Universal Sound Separation:

Distance-Aware Mixture Simulation,

Co-occurrence Conditioning,
and Chain-of-Inference

A Thesis

Submitted to the Graduate Faculty of

The University of Texas at Arlington

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science in Computer Science

by

Wonjun Park

Arlington, Texas

2026

To my parents, who made this extraordinary journey possible.

이를 가능하게 해준 부모님께.

Abstract

Universal Sound Separation (USS) – the task of disentangling arbitrary sound sources from a single-channel acoustic mixture – remains an open challenge due to the ill-posed nature of the problem and the distributional gap between synthetic training data and real-world recordings. This thesis addresses three distinct bottlenecks in the USS pipeline: training data realism, inference strategy, and conditioning richness.

We first present two knowledge-guided approaches to sound source separation. The first is a *distance-aware* mixing strategy that leverages Large Language Models (LLMs) to assign plausible loudness relationships between audio sources during training data synthesis. By querying an LLM about the natural acoustic distance between sound events, we generate Mixtures of Mixtures (MoMs) that better approximate real-world acoustic scenes. Human evaluation shows that models trained with this strategy are preferred over randomly-trained baselines in up to 75% of comparisons on three real-world benchmark categories. The second is a *co-occurrence conditioning* framework that injects information about non-target sounds present in a mixture into the encoder of AudioSep via FiLM modulation, complementing the standard target conditioning. We propose a CLAP-based estimation procedure ($\hat{a}_2$) that approximates co-occurrence embeddings at inference time from only the mixture and the target text, matching the practical setting of USS; an exploratory evaluation shows improved separation on five of six USS benchmarks.

We then introduce Chain-of-Inference (CoI), a *training-free* multi-step inference framework motivated by the human auditory system's sensitivity to sudden changes in the acoustic scene and structurally analogous to Chain-of-Thought prompting in language models. CoI iteratively re-introduces a proportion of the original mixture –

governed by cosine similarity between the current output and the input – progressively decomposing the separation problem into easier sub-problems. Without any additional training, CoI consistently improves AudioSep across all five evaluated tasks and SAM-Audio on four of five. An interactive online demonstration system is released alongside this work, allowing users to experience the perceptual improvements on arbitrary audio.

Taken together, these contributions show that USS performance can be improved from two distinct angles: incorporating external knowledge – LLM commonsense priors and contrastive audio-text embeddings – to improve training data and conditioning, and exploiting underutilised capacity already present in frozen models through principled inference-time refinement.

Acknowledgements

I would like to express my deepest appreciation to my advisor, Dr. Kenny Q. Zhu, for his invaluable guidance throughout my graduate studies. Our weekly one-on-one meetings with his insightful feedback and mentorship were instrumental in shaping every aspect of this work, and the road trip we shared remains one of my most memorable experiences during this journey.

I am likewise sincerely grateful to my committee members, Dr. Vassilis Athitsos and Dr. Shirin Nilizadeh, for their time, careful review, and constructive feedback throughout the process.

I also wish to thank my lab mates who actively participated in every weekly meeting and for the meaningful discussions we shared. Their feedback, insightful questions, and collaborative spirit have been invaluable throughout this journey.

Finally, I am grateful to the friends I have met before and during this graduate journey for their kindness, encouragement, and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Acoustic World and the Separation Problem

Every moment, the world produces an immense tapestry of overlapping sounds. In a busy café, the hiss of an espresso machine, the murmur of conversation, the clink of cutlery, and the faint melody of background music all arrive at a listener's ears simultaneously. In a forest at dawn, bird calls layer upon insect chirps, wind through leaves, and the distant rush of a stream. In a recording studio, the bleed between microphones captures not only the target instrument but also the room and its reflections. The human auditory system is extraordinarily adept at segregating these sources, directing attention to the voice of a friend across a crowded room – the *cocktail-party effect* [7] – while suppressing all else.

Replicating this capability computationally is the goal of *source separation*. More specifically, *Universal Sound Separation* (USS) [26] aims to isolate an *arbitrary* sound source from a single-channel acoustic mixture without prior knowledge of which source categories are present. This universality is what distinguishes USS from the specialised tasks it generalises: Speech Enhancement (SE), which isolates a target speaker from background noise, speech separation, which disentangles overlapping speakers, and Music Source Separation (MSS), which decomposes musical recordings into a fixed set of predetermined instrument stems. USS must generalise across the full breadth of

acoustic categories encountered in the wild.

1.2 Applications and Motivation

The practical demand for USS is wide-ranging and spans multiple scientific and commercial disciplines.

Animal language processing. Bioacoustics researchers studying animal communication must first isolate target vocalisations from field recordings contaminated by other species, environmental noise, and human interference. Projects such as the transcription of Shiba Inu vocalisations [23] and the phonetic analysis of canine speech [65] require a pre-processing step that removes all sound sources except the target animal. Similarly, biodenoising work [40] shows the pressing need to separate animal calls from background noise without access to clean isolated recordings. A robust general-purpose separator would dramatically accelerate this line of research.

Assistive listening and accessibility. Hearing-aid algorithms must suppress competing noise sources and enhance the signal of interest in real time. Multimicrophone binaural systems have shown that speech enhancement can be performed while preserving spatial cues critical for hearing-impaired users [56], and recent work exploits the wearer's own voice to identify and suppress interfering speakers [20]. Speaker-conditioned extraction algorithms have been evaluated directly with hearing-impaired listeners, demonstrating reduced listening effort [53], while location-aware models steer extraction toward arbitrary directions using low-latency architectures suitable for hearing aids [1]. Systems that can additionally accept a natural-language specification of the desired source – "amplify the speaker in front of me" – would represent a further qualitative leap in accessibility technology.

Speech and dialogue systems. Automatic speech recognition (ASR) and spoken dialogue systems degrade significantly in the presence of competing sound sources. Permutation invariant training has enabled single-channel multi-talker speech recognition, cutting word error rates by up to 45% relative on two-speaker mixtures [46]. End-to-end approaches jointly optimise source extraction and ASR, using location cues to guide target speech extraction without requiring parallel clean data [55]. Preprocessing the audio stream with a targeted separator before the ASR module is now a practical pipeline improvement, especially in far-field and multi-talker environments.

Music production and remixing. Modern creative tools increasingly allow users to remix existing recordings by isolating and manipulating individual stems. Open-source systems such as Open-Unmix [54] and Spleeter [18] have made deep-learning-based music separation accessible to non-expert users, while community benchmarks like the Music Demixing Challenge [41] continue to drive progress. USS would extend this capability beyond the studio, allowing arbitrary stems to be extracted from live recordings or archival audio without multi-track sources.

Smart-home environments. Smart-home devices must detect and classify acoustic events – a doorbell, a smoke alarm, a baby crying – in the presence of television audio, music, and conversation. Source separation has been shown to improve sound event detection in such multisource domestic environments [17], and jointly trained models that perform separation and detection simultaneously achieve stronger polyphonic detection in household settings [9, 10]. As smart speakers and home assistants proliferate, USS offers a natural front-end for robust acoustic scene understanding in everyday living spaces.

1.3 Challenges

Despite significant progress driven by deep learning, USS faces fundamental challenges that limit its deployment in real-world settings.

Ill-posedness. Given a single-channel mixture $m = s_1 + s_2 + \cdots + s_K$, recovering all K sources is an under-determined inverse problem. Even with two sources, there are infinitely many decompositions that sum to the observed mixture. A model must use learned priors to break this ambiguity, and the quality of those priors is ultimately limited by the quality and diversity of training data.

Training data scarcity. Obtaining clean, isolated recordings of arbitrary sound events is expensive or impossible. Field recordings always contain some degree of environmental contamination. The mix-and-separate paradigm [69] sidesteps this requirement by constructing synthetic training data from readily available audio, but introduces its own distributional challenges, as discussed below.

Distributional gap between training and deployment. When training MoMs are constructed by randomly pairing audio clips, the resulting mixtures often represent acoustically implausible combinations – a car horn and a whale vocalisation at equal volume, for instance. Because a model's learned priors are shaped by its training distribution, exposure to implausible mixtures during training may lead to suboptimal priors for real-world deployment.

Scalability of inference. Most pre-trained USS models are large neural networks that process audio in a single forward pass; even diffusion-based models such as SAM-Audio, which use multiple internal denoising steps, produce their output in a single invocation without revisiting the result. For complex, highly overlapping mixtures,

this one-shot inference may not be sufficient to fully disentangle all sources, yet there is no principled mechanism to allow the model to "try harder" without retraining.

Conditioning limitations. Text-guided models such as AudioSep [34] condition separation on a description of the desired output. However, the conditioning signal describes only the target source; the model receives no information about the non-target sounds also present in the mixture. This asymmetry may limit the model's ability to suppress interfering sources.

1.4 Research Questions

This thesis investigates three specific research questions that address the challenges identified above:

RQ1. Training data realism. Can the distributional gap between synthetic training MoMs and real-world acoustic scenes be reduced by incorporating world knowledge about the natural co-occurrence distances of sound sources, and does doing so improve separation performance on real-world audio?

RQ2. Conditioning richness. Can knowledge about the co-occurring, non-target sound sources in a mixture be leveraged to improve conditioned separation, and can such knowledge be estimated at inference time from the mixture and the target description alone?

RQ3. Inference strategy. Can the chain-of-thought principle – iterative problem decomposition – be transferred from language reasoning to sound separation to improve the utilisation of pre-trained model capacity, without any additional training?

1.5 Contributions

This thesis makes the following contributions in response to the research questions above:

C1. A **distance-aware audio mixing framework** (addressing RQ1) that uses an LLM to assign plausible loudness relationships between audio sources based on their natural co-occurrence distance, yielding more realistic MoMs. Human evaluation demonstrates that models trained with this strategy are preferred over random-mixing baselines in up to 75% of comparisons on real-world benchmarks, and that widely-used randomly-synthesised benchmarks systematically undervalue methods that model real acoustic scenes.

C2. A **co-occurrence conditioning framework** (addressing RQ2) that injects information about non-target co-occurring sources into the encoder of AudioSep's ResUNet architecture, together with a CLAP-based estimation procedure ($\hat{a}_2$) for recovering co-occurrence embeddings at inference time from the mixture and the target description, without ground-truth source labels.

C3. **Chain-of-Inference (CoI)** (addressing RQ3), a training-free multi-step inference framework that iteratively re-introduces the original mixture at a cosine-similarity-derived proportion. CoI consistently improves AudioSep [34] across all five evaluated tasks and SAM-Audio [52] on four of five tasks. The framework is deployed as an interactive online demonstration system at `https://redgiant.uta.edu/sound_separation`.

1.6 Publication Record

The contributions of this thesis have been disseminated in the following publications:

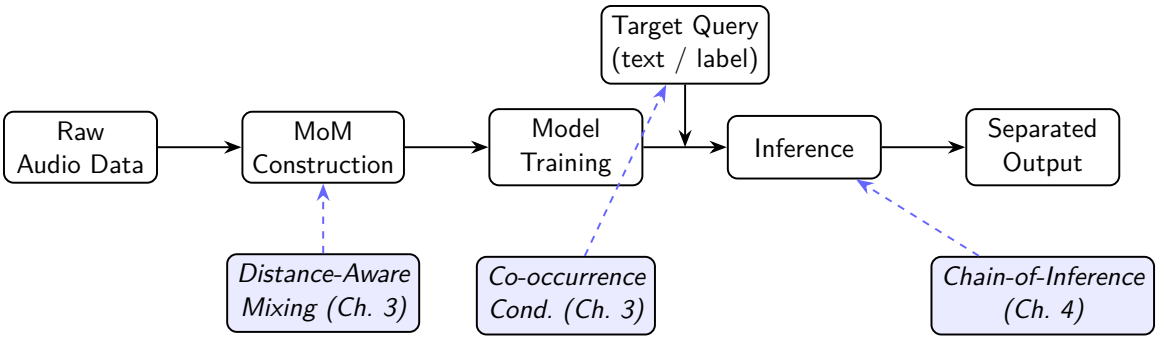


Figure 1.1: Overview of the thesis contributions mapped onto the USS pipeline. Distance-aware mixing (Section 3.1) improves training data synthesis; co-occurrence conditioning (Section 3.2) enriches the conditioning signal; Chain-of-Inference (Chapter 4) improves inference without retraining. Dashed arrows indicate the point of intervention for each contribution.

- **Chapter 3, Section 3.1:** Wonjun Park, Tuan M. Dang, and Kenny Q. Zhu. "Towards Distance-Aware Synthetic Audio Mixtures for Universal Sound Separation." *Accepted at ICASSP 2026.*

- **Chapter 3, Section 3.2:** Wonjun Park and Kenny Q. Zhu. "Guidance of Co-occurrence Knowledge and its Estimation in Latent Space for Conditioned Sound Separation." *Work in progress.*

- **Chapter 4:** Wonjun Park and Kenny Q. Zhu. "An Incremental Framework for Sound Source Separation." *Under review at ACL 2026 System Demonstration Track.*

1.7 Thesis Organisation

Figure 1.1 illustrates how the three contributions of this thesis map onto the USS pipeline.

The remainder of this thesis is organised as follows. Chapter 2 provides the technical background necessary to understand the contributions, covering deep learning for audio, the USS problem formulation, text-audio alignment models, and chain-of-thought

reasoning. Chapter 3 presents two knowledge-guided approaches: distance-aware data synthesis and co-occurrence conditioning. Chapter 4 introduces Chain-of-Inference, a training-free multi-step inference framework. Chapter 5 describes an interactive online demonstration system that unifies all contributions into a single platform. Chapter 6 discusses the broader implications, limitations, and future directions.

Chapter 2

Background

2.1 Audio Signal Representations

Raw mono audio is a one-dimensional time-domain signal $x \in \mathbb{R}^T$ sampled at a rate f_s (typically 16 kHz, 32 kHz, or 44.1 kHz). Deep learning models for audio operate on raw waveforms, time-frequency representations derived from the Short-Time Fourier Transform (STFT), or hybrid combinations of both.

Short-Time Fourier Transform. The STFT of a signal x with window function w, hop length H, and window length L is:

$$X(t, f) = \sum_{\tau=0}^{L-1} x(\tau + tH)\, w(\tau)\, e^{-j2\pi f\tau/L}, \tag{2.1}$$

producing a complex-valued spectrogram $X \in \mathbb{C}^{F\times T'}$ where $F = L/2 + 1$ is the number of frequency bins and T' is the number of time frames. The magnitude $|X|$ (or log-magnitude) is widely used as a feature for audio classification and source separation, as it captures the energy distribution of sound across time and frequency in a perceptually meaningful representation. Some modern architectures take the full complex spectrogram as input to jointly estimate magnitude and phase; for example, the ResUNet in AudioSep [34] predicts a magnitude mask $|M|$ and a phase residual $\angle M$, recovering the separated spectrogram as $\hat{S} = |M| \odot |X|\, e^{j(\angle X + \angle M)}$.

Mel spectrogram. A Mel spectrogram warps the frequency axis onto the Mel scale, which approximates the non-linear frequency resolution of the human auditory system. Due to this perceptual alignment, it is widely used in speech-related tasks including speech separation and audio classification. It is less common in general source separation pipelines that require phase information for waveform resynthesis.

2.2 Deep Learning for Source Separation

2.2.1 The U-Net Architecture

U-Net [49], originally proposed for biomedical image segmentation, has become the dominant backbone for spectrogram-domain source separation. It consists of an encoder that progressively downsamples the input, a bottleneck, and a decoder that progressively upsamples, with skip connections bridging corresponding encoder and decoder layers. Applied to spectrograms, the encoder compresses the input into a compact latent representation that captures abstract source characteristics, while the decoder reconstructs the magnitude spectrogram of the target source.

The output of a U-Net separator is typically a *soft mask* $M \in [0,1]^{F \times T'}$ applied elementwise to the input mixture spectrogram:

$$\hat{S} = M \odot X, \tag{2.2}$$

where $\hat{S}$ is the estimated target spectrogram and the separated audio is obtained by inverse STFT using the original mixture phase. More advanced variants decouple magnitude and phase estimation – predicting a magnitude mask alongside a phase residual – so that the network can modify both components rather than borrowing the noisy mixture phase (see Section 2.1 above for the AudioSep formulation).

2.2.2 Waveform-Domain Models

An alternative to spectrogram-domain processing is to operate directly on the raw waveform. Conv-TasNet [35] learns a 1-D convolutional encoder, a temporal convolutional network (TCN) for mask estimation, and a 1-D convolutional decoder. TDCN++ [69] extends this architecture with improved skip connections and improved normalisation, achieving strong performance in unsupervised USS settings.

Waveform-domain models avoid the phase reconstruction problem of spectrogram-domain models but tend to require more computational resources and are sensitive to the choice of encoder/decoder filters.

2.2.3 Conditioning Mechanisms

A *conditioned* separator takes as input not only the audio mixture but also a conditioning signal that specifies the desired output. The most common conditioning mechanism in audio is Feature-wise Linear Modulation (FiLM) [44], which applies an affine transformation to intermediate feature maps:

$$\text{FiLM}(h_i \mid \gamma_i, \beta_i) = \gamma_i \odot h_i + \beta_i, \tag{2.3}$$

where h_i is the i-th feature map, and γ_i, β_i are predicted from the conditioning signal. CUNet [39] inserts FiLM layers into the encoder of a U-Net to condition on the target instrument. LaSAFT [8] found through preliminary experiments that applying FiLM in the decoder was consistently better than the encoder, because encoder-level conditioning makes the latent space more discontinuous. LaSAFT further proposed Gated Point-wise Convolutional Modulation (GPoCM) as an extension of FiLM, and Latent Source Attentive Frequency Transformation to capture instrument-dependent frequency patterns via scaled dot-product attention.

2.3 Universal Sound Separation

2.3.1 Problem Formulation

Given a single-channel mixture $m_0 = \sum_{k=1}^{K} s_k$ where s_k are individual source signals and K is unknown, USS aims to recover a specified target source s_k using a query q. When q is absent (unconditional USS), the model aims to separate all K sources simultaneously; when q is a text description, the model performs *language-queried* USS.

2.3.2 Mixture Invariant Training

MixIT [69] is the key enabling methodology for unsupervised USS. Given two mixtures $m^{(1)}$ and $m^{(2)}$, a MoM is formed:

$$m^{(1,2)} = m^{(1)} + m^{(2)}. \tag{2.4}$$

A separation model f with C output channels produces C separated signals $\hat{s}_1, \ldots, \hat{s}_C$ from $m^{(1,2)}$. Training minimises the best assignment of model outputs to original mixtures:

$$\mathcal{L}_{\text{MixIT}} = \min_{\mathbf{A}} \sum_{c=1}^{C} \mathcal{L}\left(\sum_{c' :\, A_{cc'}=1} \hat{s}_{c'},\ m^{(c)} \right), \tag{2.5}$$

where $\mathbf{A}$ is a binary assignment matrix and $\mathcal{L}$ is a signal-level loss such as negative SI-SDR. Since $m^{(1)}$ and $m^{(2)}$ are themselves mixtures, no clean isolated sources are required. Figure 2.1 illustrates the MixIT training pipeline.

2.3.3 AudioSep and Language-Queried Separation

AudioSep [34] extends a frequency-domain ResUNet with a CLAP text encoder to perform language-queried USS. The text query q (e.g., “the sound of a dog barking”)

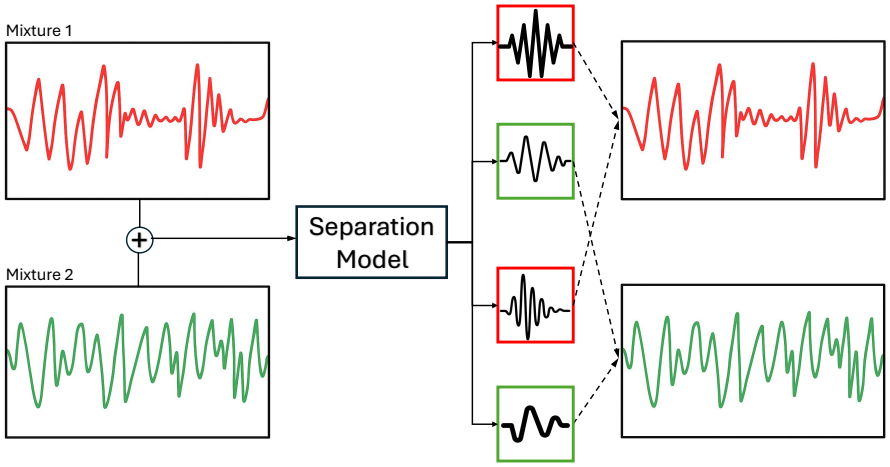


Figure 2.1: Overview of Mixture Invariant Training (MixIT). Two mixtures are combined into a Mixture of Mixtures (MoM), which is fed to the separator. Each model output is assigned to one of the original mixtures via a binary assignment matrix **A**, and the loss is minimised over all possible assignments. No isolated ground-truth sources are required.

is encoded by a pre-trained CLAP text encoder into a fixed-dimensional embedding $\phi(q) \in \mathbb{R}^d$, which conditions the ResUNet's feature maps via FiLM layers. AudioSep is trained on a broad collection of audio-text pair datasets and achieves strong zero-shot generalisation to unseen sound categories.

2.3.4 SAM-Audio

SAM-Audio [52] extends the Segment Anything paradigm [30] to the audio domain. Built on a diffusion transformer architecture trained with flow matching, it unifies text, visual, and temporal span prompting within a single framework, allowing users to specify target sources via natural language, visual masks or clicks, or time-span selection. Because of its generative diffusion-based decoder, SAM-Audio outputs are perceptually high-quality but not well-correlated with traditional waveform-domain metrics (SDR, SI-SDR).

2.4 Evaluation Metrics

SI-SDR. The Scale-Invariant Signal-to-Distortion Ratio [51] is:

$$\text{SI-SDR}(\hat{s}, s) = 10 \log_{10} \left(\frac{\|\alpha s\|^2}{\|\hat{s} - \alpha s\|^2} \right), \quad \alpha = \frac{\hat{s}^\top s}{\|s\|^2}, \tag{2.6}$$

where s is the reference source and $\hat{s}$ is the estimate. SI-SDR is scale-invariant, making it more robust to global amplitude differences than SDR.

SDR. The Signal-to-Distortion Ratio [61] measures the ratio of target energy to distortion energy without scale normalisation. It is the standard metric for MSS (MUSDB18) and DVS evaluations.

SAM-Audio Judge (SAJ). Traditional metrics such as SDR and SI-SDR compare the separated waveform against a reference signal, penalising deviations in phase and fine-grained temporal structure. Diffusion-based models like SAM-Audio generate audio through a stochastic process, producing outputs that are perceptually faithful but not waveform-aligned with the reference – leading to poor SDR/SI-SDR scores despite high perceived quality. SAJ [64] addresses this mismatch as a reference-free evaluation framework that produces an overall score in $[1, 5]$ reflecting perceptual quality of the separated output, more closely correlating with human judgement than waveform-domain metrics.

2.5 Contrastive Language-Audio Pretraining (CLAP)

CLAP [12, 70] is a joint audio-text embedding model trained with a contrastive objective, analogous to CLIP for images. Given audio a and text t, CLAP trains dual

encoders such that:

$$\phi_{\text{audio}}(a) \approx \phi_{\text{text}}(t) \quad \text{if } a \text{ and } t \text{ describe the same content,} \tag{2.7}$$

with the similarity measured by cosine similarity. This alignment enables zero-shot audio classification (by comparing audio embeddings to text label embeddings), text-guided audio retrieval, and – as exploited in this thesis – inference about the acoustic content of a mixture from its embedding.

2.6 Large Language Models and World Knowledge

Contemporary LLMs such as GPT-4 [43] and LLaMA [57] encode an enormous breadth of world knowledge acquired from text corpora spanning virtually every human domain. Wei et al. [67] showed that large-scale language models exhibit *emergent abilities* – capabilities that are absent in smaller models but appear once a critical scale is reached – including multi-step reasoning and commonsense inference, suggesting that sufficient scale enables the acquisition of broad world knowledge. This knowledge includes relationships about the physical world: which objects tend to co-occur, at what distances, and in what environments. In this thesis, we exploit LLM world knowledge to assign plausible acoustic distances between sound sources during training data synthesis (Chapter 3).

Chain-of-Thought prompting. Chain-of-Thought (CoT) prompting [68] is a technique that improves LLM performance on multi-step reasoning tasks by inducing the model to generate explicit intermediate reasoning steps before producing a final answer. Rather than asking "what is 17×23?", a CoT prompt encourages the model to first decompose the multiplication into partial products. Kojima et al. [31] showed that even the simple instruction "let's think step by step" elicits substantial reasoning

improvements. The key insight is that intermediate states scaffold the solution of hard problems by breaking them into tractable sub-problems.

2.7 Latent Space Representations

A *latent space* is the compressed, lower-dimensional representation learned by the encoder of a neural network. Variational Autoencoders (VAEs) [29] regularise the latent space to be approximately Gaussian, enabling controlled generation and interpolation. In the context of source separation, the encoder's latent representation should ideally concentrate energy at features corresponding to the desired target source while suppressing all others. Chapter 3 examines whether the latent representations of state-of-the-art conditioned separators actually achieve this goal, and how co-occurrence conditioning can improve latent-space organisation.

2.8 Related Work

This section surveys the broader landscape of source separation research. Per-chapter related work subsections in Chapters 3 and 4 provide additional context specific to each contribution.

2.8.1 Speech Separation

Speech separation has a long history predating deep learning, with classical approaches based on computational auditory scene analysis (CASA) [4, 63] and non-negative matrix factorisation (NMF) [62]. The deep learning era was catalysed by Deep Clustering [19], which learned speaker embeddings in a latent space and clustered them to assign time-frequency bins to speakers. Permutation Invariant Training (PIT) [32] resolved the label ambiguity problem by evaluating the loss over all possible output-target

assignments and selecting the minimum-cost permutation, enabling direct regression of separated signals. Conv-TasNet [35] brought waveform-domain processing to speech separation, achieving strong results without spectrogram representations. Dual-Path RNN [36] extended this by splitting input into shorter chunks and interleaving intra-chunk and inter-chunk RNNs for efficient long-sequence modelling, and more recently, Dual-Path Mamba [25] achieved competitive performance while reducing the quadratic attention cost of Transformer-based models.

2.8.2 Music Source Separation

MSS typically targets a fixed set of stems (vocals, drums, bass, other) from polyphonic recordings. The field has been anchored by the MUSDB18 benchmark [47]. Open-Unmix [54] provided an open-source deep learning baseline, while Spleeter [18] brought MSS to mainstream adoption. HTDemucs [50] introduced a hybrid temporal/spectral bi-U-Net augmented with Transformer self-attention layers, achieving state-of-the-art results. Conditioned approaches such as CUNet [39] and LaSAFT [8] unified multi-stem separation into a single network steered by instrument labels, and a few-shot framework [66] conditioned a U-Net via FiLM using audio examples of the target instrument, experimenting with different-recording examples, multi-source examples, and negative (co-occurring) examples – the latter being an early exploration of co-occurrence conditioning that Chapter 3 extends.

2.8.3 Universal Sound Separation

USS generalises beyond speech and music to arbitrary sound events. Kavalerov et al. [26] formalised the task and proposed initial benchmarks. MixIT [69] enabled unsupervised training at scale. Tzinis et al. [58] showed that adding sound classification embeddings improves universal separation. AudioSep [34] introduced language-queried USS via CLAP conditioning. SAM-Audio [52] brought the Segment Anything paradigm to audio,

unifying text, visual, and temporal span prompting within a diffusion transformer trained with flow matching. FlowSep [72] further advanced conditioned USS with rectified flow matching. Training-free approaches have recently emerged: Zang et al. [73] demonstrated multi-step inference gains with oracle access, and ZeroSep [22] leveraged diffusion denoising for zero-training separation.

2.8.4 Knowledge Transfer for Audio

A growing body of work transfers knowledge from pre-trained models to audio tasks, and this thesis draws on several such transfer mechanisms.

Cross-modal alignment. CLAP [12, 70] aligns audio and text embeddings via contrastive learning, enabling zero-shot audio understanding and text-queried separation.

Self-supervised representations for separation. Self-supervised models such as HuBERT [21] and Audio-MAE [24] learn rich contextual representations from unlabelled audio. These representations have been transferred to source separation: Pac-HuBERT [6] adapts HuBERT embeddings for MSS, and Zhao et al. [75] integrate Audio-MAE features into a USS backbone, both demonstrating that pre-trained contextual priors can substitute for task-specific supervision.

LLM world knowledge. In parallel, LLM world knowledge has been exploited for audio-related tasks. Our distance-aware mixing (Chapter 3) is, to our knowledge, the first work to transfer LLM commonsense reasoning into the audio mixing pipeline for training data synthesis.

Chapter 3

Knowledge-Guided Approaches to Sound Source Separation

This chapter presents two complementary approaches that inject external knowledge into the sound source separation pipeline. Section 3.1 introduces a distance-aware audio mixing strategy that leverages LLM world knowledge to create more realistic training data. Section 3.2 explores co-occurrence knowledge – information about non-target sounds present in a mixture – as supplementary conditioning for separation models, with a CLAP-based estimation procedure for practical deployment.

3.1 Distance-Aware Synthetic Audio Mixtures

3.1.1 Introduction

The mix-and-separate paradigm has been instrumental in scaling USS to large, unannotated audio corpora. By constructing MoMs from pairs of raw recordings and training models to recover the individual mixtures, MixIT [69] eliminated the need for isolated ground-truth sources. Subsequent work extended this framework with text conditioning [34] and richer network architectures.

However, a critical assumption has gone unquestioned in all prior work: that *random* mixing – selecting audio clips uniformly at random and summing them at equal or randomly-varied loudness – produces training data whose statistics are representative

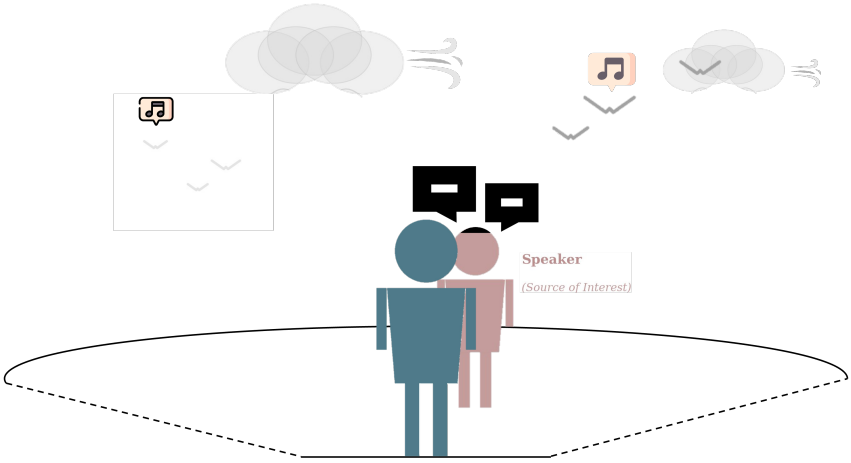


Figure 3.1: A real-world acoustic scene illustrating natural distance–loudness relationships. A nearby speaker dominates the mixture, while distant sound sources such as bird songs and wind are naturally quieter. Random mixing ignores these relationships; our distance-aware strategy models them explicitly.

of real-world acoustic scenes. We argue that this assumption is incorrect and that the resulting distributional mismatch is a fundamental obstacle to deploying USS models in practice.

Consider two audio clips: *whale vocalization* and *vehicle horn*. In a randomly-mixed training set, these might be combined at equal loudness, implying that they co-occur in the same acoustic environment at comparable distances. In reality, this combination is essentially impossible in any natural or urban setting. By training on such implausible mixtures, a model is implicitly learning that *any* two sounds can co-occur at *any* relative volume – a prior that is inconsistent with the structured statistics of real acoustic scenes.

This section introduces a distance-aware mixing strategy that uses an LLM to assign plausible relative loudness to each pair of audio sources, grounded in the physical concept of acoustic distance.

Figure 3.1 illustrates a typical real-world acoustic scene: a speaker at close range dominates the mixture, while environmental sounds such as bird songs and wind arrive from a distance and are naturally quieter. This distance–loudness relationship is what

random mixing fails to capture.

3.1.2 Acoustic Distance and Loudness

The relationship between distance and loudness in a free acoustic field follows the inverse-square law: doubling the distance from a sound source reduces its energy by 6 dB. In practice, room acoustics, directivity, and source properties complicate this relationship, but the qualitative intuition is robust: a sound that originates far from the listener is quieter than one that originates nearby.

We operationalise this physical intuition through three qualitative distance categories:

- **Far**: the candidate sound source is naturally distant from the base source, implying that it should be mixed at a lower volume.
- **Same**: the two sources naturally occur at comparable distances, implying approximately equal loudness.
- **Close**: the candidate source is naturally nearer than the base, implying that it should be mixed at a higher volume.

These categories are sufficient to capture the dominant variation in relative loudness without requiring precise numerical distance estimates, which would be difficult to obtain reliably from an LLM.

3.1.3 Methodology

LLM Distance Query

Let $\{(a_i, t_i)\}_{i=1}^{N}$ be a dataset of N audio-text pairs, where a_i is an audio clip and t_i is its associated caption. For each base clip a_i, a candidate clip $a_{i^\star}$ is drawn with $i^\star \neq i$.

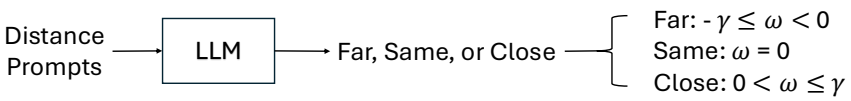


Figure 3.2: The LLM Mixing Module. Given captions t_i (base) and $t_i^\star$ (candidate), the LLM returns one of `far`, `same`, or `close`, which is mapped to a loudness offset ω sampled from the corresponding dB range $[-\gamma, 0)$, $\{0\}$, or $(0, \gamma]$.

We submit the following prompt to an LLM to determine the plausible distance of $t_{i^\star}$ relative to t_i:

> **System prompt:** You will be given two captions: (1) a caption describing a *base* audio; (2) a caption describing a *candidate* audio. Decide how loudly the candidate audio should be mixed with the base audio so the blend sounds the most natural. Return **exactly one** of the following tokens and nothing else: `far` (candidate sounds distant / noticeably quieter than the base), `same` (candidate is at roughly the same loudness as the base), `close` (candidate sounds very near / noticeably louder than the base).
>
> **Few-shot examples:**
>
> Base: *relaxed music* / Candidate: *human speech* → `close`
>
> Base: *a child laughs* / Candidate: *a woman speaks* → `same`
>
> Base: *birds chirp* / Candidate: *a branch sways in the wild* → `far`

Few-shot prompting guides both the output format and the quality of the LLM's decisions. If the LLM returns a token outside the three-class vocabulary, its output is encoded by a CLAP text encoder and the nearest of the three class embeddings (`far`, `same`, `close`) is selected by cosine similarity, guaranteeing a valid answer.

Energy-Ratio Scaling

Each distance label is mapped to a loudness offset ω in decibels:

$$\omega \sim \begin{cases} \mathcal{U}(-\gamma,\ 0) & \text{if label} = \texttt{far}, \\ 0 & \text{if label} = \texttt{same}, \\ \mathcal{U}(0,\ \gamma) & \text{if label} = \texttt{close}, \end{cases} \tag{3.1}$$

where $\gamma > 0$ is a hyperparameter controlling the dynamic range of the loudness adjustment. A scaling factor α is then computed to adjust the candidate energy relative to the base while applying the dB offset:

$$\alpha = \sqrt{\frac{E_1}{E_2} \cdot 10^{\omega/10}}, \qquad m_i = a_i + \alpha \cdot a_{i^\star}, \tag{3.2}$$

where $E_1 = \|a_i\|^2$ and $E_2 = \|a_{i^\star}\|^2$ are the signal energies of the base and candidate clips respectively, and m_i is the resulting MoM.

The scaling factor α jointly normalises the candidate energy to match the base (E_1/E_2) and applies the dB offset ($10^{\omega/10}$), all under a single square root since the multiplication by α scales the waveform amplitude. For example, if $\omega = +5\,\text{dB}$ and label is `close`, the candidate will be 5 dB *louder* than the base after energy normalisation.

The mixing procedure is executed on-the-fly for every mini-batch, so the model effectively sees a different MoM at each training iteration. This on-line data augmentation prevents overfitting to a fixed set of mixtures. Note that the entire mixing strategy is model-agnostic: it operates purely on audio-text pairs and can be used independently to generate a static training dataset for any downstream separation model, not only the ones evaluated in this work.

Implementation Details

We set $\gamma = 15\,\text{dB}$ in all experiments, covering a 30 dB dynamic range across the three distance categories, following the loudness range used in AudioSep [34]. We use Meta Llama 3.2 1B as the LLM, which is lightweight enough to run on the same GPU cluster used for model training. All experiments are performed on a server with 2 NVIDIA RTX 4090 GPUs (24 GB each) and two computational nodes with 3 and 1 NVIDIA A100 GPUs (40 GB each) respectively.

Text encoders in our experiments use pre-trained weights from CLAP [12], which aligns audio and text modalities via a contrastive objective and provides robust cross-modal embeddings for both the fallback token selection and the AudioSep conditioning.

3.1.4 Experimental Setup

Models

We evaluate the distance-aware mixing strategy on two architecturally distinct separation models to demonstrate that the benefits are not specific to any particular architecture:

- **TDCN++** [69]: a temporal dilated convolutional network operating in the waveform domain, trained with MixIT. This is an *unconditional* model that attempts to separate all sources in the input.
- **AudioSep** (ResUNet) [34]: a U-Net operating in the spectrogram domain, trained with text-conditioned supervision via CLAP. This is a *conditional* model guided by a natural-language query.

Training steps for each strategy and each model are kept identical to ensure a fair comparison. All other training hyperparameters follow the original implementations.

Training Data

Both models are trained on the combination of Clotho v2.1 [11] (development, validation, and evaluation splits) and FSD50K [14] (development and evaluation splits). Audio clips range from 0.3 to 30 seconds in length. All signals are resampled to 16 kHz and downmixed to mono. AudioSep randomly crops 10-second segments; MixIT randomly crops 5-second segments. The batch size is 18 for all experiments.

Evaluation Benchmarks

Real-world benchmarks (human evaluation). Three benchmarks are curated manually from publicly available datasets, each containing 100 audio mixtures:

- **Indoor+City**: sourced from the AudioCaps [28] test set, covering indoor scenes (speech in a café, water running, etc.) and urban sounds (car horns, traffic).
- **Outdoor+Wild**: also from AudioCaps, covering outdoor and natural environments (bird calls, thunderstorms, rivers).
- **Music**: sourced from the MUSIC dataset [74], pairing two of 11 instruments (flute, violin, trumpet, etc.) to form each mixture. This benchmark is more diverse and realistic than MUSDB18 [47] for evaluating general USS models.

Synthesised benchmark (automatic evaluation). We also evaluate on a benchmark created by randomly mixing two sound events from ESC50 [45], which consists of 5-second audio clips spanning 50 environmental sound categories. SI-SDR is used as the metric.

Human Evaluation Protocol

We develop a blinded Jupyter notebook-based voting application for human evaluation. Each trial presents voters with:

Table 3.1: Human preference (%) and SI-SDR (dB) for AudioSep and MixIT, comparing Random vs. Distance-aware mixing. Bold denotes the preferred or better result. 4 voters; overall agreement rate: 69.1%.

Model	Benchmark	Random (%)	Distance (%)
AudioSep	Indoor+City	28.7	**71.3**
	Outdoor+Wild	25.0	**75.0**
	Music	38.3	**61.7**
	ESC50 (SI-SDR, dB)	1.597	**3.029**
MixIT	Indoor+City	43.3	**56.7**
	Outdoor+Wild	47.7	**52.3**
	Music	47.0	**53.0**
	ESC50 (SI-SDR, dB)	**8.292**	4.961

1. The original mixture audio and its textual caption.

2. Two separated audio outputs, labelled *Option 1* and *Option 2* (randomly shuffled between the *Random* and *Distance* strategies).

3. Spectrogram visualisations of the original audio and both options, for cases where audio alone is insufficient.

Voters are instructed to select the output that better corresponds to the target caption, or to choose *Tie* if the quality is indistinguishable. A definitive vote contributes 1 point to the chosen option; a Tie vote contributes 0.5 points to each.

Inter-annotator agreement is calculated per pair of voters over all clips. For each clip, the agreement contribution is: 1 if both annotators choose the same method, 0.5 if one chooses a method and the other chooses Tie, and 0 if one chooses Random and the other chooses Distance. The overall agreement rate is the mean across all pairs.

3.1.5 Results and Analysis

Table 3.1 reports the full results. We discuss the findings for each model in turn.

AudioSep Results

The distance-aware strategy yields substantial and consistent improvement for AudioSep on all three real-world benchmarks. The gains are particularly pronounced for Outdoor+Wild (75% vs. 25%), where natural sound environments are characterised by strong distance-based loudness relationships – distant bird calls, nearby rustling leaves – that the distance-aware strategy models accurately. Indoor+City (71.3%) and Music (61.7%) also show strong preferences for the distance strategy.

On the synthesised ESC50 benchmark, AudioSep trained with the distance-aware strategy achieves an SI-SDR of 3.029 vs. 1.597, a gain of +1.432 dB. This improvement contrasts with the MixIT result below and reflects that text-conditioned models are less sensitive to the training-test distributional mismatch.

MixIT Results

MixIT also shows a preference for the distance strategy on all three real-world benchmarks, though the margins are smaller (56.7%, 52.3%, 53.0%). Crucially, MixIT with distance-aware mixing performs *worse* on the synthesised ESC50 benchmark (4.961 vs. 8.292), showing a degradation of −3.331 dB.

The Synthesis Benchmark Inversion

The degradation of MixIT on ESC50 merits careful discussion. The ESC50 benchmark is constructed by randomly pairing two clips from a 50-class environmental sound dataset at equal loudness – precisely the distribution that random training mimics. Distance-aware training biases the model toward realistic loudness relationships; when evaluated on a benchmark whose source pairs are acoustically implausible and uniformly loud, the distance-aware model is naturally at a disadvantage.

This inversion reveals a deeper issue: widely-used randomly-synthesised benchmarks may *systematically undervalue* methods that model real acoustic scenes. A model that

performs well in the real world – as measured by human evaluators – may appear worse than a randomly-trained baseline on synthetic benchmarks whose statistics are inconsistent with nature. We view this as an important finding for the USS field: benchmark construction choices can obscure genuine improvements in real-world performance.

Conditional models such as AudioSep are less affected because the text conditioning provides an additional axis along which to distinguish target from background, partially decoupling the model from training-distribution priors.

3.1.6 Related Work

Early deep learning approaches to single-channel source separation relied on supervised training with labelled and isolated audio sources, which severely limited training data scale and diversity. MixIT [69] introduced the mix-and-separate paradigm, enabling unsupervised training on arbitrary audio collections. AudioSep [34] added text conditioning, dramatically expanding the scope of USS. HTDemucs [50] addressed mixing realism in the *music* domain by using beat tracking and tempo estimation to construct melodically coherent multi-track mixtures for fine-tuning, achieving state-of-the-art MSS on MUSDB18 at the time of its release. Heterogeneous target speech separation [59] showed that real-world features such as microphone distance and speaker language are useful for conditional speech separation. Our work generalises the mixing-realism insight from music and speech to arbitrary sound categories, and provides a domain-agnostic mechanism via LLM world knowledge.

3.1.7 Limitations and Future Work

The primary practical limitation of the distance-aware strategy is the inference overhead introduced by querying an LLM for every training pair in every mini-batch. While we use a lightweight 1B-parameter model, this still adds wall-clock time compared to the

essentially cost-free random strategy. Future directions include: (i) distilling the LLM's distance decisions into a lightweight lookup table or a small classifier trained on a set of pre-computed labels; (ii) extending the distance categories beyond three to capture finer-grained loudness relationships; (iii) incorporating additional acoustic factors such as frequency-dependent attenuation (sounds at large distances lose high-frequency energy) and room acoustics.

3.2 Co-occurrence Knowledge for Conditioned Sound Separation

3.2.1 Introduction

Conditioned source separation models receive an explicit description of the desired output – a text caption, a class label, or an example audio – and use it to steer the separator toward the specified target. The implicit assumption is that knowing *what* you want is sufficient: if you specify "violin", the model should amplify violin features and suppress everything else.

However, this framing ignores a rich source of information that is available in the mixture itself: knowledge about *what else is present*. In a mixture of violin, piano, and speech, a separator targeting violin needs not only to know that violin is the desired output, but also that piano and speech are the *undesired* outputs. Providing both kinds of information – target and co-occurrence – gives the model a more complete specification of the separation problem. Prior work on few-shot MSS [66] has shown that conditioning on audio exemplars of co-occurring sources can improve separation, supporting this intuition.

This section presents an exploratory investigation into whether co-occurrence knowledge can be estimated and injected into the encoder of USS models without

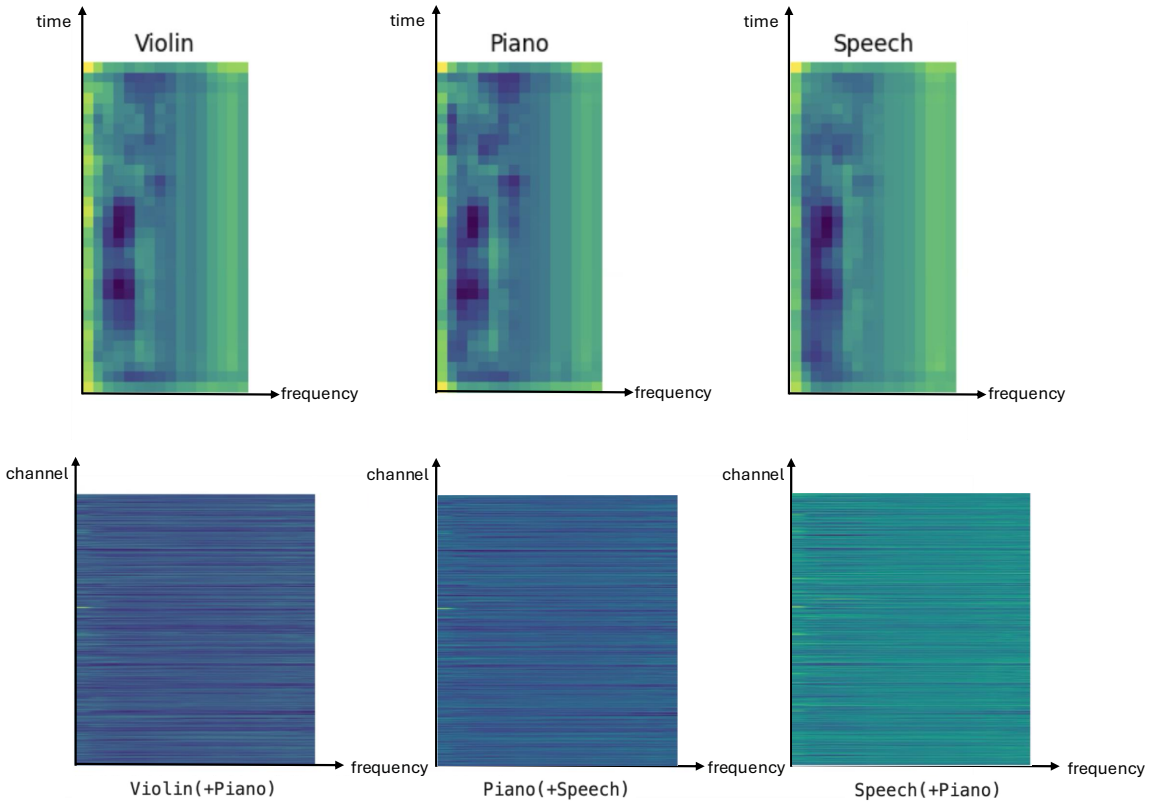


Figure 3.3: Visualisation of the AudioSep encoder's latent space for the same audio mixture (Violin + Piano + Speech) conditioned on three different targets. The first row shows the mean over the channel axis; the second row shows the mean over the time axis. Despite different conditioning signals, the latent feature maps show limited differentiation across targets, suggesting that target-only encoder conditioning is insufficient to shape the latent representation.

requiring ground-truth source labels. We propose a CLAP-based estimation procedure that approximates co-occurrence embeddings from the mixture audio alone and evaluate its effect on AudioSep across six benchmarks. The estimator improves separation on five of six benchmarks, with one exception that we analyse in detail.

3.2.2 Motivation: Latent Space Analysis

To motivate the co-occurrence hypothesis, we inspect the latent representations of a pre-trained AudioSep model when applied to the same audio mixture with three different conditioning targets: Violin, Piano, and Speech (Figure 3.3). If target conditioning were

shaping the encoder effectively, we would expect visually distinct latent representations for each target.

Instead, the latent feature maps show surprisingly little differentiation across targets. LaSAFT [8] reported a related finding through ablation: conditioning the encoder on the *target* identity actually *confuses* the latent space, degrading separation performance relative to decoder-only conditioning. We hypothesise that this occurs because the encoder's primary role is lossy compression of the input, not target-specific feature selection – a view that our independent observation supports and that motivates the alternative conditioning strategy explored below.

This raises a natural question: if the encoder is not the right place for target information, could it instead benefit from a different kind of signal? We explore the idea that the encoder should receive *co-occurrence* information – a description of what is *not* wanted – rather than target information. The intuition is that co-occurrence knowledge specifies what to filter out during compression, which aligns more naturally with the encoder's role than target preservation does. Whether this idea holds in practice is the subject of the experiments below.

3.2.3 Methodology

Co-occurrence Conditioning in USS

Building on the AudioSep training framework, a MoM M_1 is constructed from two mixtures A_1 and A_2. The target caption C_1 (from A_1) and co-occurrence caption C_2 (from A_2) are both available during training. We encode C_2 with the CLAP text encoder to obtain the co-occurrence embedding $c_2 = \phi_{\text{text}}(C_2)$, which is injected into the encoder via FiLM modulation alongside the standard decoder-level target conditioning.

At inference time, however, C_2 is unavailable. This motivates the estimation

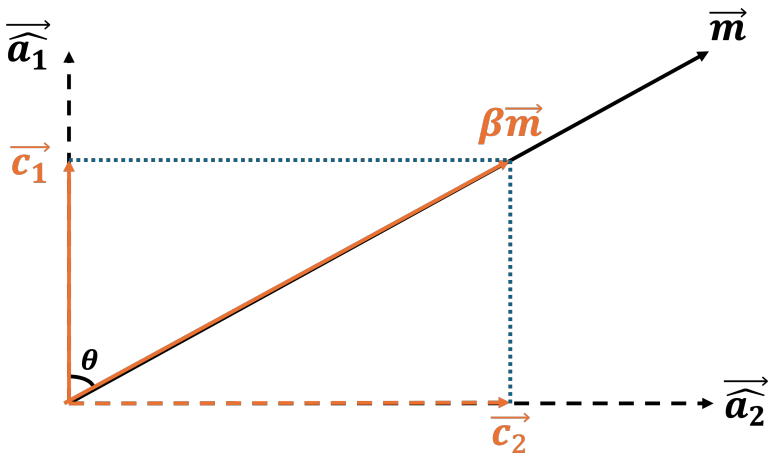


Figure 3.4: Vector-space view of CLAP-based co-occurrence estimation. Under the CLAP alignment objective, the mixture audio embedding m lies approximately along the sum of individual source embeddings $a_1 + a_2$. Given m and the target text embedding c_1 (cross-modally aligned with a_1), one can estimate the co-occurrence text embedding as $\hat{c}_2 = \beta \cdot m - c_1$ by scaling the mixture onto the text space. Similarly, scaling c_1 onto the audio space yields the co-occurrence audio embedding $\hat{a}_2 = m - \alpha \cdot c_1$, which is the estimator evaluated in this work.

procedure below.

CLAP-Based Co-occurrence Estimation

In practical deployment, C_2 is unavailable; only the mixture audio M and the target text C_1 are observed. We propose to estimate the co-occurrence embedding from these two quantities.

We adopt the working hypothesis that, under the CLAP alignment objective, the audio embeddings of sources are approximately additive: $\phi_{\text{audio}}(M) \approx \phi_{\text{audio}}(A_1) + \phi_{\text{audio}}(A_2)$. This property is not explicitly established in the CLAP literature [12] but is a plausible consequence of the contrastive training objective; we treat it as an empirical assumption whose validity is reflected in the downstream separation results. Let $m = \phi_{\text{audio}}(M)$, $a_1 = \phi_{\text{audio}}(A_1)$, $a_2 = \phi_{\text{audio}}(A_2)$, and $c_1 = \phi_{\text{text}}(C_1)$ with $c_1 \approx a_1$ by cross-modal alignment. We derive an estimator of c_2 (Figure 3.4).

The simplest approach would be the raw difference $m - c_1$, but this is potentially

noisy because m is an audio embedding and c_1 is a text embedding in a shared but not identical space. To bridge the modality gap, one can scale the mixture onto the text space, yielding the estimated co-occurrence caption $\hat{c}_2 = \beta \cdot m - c_1$. Alternatively, one can scale the target text embedding onto the audio space:

$$\hat{a}_2 = m - \alpha \cdot c_1, \tag{3.3}$$

where $\alpha = \text{cosim}(m, c_1)$ maps the target text embedding onto the mixture audio embedding scale. This estimator approximates the audio embedding of the non-target sources. We evaluate $\hat{a}_2$ in the experiments below; it replaces the oracle c_2 as the co-occurrence conditioning signal injected into the encoder.

3.2.4 Experimental Setup

Model

We use AudioSep [34] extended to accept an additional co-occurrence conditioning vector at the encoder via FiLM modulation. The base target conditioning at the decoder is preserved; the only architectural change is the addition of the co-occurrence input to the encoder.

Datasets

We evaluate on six USS benchmarks: AudioSet [15], AudioCaps [28], VGGSound [5], MUSIC [74], ESC50 [45], and Clotho [11].

3.2.5 Results

Table 3.2 reports AudioSep results with the $\hat{a}_2$ co-occurrence estimator. The estimator improves performance on five of six benchmarks over target-only conditioning, with the largest gain on ESC50 (+0.86 SDR, +0.99 SI-SDR). The exception is MUSIC,

Table 3.2: AudioSep USS results (SDR / SI-SDR) comparing target-only conditioning against the $\hat{a}_2$ co-occurrence estimator (Equation 3.3). Bold denotes the better result per benchmark.

Benchmark	Target-Only	$\hat{a}_2$
AudioSet	4.829 / 3.046	**5.368 / 3.562**
AudioCaps	5.454 / 3.995	**5.777 / 4.424**
VGGSound	7.270 / 5.933	**7.438 / 6.130**
MUSIC	**3.847 / 1.499**	3.331 / 0.980
ESC50	5.984 / 4.494	**6.844 / 5.486**
Clotho	4.377 / 2.479	**4.502 / 2.677**

where performance drops (-0.52 SDR, -0.52 SI-SDR). This may reflect the particular nature of the MUSIC dataset (paired musical instruments with balanced energy), where the mixture embedding m and target text embedding c_1 occupy similar regions of the CLAP space and are difficult to disentangle reliably.

These results suggest that co-occurrence information can benefit USS in most settings, but that the estimator is not robust across all source types. Alternative scaling strategies – for instance, projecting the mixture embedding onto the text space before subtraction rather than scaling the text embedding onto the audio space – may address this limitation and are left to future work.

3.2.6 Related Work

Conditioned separation. CUNet [39] introduced FiLM conditioning for multi-stem MSS. LaSAFT [8] extended this with latent source attention and showed that decoder conditioning is preferable. A few-shot MSS framework [66] conditioned on audio exemplars including co-occurring examples, but required example audios at inference time. AudioSep [34] applies CLAP text conditioning at both encoder and decoder for open-domain USS. FlowSep [72] brings rectified flow matching to conditioned USS, improving separation quality in the generative setting.

Context-aware separation. Several architectures implicitly capture co-occurrence structure. Dual-Path RNN [36] models long-range context via bidirectional RNN over chunked sequences. Pac-HuBERT [6] leverages HuBERT [21] context-rich SSL embeddings to enhance MSS. A-MAE [24] integrates audio masked autoencoders as contextual priors in a USS backbone. Dual-Path Mamba [25] achieves efficient long-context modelling for separation. Our work differs from all of these in that it explicitly and directly injects co-occurrence information as a conditioning signal, rather than implicitly capturing context through architecture design.

3.2.7 Limitations and Future Work

Several open questions remain: (i) Do the improvements from the $\hat{a}_2$ estimator hold across architectures beyond AudioSep? (ii) Can alternative scaling strategies further improve estimation accuracy? (iii) How should co-occurrence conditioning interact with target conditioning to avoid disrupting learned representations? Future work will also explore whether the CLAP embedding space's assumed additivity ($m \approx a_1 + a_2$) can be enforced via fine-tuning to improve estimation accuracy.

Co-occurrence-aware mixing. It is worth distinguishing the co-occurrence *conditioning* investigated in Section 3.2 from a complementary but distinct idea: co-occurrence-aware *mixing*, i.e., selecting which audio clips to pair during MoM construction based on how likely those sound events are to co-occur in practice. Section 3.1 addresses the *relative loudness* of a pair once selected, but the selection itself remains random. A natural extension is to also make the pairing realistic. One possible approach is to leverage pre-trained audio tagging models such as PANNs [33]: frame-level sound event detection yields per-clip event probabilities, and pairwise cosine similarities between these probability vectors can be accumulated across a large, label-balanced corpus to produce a co-occurrence frequency distribution. At training time, candidate

clips are then sampled according to this distribution rather than uniformly at random, yielding MoMs whose source pairings reflect genuine environmental statistics. An alternative route is to prompt an LLM to generate plausible co-occurring sound sources given a base caption. Both routes are complementary to the distance-aware strategy presented in this chapter and could further close the gap between synthetic and real-world mixture statistics. We leave this investigation to future work.

Chapter 4

Chain-of-Inference: A Training-Free Incremental Framework

4.1 Introduction

As discussed in Chapter 1, current USS models process a mixture in a single forward pass, yet complex real-world mixtures may exceed what a single inference step can resolve. Understanding *why* iterative refinement helps requires a perceptual lens. A key insight is that the human auditory system is particularly sensitive to sudden changes in the acoustic scene – transient onsets, abrupt level shifts, and novel events that capture involuntary attention [13, 27, 42]. This sensitivity suggests that a framework which progressively sharpens the contrast between target and background produces increasingly salient waveform changes at each step, effectively transforming a hard separation problem into a sequence of easier ones.

This observation motivates a connection to recent advances in language modelling. The chain-of-thought paradigm [31, 68] has proven to be one of the most effective techniques for improving LLM performance on complex reasoning tasks. By generating explicit intermediate steps rather than jumping directly to the final answer, a model can decompose a hard problem into a sequence of easier sub-problems, each within its individual competence. The gains are well-documented across arithmetic, symbolic reasoning, commonsense inference, and multi-step planning.

We observe that the USS problem shares a structural analogy with these hard reasoning tasks: separating a complex multi-source mixture in a single step is analogous to solving a multi-step arithmetic problem in a single jump. In both cases, the difficulty exceeds the capacity of any single-step computation, and decomposition into intermediate states is a natural remedy.

However, transferring CoT to audio is non-trivial. In language, CoT works because the model's context window can accumulate intermediate results; each token generation is conditioned on the full sequence of prior reasoning steps. Audio separation models lack this sequential input structure: they take a mixture as input and a query, and produce a separated output. There is no natural mechanism for feeding intermediate results back into the model in a principled way.

This chapter introduces Chain-of-Inference (CoI), a formulation of multi-step inference for USS that addresses this gap. The key idea is to construct a sequence of *auxiliary mixtures* that gradually transition from the original (hard) problem to an easier one, using the model's own output at each step to shape the input at the next.

4.2 The Vector-Geometric View of Sound Separation

Before presenting the CoI formulation, it is useful to develop a geometric intuition for what "progress" in separation looks like. Consider vectors in the signal space $\mathbb{R}^T$. The original mixture $\vec{m}_0$ can be decomposed as:

$$\vec{m}_0 = \vec{g} + \vec{n}, \tag{4.1}$$

where $\vec{g}$ is the target source and $\vec{n}$ is the sum of all non-target sources. An ideal separator would produce $\hat{s} = \vec{g}$, i.e., the separated output would be exactly the target.

In practice, the separator produces an estimate $\vec{s}_t$ that lies between $\vec{m}_0$ (no separation) and $\vec{g}$ (perfect separation) in the signal space.

The angle θ_t between $\vec{m}_0$ and $\vec{s}_t$ captures this geometric progress:

- When $\theta_t \to 0$: $\vec{s}_t \approx \vec{m}_0$ (no separation has occurred).
- As θ_t increases: $\vec{s}_t$ rotates away from $\vec{m}_0$ toward $\vec{g}$ (separation is making progress).
- Theoretically $\theta_t < 90°$, since $\vec{m}_0$ and $\vec{s}_t$ share the common component $\vec{g}$.

Working assumption. The geometric picture sketched in Figure 4.1 is not automatic: because θ_t is an unsigned angle, a non-zero θ_t only indicates that $\vec{s}_t$ has rotated *somewhere* away from $\vec{m}_0$, not necessarily in the useful direction. CoI therefore relies on a base-model competence assumption: given $\vec{m}_0$, the separator f rotates its output *toward* the target $\vec{g}$ rather than in an arbitrary direction. Concretely, we assume (i) $\theta_0 > 0$, so the first forward pass produces at least some separation ($\vec{s}_0 \not\approx \vec{m}_0$); and (ii) $\vec{s}_t$ lies in the half-space containing $\vec{g}$, i.e., $\vec{s}_t \cdot \vec{g} > 0$, so that a larger θ_t corresponds to being *closer* to the target rather than further from it. Together these conditions guarantee the $\theta_t < 90°$ bound listed in the third bullet and imply that θ_t is expected to increase monotonically until the model exhausts its capacity on the current input; the stopping criterion in Section 4.3.3 detects the first violation of this monotonicity. We take this assumption as given for well-pretrained separators such as AudioSep and SAM-Audio, which are trained on large-scale supervised USS data; CoI offers no guarantees when it fails (e.g., for a mis-specified model that outputs content in the direction of $-\vec{g}$ or of the non-target $\vec{n}$).

This geometric view motivates two design choices in CoI: the use of cosine similarity as the ratio schedule, and the cosine-similarity stopping criterion.

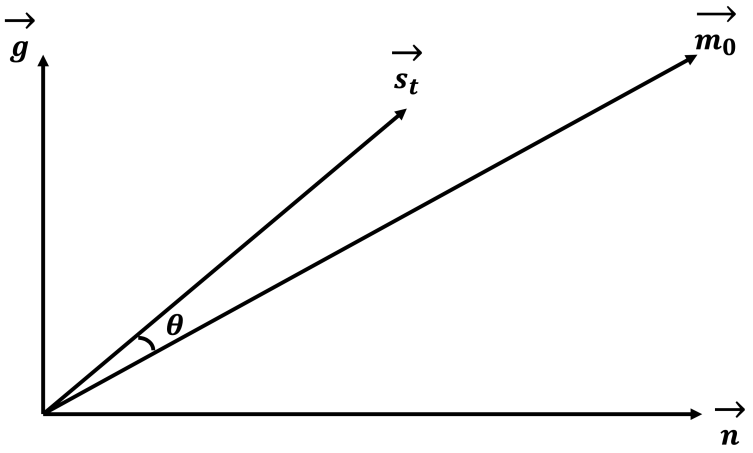


Figure 4.1: Vector-space view of sound separation. The original mixture $\vec{m}_0$ decomposes into target $\vec{g}$ and non-target $\vec{n}$. The separator's estimate $\vec{s}_t$ rotates away from $\vec{m}_0$ toward $\vec{g}$; the angle θ_t measures separation progress.

4.3 CoI Formulation

4.3.1 Auxiliary Mixture Construction

Given the original mixture m_0 and the separated output s_{t-1} from the previous step, we construct the auxiliary mixture for step t as:

$$m_t = r_t\, m_0 + (1 - r_t)\, s_{t-1}, \tag{4.2}$$

where $r_t \in [0, 1]$ is the mixing ratio. This construction has an intuitive interpretation: m_t blends the original problem m_0 (which contains all the sources) with the previous estimate s_{t-1} (which should be closer to the target), progressively reducing the proportion of non-target sources in the input. When $r_t = 1$, $m_t = m_0$ (no progress used); when $r_t \to 0$, $m_t \to s_{t-1}$ (problem is essentially solved).

4.3.2 Cosine Similarity Ratio Schedule

The most important design choice is the ratio r_t. We set:

$$r_t = \text{cosim}(\vec{m}_0,\, \vec{s}_{t-1}) = \frac{\vec{m}_0 \cdot \vec{s}_{t-1}}{\|\vec{m}_0\|\,\|\vec{s}_{t-1}\|} = \cos\theta_{t-1}. \tag{4.3}$$

The rationale is geometric: the cosine similarity $\cos\theta_{t-1}$ naturally maps to $(0, 1)$ in this setting (since the mixture and any separated output share the target component, they are never perpendicular). Moreover, as separation progresses (θ_{t-1} increases), r_t decreases, which is exactly the desired behaviour: the better the previous estimate, the less we need to reintroduce the original problem.

Adaptive difficulty annealing. The cosine ratio schedule adapts to the *difficulty* of the input. For an easy mixture (where the target is already well-separated after one step), θ_0 is large, r_1 is small, and subsequent steps introduce little new information – the algorithm converges quickly. For a hard mixture (where the target is deeply entangled with non-target sources), θ_0 is small, r_1 is large, and the algorithm iterates more aggressively.

4.3.3 Stopping Criterion

Iteration terminates when the cosine similarity between $\vec{m}_0$ and $\vec{s}_t$ ceases to increase:

$$\text{Stop if} \quad r_t < r_{t-1}, \tag{4.4}$$

i.e., when $\cos\theta_t > \cos\theta_{t-1}$, meaning θ has started to decrease. This indicates that the model has reached its capacity on the current input and further iterations would not improve separation. The stopping criterion is computed with zero additional forward passes, making it computationally free.

Algorithm 1 Chain-of-Inference (CoI)

Require: Mixture m_0, separation model $f(\cdot, \cdot)$, text prompt q

$s_0 \leftarrow f(m_0,\ q)$
$r_0 \leftarrow 1$
for $t = 1, 2, 3, \ldots$ **do**
  $r_t \leftarrow \text{cosim}(\vec{m}_0,\ \vec{s}_{t-1})$
  $m_t \leftarrow r_t\, m_0 + (1 - r_t)\, s_{t-1}$
  $s_t \leftarrow f(m_t,\ q)$
  if $r_t > r_{t-1}$ **then**
    break
  end if
end for
return s_t

Algorithm 1 presents the complete CoI procedure.

4.4 Experimental Setup

4.4.1 Datasets

- **Dog Vocal Separation (DVS)** [2]: a challenge dataset from the Barkopedia competition at IJCAI 2025, containing dog vocalisations mixed with environmental background sounds. This dataset lies partially within AudioSep's training distribution.

- **VCTK-DEMAND** [60] (Speech Enhancement, SE): clean speech from the VCTK corpus mixed with noise from the DEMAND dataset. Speech enhancement is a well-studied task with clear perceptual objectives.

- **MUSDB18** [47] (Music Source Separation): 150 multi-track music recordings in four stems (vocals, drums, bass, other). We evaluate three stems (vocals, drums, bass) and exclude "other" since target-extraction models are primarily trained to disentangle text-specified sources from the rest. MUSDB18 is largely out-of-distribution for both AudioSep and SAM-Audio, making it a challenging

test of generalisation.

4.4.2 Models

- **AudioSep** [34]: a language-queried generalist USS model trained on broad open-domain audio datasets. Used as a frozen checkpoint.
- **SAM-Audio** [52]: a diffusion-based audio separator extending the Segment Anything paradigm. Used as a frozen checkpoint.

4.4.3 Baselines

The baseline for each model is the standard one-step inference: $s = f(m_0, q)$. No fine-tuning or additional training is performed for the CoI condition.

4.4.4 Text Prompts

Following the NP/VP prompting effectiveness findings of Shi et al. [52], we use the following fixed prompts for SAM-Audio: "dog barking" (DVS), "human speaking" (SE), "singer delivering" (MUSDB18 vocals), "drums playing" (MUSDB18 drums), "bass slapping" (MUSDB18 bass). The same prompts are used for AudioSep to ensure a fair comparison.

4.5 Results

4.5.1 AudioSep Results

CoI improves AudioSep across all five tasks. Note that AudioSep was not trained on MUSDB18, and the strongly negative SDR values for those tasks (e.g., $-75\,\mathrm{dB}$ for vocals) indicate that the one-step baseline is essentially non-functional on this out-of-distribution data; improvements in this regime should be interpreted cautiously.

Table 4.1: Separation quality under one-step and multi-step (CoI) inference for AudioSep and SAM-Audio. Bold denotes the better result per task and model. AudioSep uses SDR and SI-SDR (dB, ↑); SAM-Audio uses SAJ Overall ($\in [1, 5]$, ↑). Avg. Steps reflects the mean number of forward passes per sample under CoI.

		AudioSep			SAM-Audio	
Task	Method	SDR	SI-SDR	Avg. Steps	SAJ Ovr.	Avg. Steps
DVS	One-step	10.075	8.713	1.00	3.742	1.00
	Multi-step	**10.226**	**8.785**	3.31	**3.764**	2.40
SE	One-step	17.320	17.237	1.00	**3.620**	1.00
	Multi-step	**17.393**	**17.294**	3.43	3.369	2.65
MUSDB18 Vocals	One-step	−75.578	−77.166	1.00	3.216	1.00
	Multi-step	**−75.422**	**−76.863**	4.68	**3.342**	2.63
MUSDB18 Drums	One-step	−27.760	−29.590	1.00	3.789	1.00
	Multi-step	**−27.748**	**−29.555**	3.24	**3.975**	2.80
MUSDB18 Bass	One-step	−56.473	−59.226	1.00	2.944	1.00
	Multi-step	**−54.338**	**−58.735**	5.34	**3.407**	3.13

The largest absolute gains are on MUSDB18 bass (+2.14 SDR, +0.49 SI-SDR, 5.34 average steps). On in-distribution tasks where the baseline is already strong, CoI still yields consistent improvements. Figure 4.2 illustrates the progressive refinement on a representative sample.

Gains on DVS (+0.15 SDR) and SE (+0.07 SDR) are smaller but consistent, reflecting that these tasks are closer to AudioSep's training distribution and therefore the one-step baseline is already relatively strong. MUSDB18 drums shows the smallest improvement (+0.01 SDR, 3.24 steps), indicating near-capacity performance for drums in one step.

4.5.2 SAM-Audio Results

CoI improves SAM-Audio on four of five tasks. The largest gains are on MUSDB18 bass (+0.46 SAJ) and drums (+0.19 SAJ). The exception is SE, where the one-step baseline (3.620) outperforms multi-step (3.369). This is consistent with a failure mode of this specific benchmark reported independently in the literature [71, 73]: SAM-

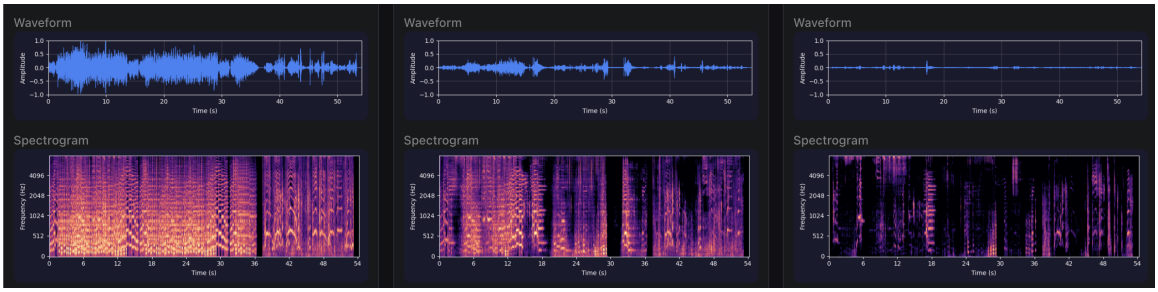


Figure 4.2: Waveforms and spectrograms for a dog-barking sample mixed with human speech and background music (AudioSep). Left: original mixture. Middle: one-step output. Right: multi-step (CoI) output. CoI produces a cleaner separation with reduced residual interference.

Audio's strong speech alignment means the first-pass output is already near-optimal, and reintroducing the original mixture in subsequent steps introduces interference rather than refinement. This points to *adaptive stopping* as a productive direction: the stopping criterion should detect not only when the cosine similarity stops improving, but also when the absolute quality of the output has already reached the model's ceiling for a given source type.

4.5.3 Computational Cost

The average step counts in Table 4.1 confirm that the cosine-similarity stopping criterion adapts the number of iterations to task difficulty. AudioSep requires 3.24–5.34 steps; SAM-Audio requires 2.40–3.13 steps. Bass consistently requires the most iterations across both models, consistent with its being the most out-of-distribution stem. The stopping criterion incurs no additional forward passes and adapts the total computation to the input, avoiding unnecessary iterations on easier tasks.

4.6 Related Work

Training-free separation. Zang et al. [73] demonstrated that multi-step inference can yield a "free lunch" in separation quality. Their method shares the same high-level

structure as CoI: at each step, the original mixture is blended with the model's own prior output to form an intermediate input. The key difference lies in how the blending ratio is determined. Zang et al. select the ratio by maximising a separation metric at each step, which requires evaluating the metric for multiple candidate ratios. CoI instead derives the ratio directly from the cosine similarity between the mixture and the current output (Equation 4.3), which is computed in closed form with no additional forward passes. This makes CoI's ratio schedule computationally free and fully deterministic, whereas the metric-maximisation approach incurs additional evaluation cost per step.

ZeroSep [22] proposes training-free USS via the iterative denoising process of diffusion-based generative models. While effective, this approach is architecturally specific to diffusion models and cannot be applied to discriminative models such as AudioSep. CoI places no architectural constraints on the underlying model.

Universal Sound Separation models. Conv-TasNet [35] established the one-step inference paradigm for waveform-domain separation. AudioSep [34] and SAM-Audio [52] are the two primary publicly available checkpoints for large-scale USS. Both operate under one-step inference; CoI is directly applicable to both without modification.

4.7 Limitations and Future Work

CoI's primary limitation is the increased inference cost relative to one-step processing: each sample requires multiple forward passes through potentially large models. In latency-sensitive applications (e.g., real-time hearing aids), this overhead may be prohibitive. Future directions include distilling CoI's multi-step reasoning into a single forward pass via sequence-to-sequence learning, and developing learned adaptive stopping strategies that terminate earlier without sacrificing quality. The precise text prompt is also increasingly consequential in the multi-step setting – each step amplifies

the model's capacity in the direction of the query – motivating joint optimisation of prompts and step count. A further promising direction is to develop training frameworks that explicitly encourage the model to attend to sudden changes – transient onsets and abrupt level shifts – in the waveform. Incorporating such saliency-aware objectives into the training process could yield models whose outputs are not only objectively better-separated but also perceptually more natural for downstream assistive-listening applications.

Connection to perceptual mechanism and distance-aware training. As noted in the introduction, CoI's progressive strengthening of the target source aligns with the perceptual mechanisms that hearing aids seek to exploit. An open question is whether models trained on the distance-aware synthetic mixtures introduced in Chapter 3 would benefit more or less from CoI at inference time. Distance-aware mixing produces training data with more realistic amplitude contrasts and onset profiles – effectively preserving the sudden changes that characterise real-world source interactions. A model trained on such data may already produce sharper first-step estimates, allowing CoI to converge in fewer steps or achieve higher final quality. We leave this investigation to future work.

Stopping criterion refinement. Preliminary experiments with a patience counter – allowing the algorithm to continue for several steps after the first decrease in r_t rather than stopping immediately – showed further improvements in separation quality at the cost of higher average step counts. This suggests that the current single-step stopping criterion may be overly conservative, and that more sophisticated termination strategies (e.g., patience-based or learned criteria) could unlock additional gains. Characterising the trade-off between step budget and output quality under different stopping policies is a promising direction for future work.

Chapter 5

System Demonstration

5.1 Introduction

An important complement to algorithmic contributions is making them accessible and tangible to the research community. To this end, we develop and deploy an interactive online demonstration system that showcases the Chain-of-Inference framework (Chapter 4) through a web-based platform. The system provides two interaction modes – a sample page with pre-computed results and an upload page for on-demand separation – allowing users to experience the perceptual impact of multi-step inference on arbitrary audio, side by side with the one-step baseline, without requiring any local installation or GPU resources. Integration of the distance-aware synthesis and co-occurrence conditioning contributions from Chapter 3 is planned for future work.

The demonstration system is publicly available at `https://redgiant.uta.edu/sound_separation`.

5.2 System Architecture

The system is built entirely in Python using Reflex [48] (v0.8.26), a full-stack web framework that compiles the UI to the browser without requiring a separate JavaScript codebase. This single-language design simplifies deployment and maintenance: the same Python codebase defines the user interface, handles file uploads, orchestrates

GPU inference, and renders results. The server runs on a machine equipped with two NVIDIA RTX 4090 GPUs (24 GB VRAM each), providing sufficient memory to host multiple separation models simultaneously.

The audio processing pipeline is built on the following components:

- **librosa** [38] for audio loading, resampling, and computing the inter-vector angle θ that drives the CoI ratio schedule (Equation 4.3).
- **soundfile** [3] for reading uploaded files in various formats and encoding separated outputs as WAV.
- **matplotlib** [37] for rendering waveform and spectrogram visualisations displayed alongside each audio player.
- **NumPy** [16] for all signal arithmetic, energy computation, and cosine-similarity calculations.

5.3 Interaction Modes

The system provides two complementary interaction modes (Figure 5.1).

5.3.1 Sample Page

The sample page presents a curated collection of audio examples spanning the three benchmark categories used throughout this thesis: indoor/city, outdoor/wild, and music. For each sample, the interface displays:

1. The **original mixture** audio with its textual caption and spectrogram.
2. The **one-step baseline** output (standard AudioSep inference).
3. The **Chain-of-Inference** multi-step output, along with the number of steps taken and the final cosine-similarity ratio.

All outputs are pre-computed and cached, allowing instant playback without GPU latency. Users can listen to all three side by side and visually compare spectrograms to observe how CoI progressively removes non-target energy from the output.

5.3.2 Upload Page

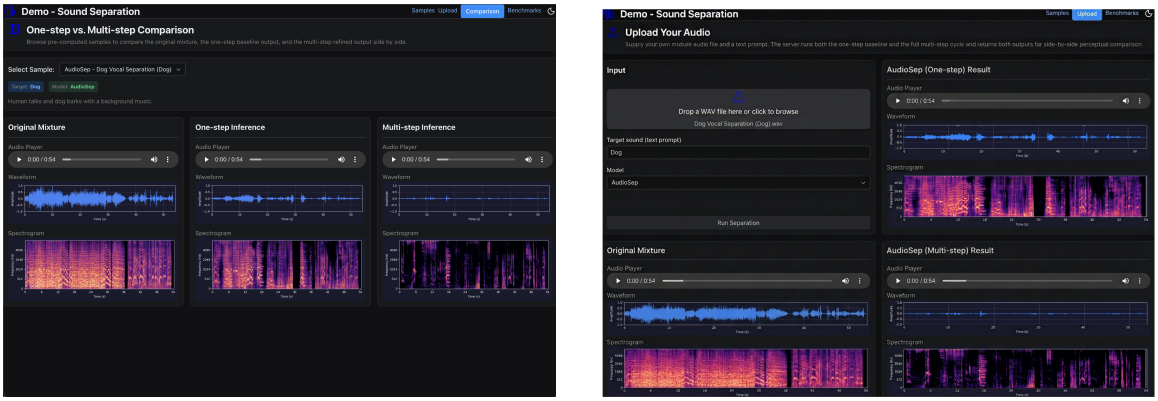
The upload page provides a fully interactive experience. Users supply their own mixture audio file (WAV or MP3, up to about 60 seconds due to the GPU capacity) and specify a separation target via a free-form text query (e.g., "the sound of a violin", "speech"). Upon submission, the server executes the full separation pipeline:

1. The mixture is loaded, resampled to the target model's native sample rate (32 kHz for AudioSep, 48 kHz for SAM-Audio), and downmixed to mono.
2. **One-step separation**: AudioSep processes the mixture with the text query in a single forward pass.
3. **Multi-step separation (CoI)**: The iterative CoI procedure (Algorithm 1) is executed until the stopping criterion is met.
4. Both outputs are returned with waveform and spectrogram visualisations for side-by-side perceptual comparison.

The upload page demonstrates that CoI works on arbitrary user-supplied audio, not just the curated benchmarks, reinforcing the generality of the framework.

5.4 Deployment and Availability

The demonstration system has been online since February 2026 and has been used by reviewers, collaborators, and members of the research community. The system is hosted at `https://redgiant.uta.edu/sound_separation` and is freely accessible without authentication. Figure 5.1 shows screenshots of the two interaction modes.



(a) Sample page (b) Upload page

Figure 5.1: Screenshots of the online demonstration system. *Left*: The sample page presents curated audio examples with pre-computed outputs from multiple methods for direct listening. *Right*: The upload page accepts a user-supplied audio file and a free-form text query, runs separation with multiple methods on-demand, and returns outputs side by side for perceptual comparison.

Chapter 6

Discussion and Conclusion

6.1 Summary of Contributions

This thesis has addressed Sound Separation from three complementary perspectives, each targeting a distinct bottleneck in the path from raw audio to deployed separation system.

Data quality (Chapter 3). The first contribution demonstrated that the distribution of randomly-synthesised training mixtures diverges substantially from real-world acoustic statistics, and that this gap can be reduced by incorporating LLM world knowledge about acoustic distance. The practical implication is that separation models trained on distance-aware MoMs are consistently preferred by human evaluators over their randomly-trained counterparts on real-world audio, while performing comparably or better on standard benchmarks with conditional models.

Inference strategy (Chapter 4). The second contribution showed that the one-step inference paradigm, universally adopted in USS, leaves substantial model capacity underutilised on complex mixtures. Chain-of-Inference provides a principled, training-free mechanism to exploit this capacity by iteratively decomposing the separation problem, governed by cosine-similarity geometry in the signal space. The online demonstration system makes this improvement accessible and tangible to the research community.

Conditioning richness (Chapter 3). The third contribution opened the question of co-occurrence conditioning: can knowledge about what is *not* wanted in the output improve separation beyond what target-only conditioning achieves? An exploratory investigation showed that a CLAP-based co-occurrence estimator ($\hat{a}_2$) improves AudioSep on five of six benchmarks, providing initial evidence that such conditioning can be estimated without oracle source labels using the shared embedding space of CLAP. The MUSIC benchmark degrades, indicating that further work is needed to make co-occurrence conditioning robust across all source types.

6.2 Cross-Cutting Themes

Improving USS without new annotations. A recurring theme across all three contributions is that substantial gains in separation quality can be achieved without collecting new annotated data or isolated ground-truth sources. Chapter 3 repurposes LLM world knowledge and CLAP embeddings – resources trained for general-purpose language and audio-language alignment – to improve training data realism and model conditioning. Chapter 4 requires no additional data or training at all: it extracts latent capacity from frozen models through a principled inference-time procedure. In each case, the key insight is to identify and exploit underutilised resources – whether external (LLMs, CLAP) or internal (model capacity) – rather than scaling up supervision.

Training-free vs. training-time improvements. The three contributions operate at different points in the model lifecycle. Distance-aware mixing is a training-time intervention: although the initial LLM inference adds a one-time preprocessing cost, the resulting distance labels can be cached and reused across all subsequent training runs, so the amortised overhead is negligible once the preprocessing is complete. The benefits are then baked into the model weights. CoI is a pure inference-time intervention; it improves any frozen checkpoint with no retraining. Co-occurrence conditioning is both:

the model is fine-tuned to accept co-occurrence input, but the estimation procedure is applied at inference time. Together, they illustrate that improvement opportunities exist throughout the model lifecycle.

Evaluation methodology. Chapters 3 and 4 both grapple with the inadequacy of randomly-synthesised benchmarks for evaluating models that target real-world performance. This is a broader methodological concern for the USS field: as models improve at modelling real acoustic scenes, they may appear to regress on synthetic benchmarks whose construction assumptions are inconsistent with nature. Developing better evaluation protocols – including human evaluation at scale and real-world test sets with diverse acoustic environments – is an important direction for the field.

6.3 Limitations

Computational cost. Distance-aware mixing adds LLM inference overhead during training. CoI adds inference-time forward passes proportional to the number of steps. Co-occurrence estimation adds a CLAP encoding step at inference. None of these are prohibitive in batch processing, but they may limit applicability in real-time systems. Notably, applying CoI during training (rather than only at inference) requires additional care: the gradient graph must be disconnected at intermediate steps, and the multi-step output can be blended with the original one-step result via a weighted sum to stabilise training.

Scope of evaluation. The human evaluation in Chapter 3 uses 4 voters and 300 clips per benchmark; larger-scale evaluation with diverse annotator backgrounds would strengthen the conclusions.

Exploratory scope of co-occurrence conditioning. The co-occurrence investigation in Chapter 3 is exploratory in nature. The $\hat{a}_2$ estimator demonstrates that co-occurrence information can improve separation on five of six benchmarks, establishing the viability of the approach. However, whether the improvements generalise across architectures beyond AudioSep remains an open question, and alternative estimation strategies may further improve robustness.

6.4 Future Directions

Unified framework. The three contributions of this thesis are largely independent. An intriguing direction is to combine them: training a model with distance-aware MoMs and co-occurrence conditioning, then applying CoI at inference. The interactions between these three axes of improvement remain unexplored.

Learning the ratio schedule. The cosine-similarity ratio schedule in CoI is hand-designed. A learned schedule – trained to minimise separation error across a diverse set of mixtures – could outperform the cosine schedule, especially for models with non-standard output statistics.

6.5 Conclusion

This thesis has presented three complementary approaches to improving Universal Sound Separation. A distance-aware mixing strategy and a co-occurrence conditioning scheme leverage external knowledge – from LLMs and CLAP – to make training data more realistic and source specifications more complete. A chain-of-inference framework addresses a different bottleneck entirely: it unlocks latent model capacity at inference time through principled iterative refinement, without requiring any additional knowledge or training. Together, these contributions demonstrate that

the gap between current USS systems and real-world acoustic performance can be narrowed by intervening at multiple points in the model lifecycle – data construction, model conditioning, and inference strategy – without requiring new annotated data.

The broader message is that the mix-and-separate paradigm, while powerful, is not a complete solution to the USS problem. The quality of training data, the sophistication of inference, and the richness of conditioning each independently constrain what is achievable. Addressing all three simultaneously, and developing evaluation methodologies that faithfully measure real-world performance, are essential steps toward universal sound separation that truly works in the wild.

Bibliography

[1] Daniel-José Alcala Padilla, Nils L Westhausen, Swati Vivekananthan, and Bernd T Meyer. Location-aware target speaker extraction for hearing aids. In *Proc. Interspeech 2025*, pages 2975–2979, 2025.

[2] Barkopedia. Dog vocal separation. `https://huggingface.co/datasets/ArlingtonCL2/Dog-Vocal-Separation`, 2025. Hugging Face Datasets.

[3] Bastian Bechtold. python-soundfile. `https://github.com/bastibe/python-soundfile`, 2025.

[4] Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

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

[6] Ke Chen, Gordon Wichern, François G. Germain, and Jonathan Le Roux. Pac-HuBERT: Self-supervised music source separation via primitive auditory clustering and hidden-unit BERT. In *ICASSPW*, pages 1–5. IEEE, 2023.

[7] Edward Collin Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of the acoustical society of America*, 25:975–979, 1953.

[8] Woosung Choi, Minseok Kim, Jaehwa Chung, and Soonyoung Jung. LaSAFT: Latent source attentive frequency transformation for conditioned source separation. In *ICASSP*, pages 171–175. IEEE, 2021.

[9] Diego de Benito-Gorrón, Katerina Zmolikova, and Doroteo T. Toledano. Source separation for sound event detection in domestic environments using jointly trained models. In *IWAENC*. IEEE, 2022. doi: 10.1109/IWAENC53105.2022.9914755.

[10] Diego de Benito-Gorrón, Katerina Zmolikova, and Doroteo T. Toledano. Analysis and interpretation of joint source separation and sound event detection in domestic environments. *PLoS ONE*, 19(7):e0303994, 2024. doi: 10.1371/journal.pone.0303994.

[11] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP*, pages 736–740. IEEE, 2020.

[12] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 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, 2023.

[13] Carles Escera, Kimmo Alho, István Winkler, and Risto Näätänen. Neural mechanisms of involuntary attention to acoustic novelty and change. *Journal of cognitive neuroscience*, 10(5):590–604, 1998.

[14] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K, October 2020.

[15] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, pages 776–780. IEEE, 2017.

[16] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg,

Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

[17] Toni Heittola, Annamaria Mesaros, Tuomas Virtanen, and Antti Eronen. Sound event detection in multisource environments using source separation. In *Workshop on Machine Listening in Multisource Environments (CHiME)*, 2011.

[18] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: A fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020. doi: 10.21105/joss.02154.

[19] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *ICASSP*, pages 31–35. IEEE, 2016.

[20] Poul Hoang, Jan Mark de Haan, Zheng-Hua Tan, and Jesper Jensen. Multichannel speech enhancement with own voice-based interfering speech suppression for hearing assistive devices. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 30: 706–720, 2022. doi: 10.1109/TASLP.2022.3145294.

[21] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 29:3451–3460, 2021.

[22] Chao Huang, Yuesheng Ma, Junxuan Huang, Susan Liang, Yunlong Tang, Jing Bi,

Wenqiang Liu, Nima Mesgarani, and Chenliang Xu. Zerosep: Separate anything in audio with zero training. *arXiv preprint arXiv:2505.23625*, 2025.

[23] Jieyi Huang, Chunhao Zhang, Mengyue Wu, and Kenny Zhu. Transcribing vocal communications of domestic shiba inu dogs. In *Findings of ACL*, pages 13819–13832. Association for Computational Linguistics, 2023.

[24] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *NeurIPS*, 2022.

[25] Xilin Jiang, Cong Han, and Nima Mesgarani. Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation. In *ICASSP*, pages 1–5. IEEE, 2025.

[26] Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R. Hershey. Universal sound separation. In *WASPAA*, pages 175–179. IEEE, 2019.

[27] Emine Merve Kaya and Mounya Elhilali. Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714), 2017.

[28] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *NAACL-HLT*, pages 119–132. Association for Computational Linguistics, 2019.

[29] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al.

Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[31] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022.

[32] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 25 (10):1901–1913, 2017.

[33] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[34] Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D. Plumbley, and Wenwu Wang. Separate anything you describe. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 33:458–471, 2024.

[35] Yi Luo and Nima Mesgarani. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 27(8):1256–1266, 2019.

[36] Yi Luo, Zhuo Chen, and Takuya Yoshioka. Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP*, pages 46–50. IEEE, 2020.

[37] Matplotlib Development Team. Matplotlib: Visualization with Python. `https://github.com/matplotlib/matplotlib`, 2012.

[38] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in Python. In *SciPy*, 2015.

[39] Gabriel Meseguer-Brocal and Geoffroy Peeters. Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations. *arXiv preprint arXiv:1907.01277*, 2019.

[40] Marius Miron, Sara Keen, Jen-Yu Liu, Benjamin Hoffman, Masato Hagiwara, Olivier Pietquin, Felix Effenberger, and Maddie Cusimano. Biodenoising: Animal vocalization denoising without access to clean data. In *ICASSP*, pages 1–5. IEEE, 2025.

[41] Yuki Mitsufuji, Giorgio Fabbro, Stefan Uhlich, Fabian-Robert Stöter, Alexandre Défossez, Minseok Kim, Woosung Choi, Chin-Yun Yu, and Kin-Wai Cheuk. Music demixing challenge 2021. *Frontiers in Signal Processing*, 1, 2022. doi: 10.3389/frsip.2021.808395.

[42] Risto Näätänen, Petri Paavilainen, Teemu Rinne, and Kimmo Alho. The mismatch negativity (mmn) in basic research of central auditory processing: a review. *Clinical neurophysiology*, 118(12):2544–2590, 2007.

[43] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[44] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[45] Karol J. Piczak. ESC: Dataset for environmental sound classification. In *ACM Multimedia*, pages 1015–1018. Association for Computing Machinery, 2015.

[46] Yanmin Qian, Xuankai Chang, and Dong Yu. Single-channel multi-talker speech recognition with permutation invariant training. *Speech Communication*, 104: 1–11, 2018. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2018.09.003.

[47] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The MUSDB18 corpus for music separation, December 2017.

[48] Reflex Development Team. Reflex: Web apps in pure Python. `https://github.com/reflex-dev/reflex`, 2022.

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[50] Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP*, pages 1–5. IEEE, 2023.

[51] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR – half-baked or well done? In *ICASSP*, pages 626–630. IEEE, 2019.

[52] Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. Sam audio: Segment anything in audio. *arXiv preprint arXiv:2512.18099*, 2025.

[53] Ragini Sinha, Ann-Christin Scherer, Simon Doclo, Christian Rollwage, and Jan Rennies. Evaluation of speaker-conditioned target speaker extraction algorithms for hearing-impaired listeners. *Trends in Hearing*, 29:1–15, 2025. doi: 10.1177/23312165251365802.

[54] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. Open-unmix – a reference implementation for music source separation. *Journal of Open Source Software*, 4(41):1667, 2019. doi: 10.21105/joss.01667.

[55] Aswin Shanmugam Subramanian, Chao Weng, Meng Yu, Shi-Xiong Zhang, Yong Xu, Shinji Watanabe, and Dong Yu. Far-field location guided target speech extraction using end-to-end speech recognition objectives. In *ICASSP*. IEEE, 2020.

[56] Joachim Thiemann, Menno Müller, Daniel Marquardt, Simon Doclo, and Steven van de Par. Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene. *EURASIP Journal on Advances in Signal Processing*, 2016(12), 2016. doi: 10.1186/s13634-016-0314-6.

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

[58] Efthymios Tzinis, Scott Wisdom, John R. Hershey, Aren Jansen, and Daniel P. Ellis. Improving universal sound separation using sound classification. In *ICASSP*, pages 96–100. IEEE, 2020.

[59] Efthymios Tzinis, Gordon Wichern, Aswin Subramanian, Paris Smaragdis, and Jonathan Le Roux. Heterogeneous target speech separation. *arXiv preprint arXiv:2204.03594*, 2022.

[60] Cassia Valentini-Botinhao. Noisy reverberant speech database for training speech enhancement algorithms and TTS models, 2017.

[61] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(4):1462–1469, 2006.

[62] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix

factorization with temporal continuity and sparseness criteria. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(3):1066–1074, 2007.

[63] DeLiang Wang and Guy J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[64] Helin Wang, Bowen Shi, Andros Tjandra, John Hoffman, Yi-Chiao Wu, Apoorv Vyas, Najim Dehak, Ann Lee, and Wei-Ning Hsu. Sam audio judge: A unified multimodal framework for perceptual evaluation of audio separation. *arXiv preprint arXiv:2601.19702*, 2026.

[65] Theron S. Wang, Xingyuan Li, Chunhao Zhang, Mengyue Wu, and Kenny Q. Zhu. Phonetic and lexical discovery of canine vocalization. In *Findings of EMNLP*, pages 13972–13983. Association for Computational Linguistics, 2024.

[66] Yun Wang, Daniel Stoller, Rachel M. Bittner, and Juan Pablo Bello. Few-shot musical source separation. In *ICASSP*, pages 121–125. IEEE, 2022.

[67] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[68] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

[69] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron Weiss, Kevin Wilson, and John Hershey. Unsupervised sound separation using mixture invariant training. In *NeurIPS*, 2020.

[70] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[71] Jianwei Yu, Yi Luo, Hangting Chen, Rongzhi Gu, and Chao Weng. High fidelity speech enhancement with band-split RNN. *arXiv preprint arXiv:2212.00406*, 2022.

[72] Yi Yuan, Xubo Liu, Haohe Liu, Mark D. Plumbley, and Wenwu Wang. FlowSep: Language-queried sound separation with rectified flow matching. In *ICASSP*, pages 1–5. IEEE, 2025.

[73] Yongyi Zang, Jingyi Li, and Qiuqiang Kong. Training-free multi-step audio source separation. *arXiv preprint arXiv:2505.19534*, 2025.

[74] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, pages 570–586, 2018.

[75] Jianwei Zhao, Xubo Liu, Jianwei Zhao, Yi Yuan, Qiuqiang Kong, Mark D. Plumbley, and Wenwu Wang. Universal sound separation with self-supervised audio masked autoencoder. In *EUSIPCO*, pages 1–5. IEEE, 2024.