies, and a relatively constant amplitude, often accompanied by subtle variations in pitch and timbre, and a relatively short duration of around 0.2-0.5 seconds, with a moderate overall energy level, and a distinctive spectral shape featuring a prominent peak in the 2-3 kHz range.
Playing accordion	The sound of playing an accordion is represented by a rich, dynamic, and complex mixture of sounds, characterized by a wide frequency range of 50-2000 Hz, featuring a prominent low-frequency foundation (around 50-100 Hz) provided by the instrument's bellows and the diatonic reeds, overlaid with a series of bright, piercing, and harmonically-rich tones produced by the instrument's buttons and keys, with a distinctive "oom-pah" or "chord-like" quality, and often accompanied by subtle vibrato and tremolo effects, as well as occasional percussive attacks and releases, all blending to create a lively, folk-inspired, and emotive sound.
Train whistling	The sound of a train whistling is represented by a distinctive, high-pitched, and piercing tone, typically in the frequency range of 2,000-4,000 Hz, with a sharp attack and decay, and a characteristic "siren-like" shape, often accompanied by a subtle tremolo effect, and a relatively long duration of around 1-3 seconds, with a moderate to high overall energy level, and a sense of spatiality and distance, as if the sound is coming from a specific location.
Vacuum cleaner cleaning floors	The sound of a vacuum cleaner cleaning floors is represented by a dominant low-frequency hum, typically in the range of 50-200 Hz, with a strong amplitude modulation caused by the motor's rotation and the movement of the vacuum head, often accompanied by a series of high-frequency clicks and rattles from the brushes and wheels, and a gentle whooshing sound from the airflow, with a moderate overall energy level and a relatively consistent spectral shape, punctuated by brief, high-amplitude transients when the vacuum head encounters obstacles or changes direction.
Waterfall burbling	The sound of a waterfall burbling is represented by a continuous, gentle, and soothing audio signal, featuring a prominent low-frequency energy peak in the range of 20-50 Hz, accompanied by a series of soft, repetitive, and varying pitched tones, typically in the frequency range of 100-500 Hz, with a gradual spectral roll-off towards higher frequencies, and a characteristic "chirping" or "bubbling" quality, often punctuated by occasional, brief, and low-amplitude bursts of energy, and a relatively long duration of several seconds to minutes, with a moderate to high overall energy level.
Writing on blackboard with chalk	The sound of writing on a blackboard with chalk is represented by a series of sharp, scratchy, and percussive sounds, typically in the frequency range of 100-500 Hz, with a characteristic "scratch-and-scrabble" texture, featuring a mix of high-amplitude, short-duration events (corresponding to the chalk striking the board) and lower-amplitude, longer-duration events (corresponding to the chalk gliding across the board), often accompanied by subtle variations in tone and timbre depending on the chalk's velocity and angle of incidence, and a relatively low overall energy level due to the soft and dry nature of the writing surface.

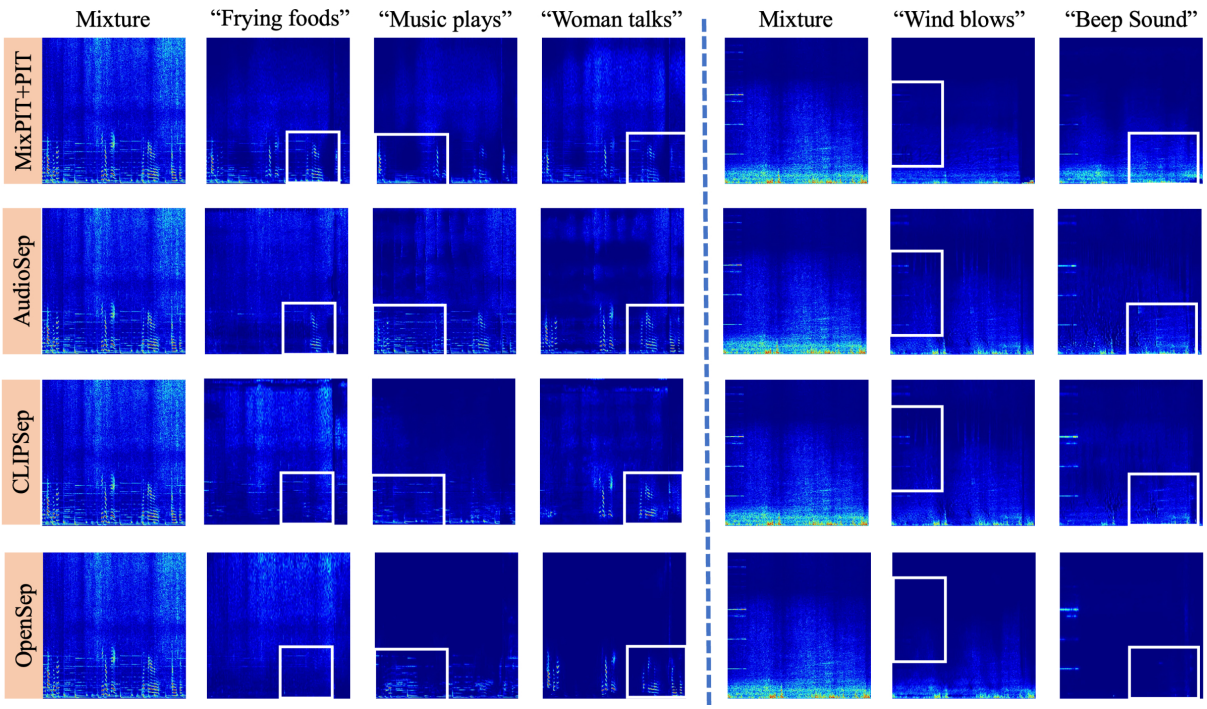

Figure 5: Qualitative results on natural mixtures from AudioCaps. (Left) We can observe the dominant "*woman talks*" spectral content in "*frying foods*" for most baselines. However, in CLIPSep, such overlap is largely reduced, but horizontal spectral contents from "*music plays*" is visible. In contrast, OpenSep largely reduces such spectral overlap in all three components while preserving all details. (Right) In this mixture, the "*beep sound*" is only present at the beginning, with large noisy sound of "*wind blows*" over the spectrogram. Most baseline methods contain noisy spectral contents in the "*beep sound*", while losing spectral contents in the "*wind blows*" prediction. In contrast, OpenSep disentangles this noisy mixture with significant reduction of spectral overlaps.

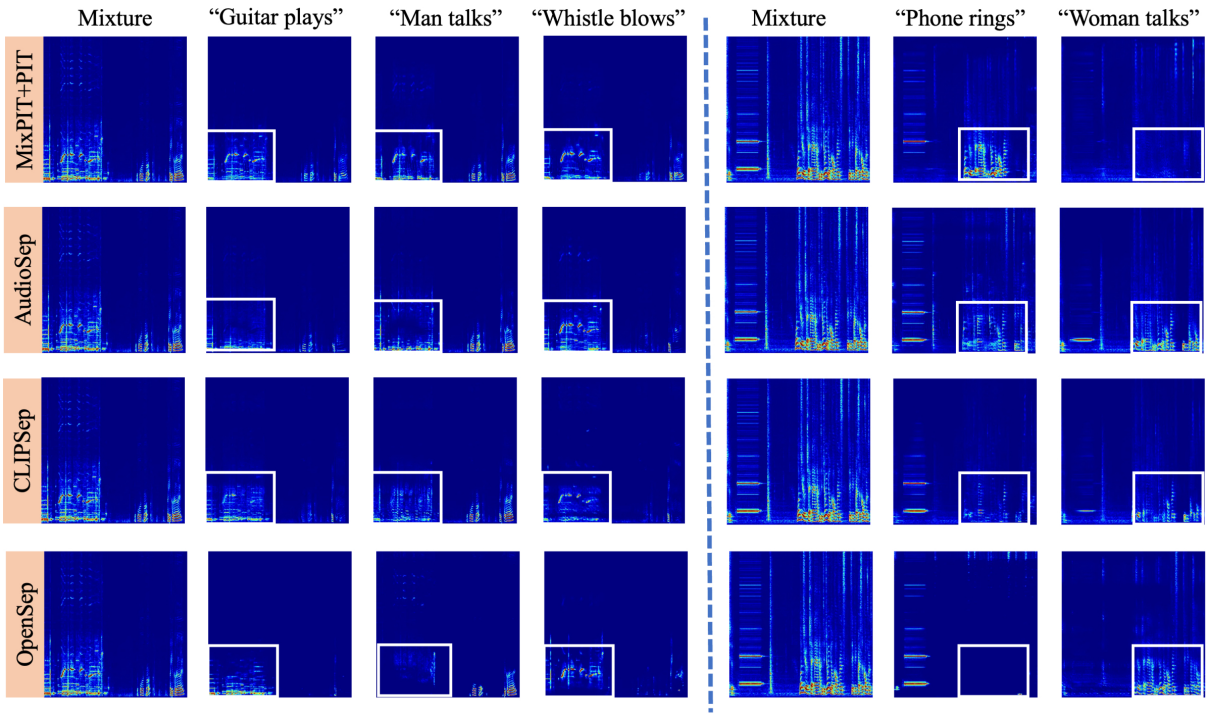

Figure 6: Qualitative results on natural mixtures from AudioCaps. (Left) In the initial phase, the "*whistle blows*" and "*guitar plays*" sounds are present, while the "*man talks*" sound appears at the end. In all baselines, we can see significant spectral overlaps of the "*whistle blows*" and "*guitar plays*" sounds. In contrast, OpenSep precisely separates both of these challenging components, while also reducing background contents in the "*man talks*" prediction. (Right) The "*phone rings*" sound mostly appears at the beginning followed by the "*woman talks*" sound at the end. Compared to all baselines, OpenSep more sharply disentangles both of these sources from this challenging mixture highlighting its effectiveness in practice.

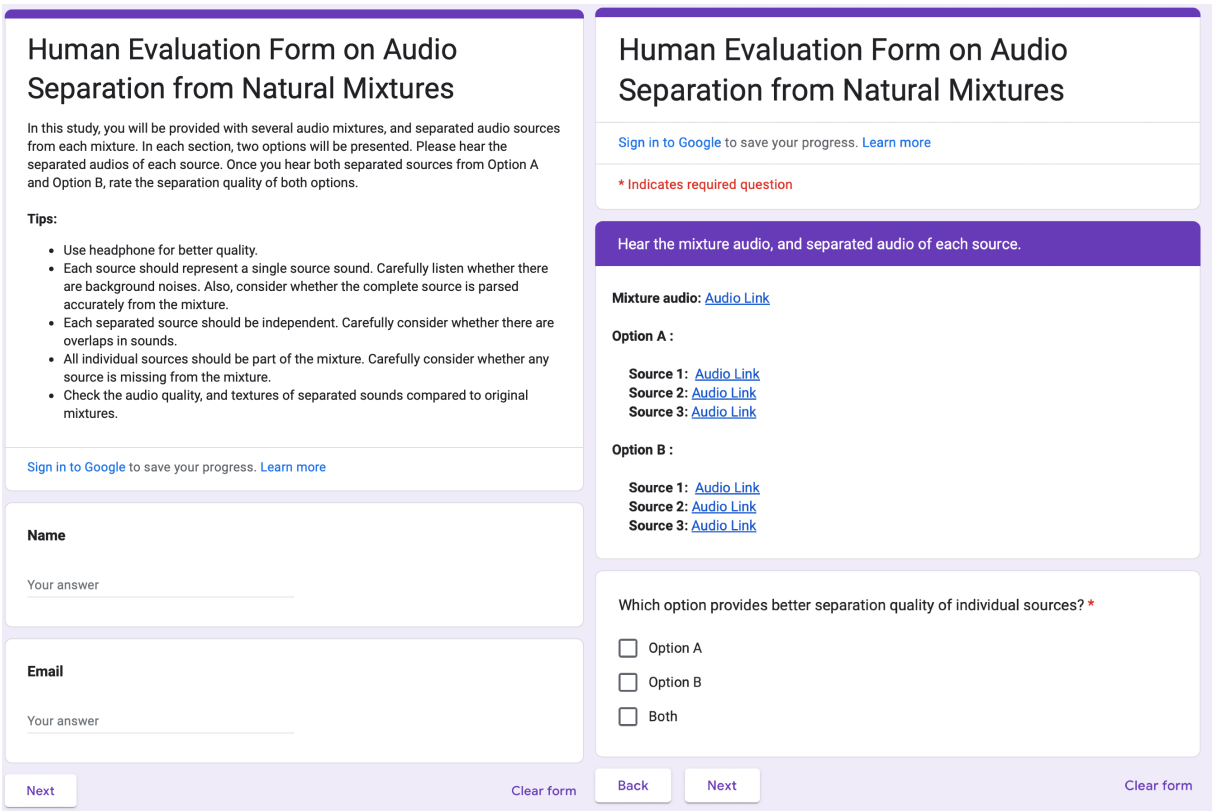

Figure 7: Screenshot of human evaluation form used for the user study. We assess the performance of audio separation from natural mixtures by human evaluation, without having access to any ground truth audios of individual sources. Each user compares the separation quality of two competitive models for several mixtures.