The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Energy Aware Computing Research Group

WEAKLY-SUPERVISED AUDIO SEPARATION VIA BIMODAL SEMANTIC SIMILARITY

(ICLR 2024)

Tanvir Mahmud[*1†], Saeed Amizadeh[†2], Kazuhito Koishida[2], and Diana Marculescu[1]

[1]University of Texas at Austin, [2]Microsoft, [†]Equal Contribution

∗Work done in part during an internship at Microsoft Corporation, Redmond, USA

Microsoft

Overview

- **Challenges of sound separation in mixtures**
- **Limitations of prior works**
- **Introduction to proposed hypothesis**
- **Proposed methodology:**
  - Language-conditioned Unsupervised Sound Separation
  - Hierarchical Reconstruction Loss
- **Experiments:**
  - Datasets
  - Experimental Setup
  - Ablation Study
- **Conclusion**
- **Future Study**

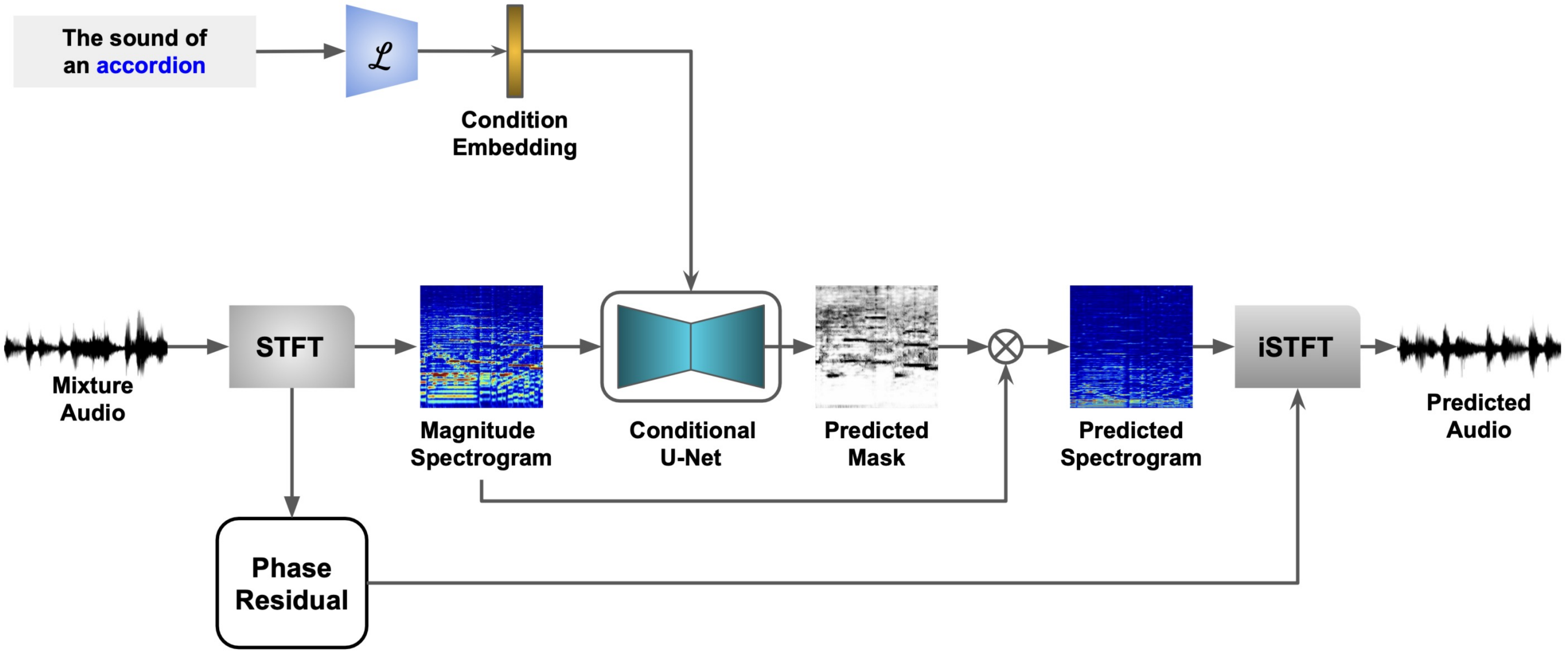
Challenges of sound separation in mixtures

- **Environmental sounds comes in natural mixtures**
  - Example 1:
    - Caption: A man talking while wood clanks on a metal pan followed by gravel crunching as food and oil sizzle
  - Example 2:
    - Caption: An adult female speaks and several people laugh, while slight rustling occurs in the background
- **It is not always feasible to gather clean-paired sounds of each source for training**
- **However, captions can represent the complex sounding events**
- **Is it possible to incorporate captions in order to use large-scale natural mixtures for training?**

Training Data and Configurations

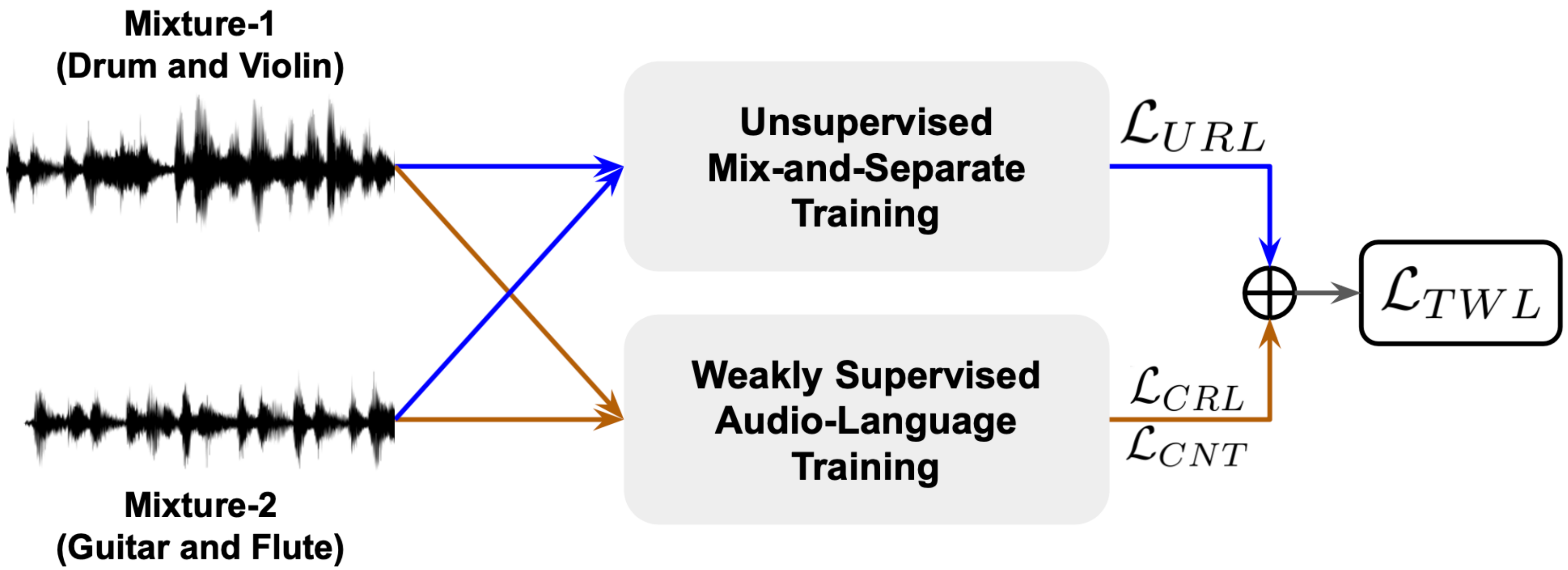
- **Supervised Single-Source:**
  - Single-source clean data is available for each source
- **Unsupervised Multi-Source:**
  - No single source clean data is available
  - Every sample represents mixture of numer of single source sounds
  - A representative caption can be available
- **Semi-supervised:**
  - Small to large fraction of single source sound is available
  - Multi-source mixture only data are available with representative captions

Inference Pipeline



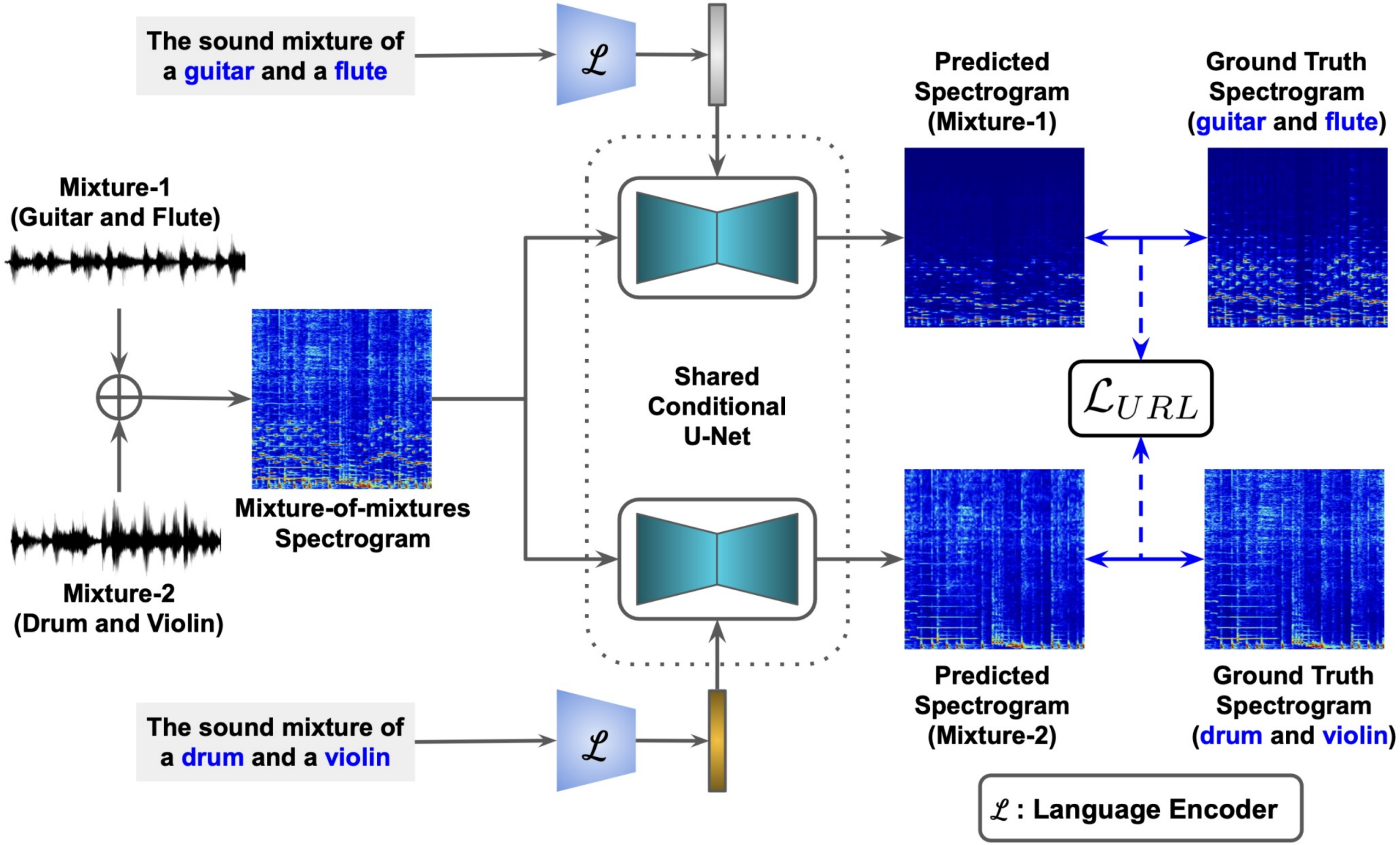
How to get the supervision on single-source separation predictions, when only mixture audio is available for training?

Proposed Framework



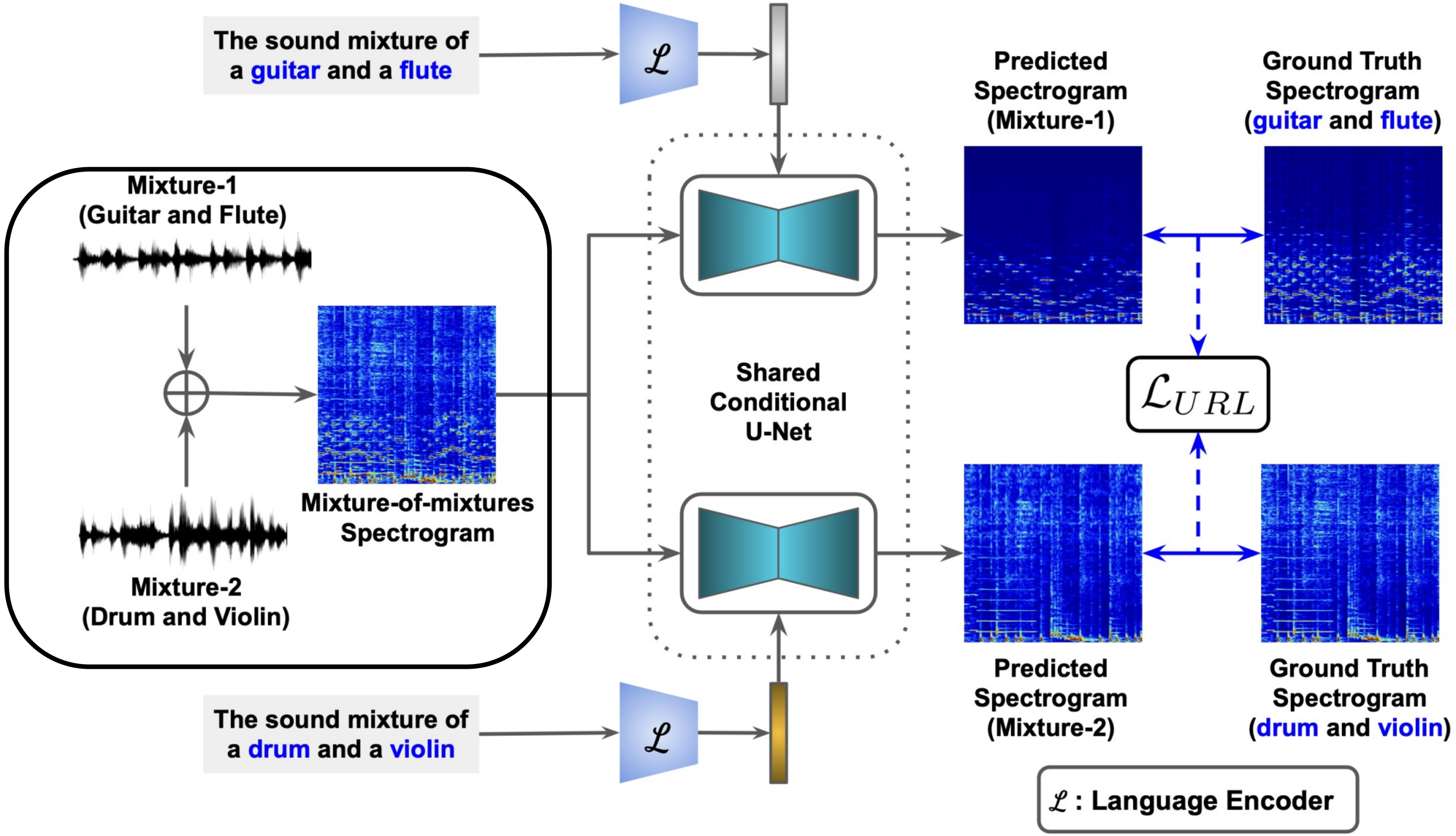
We propose an weakly supervised audio-language training method, to overcome limitations of multi-source natural mixtures

Baseline: Unsupervised Mix-and-Separate Framework



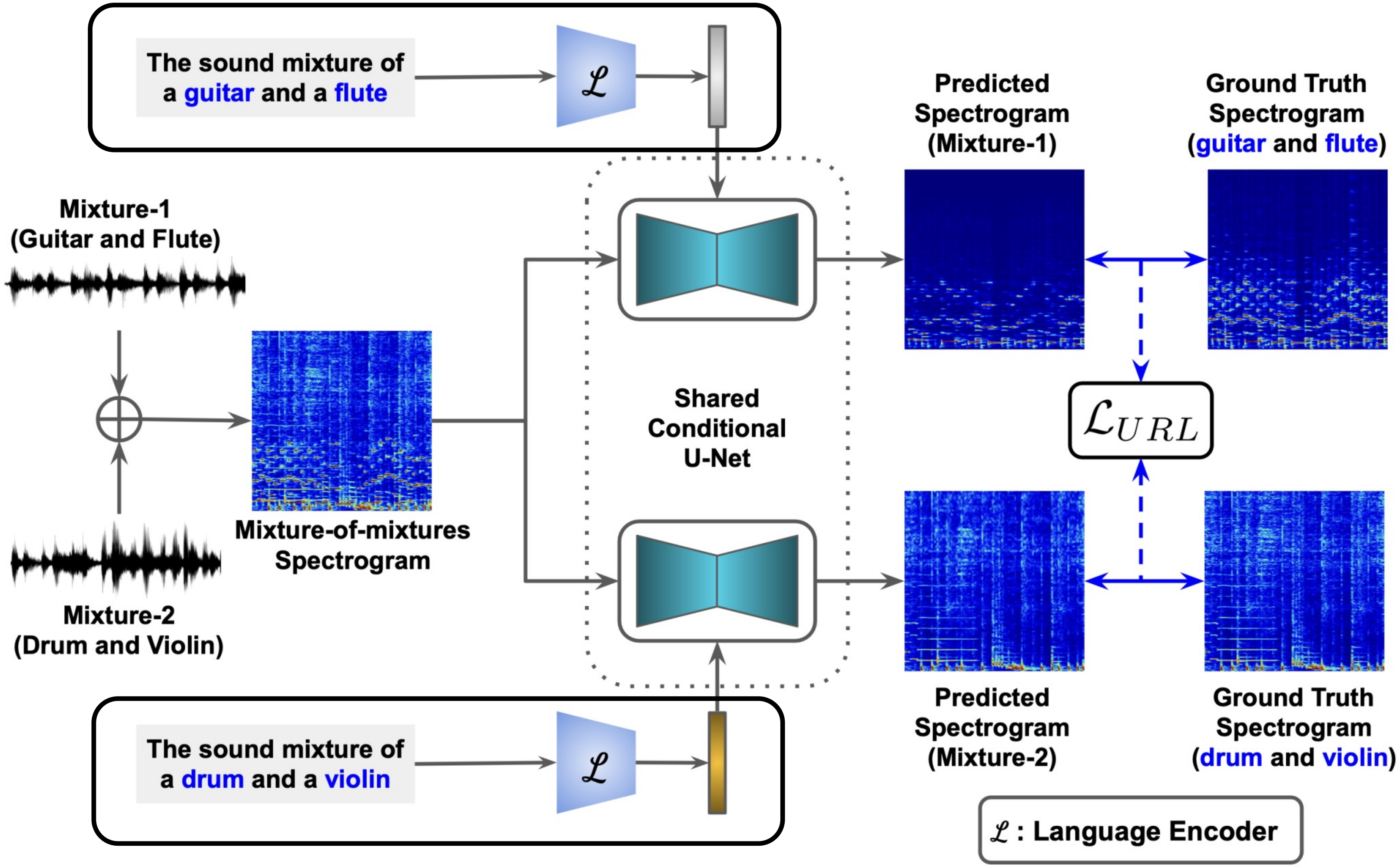
CLIPSep (ICLR'22), CCoL(CVPR'21), CoSep(ICCV'19), SOP(ECCV'18)

Baseline: Unsupervised Mix-and-Separate Framework



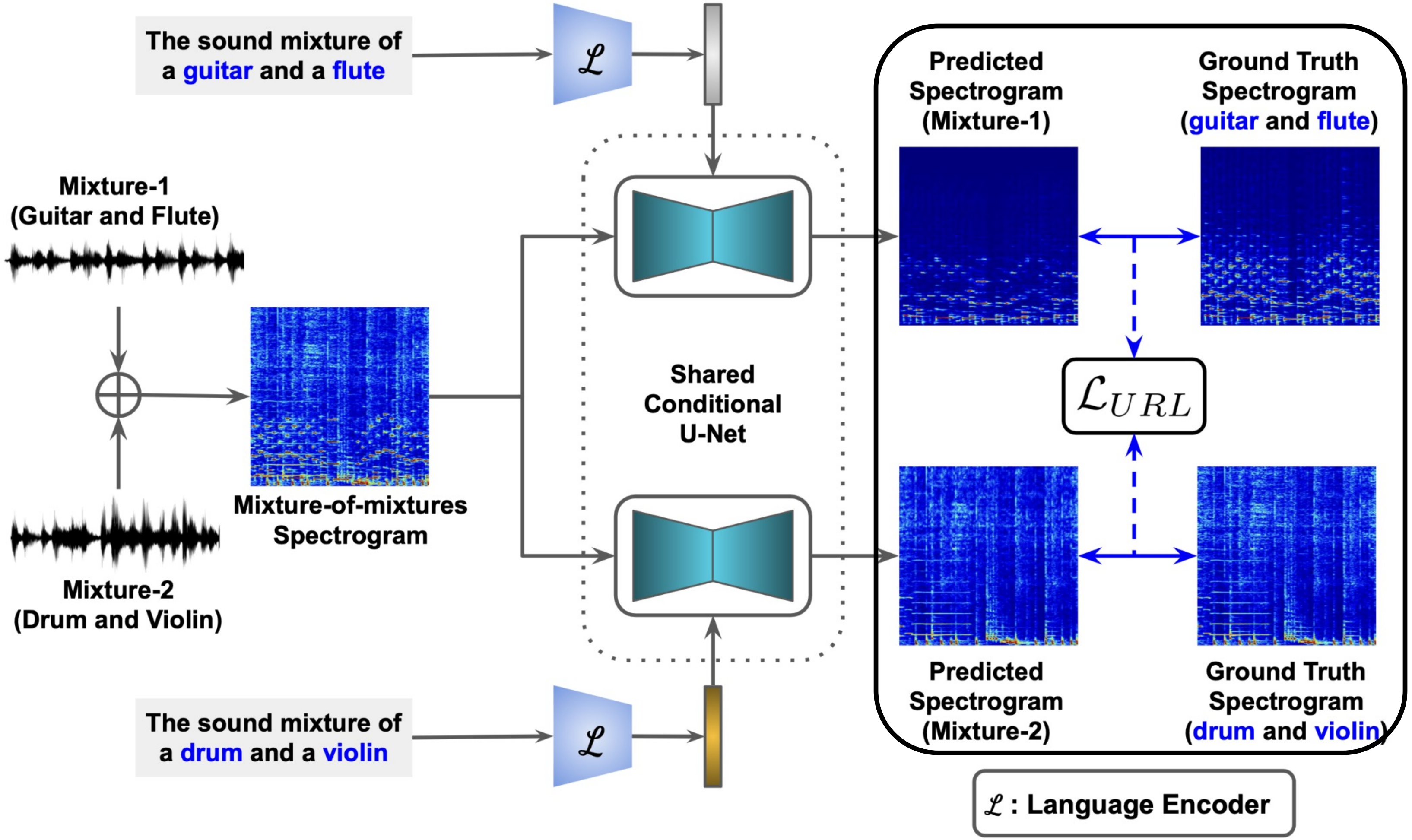
CLIPSep (ICLR'23), CCoL(CVPR'21), CoSep(ICCV'19), SOP(ECCV'18)

Baseline: Unsupervised Mix-and-Separate Framework



CLIPSep (ICLR'23), CCoL(CVPR'21), CoSep(ICCV'19), SOP(ECCV'18)

Baseline: Unsupervised Mix-and-Separate Framework



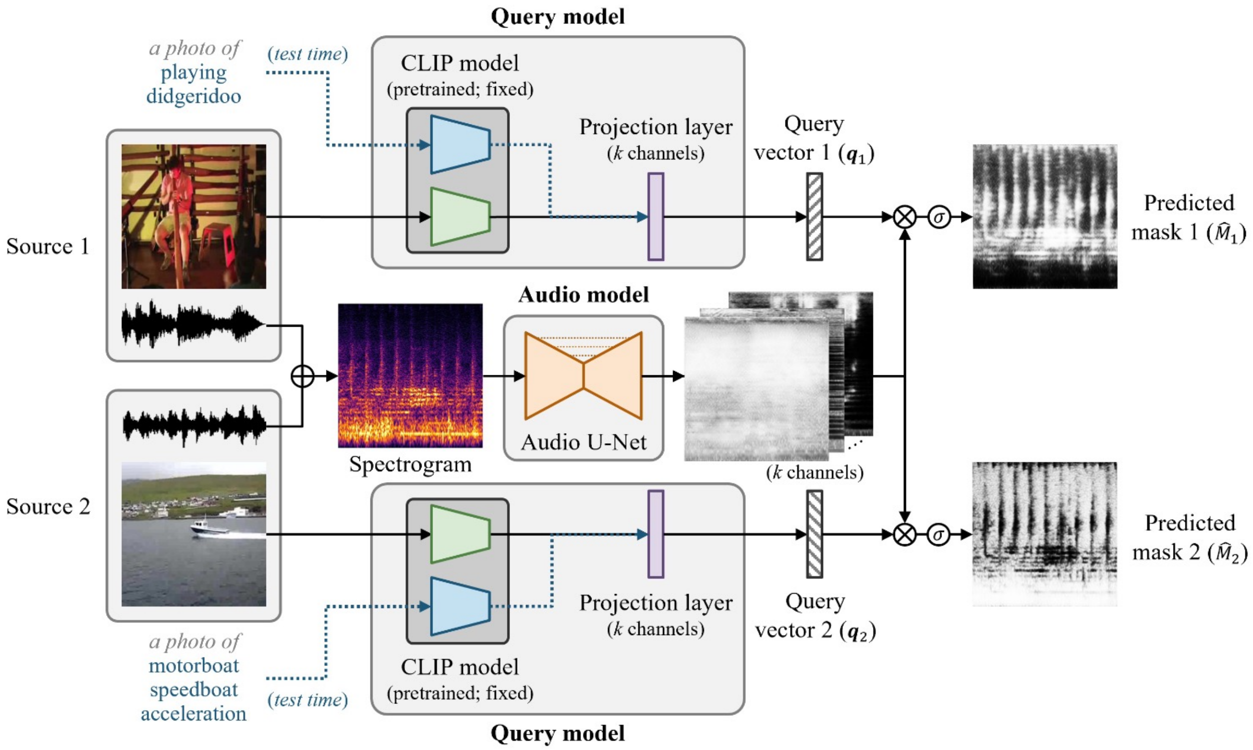
CLIPSep (ICLR'23), CCoL(CVPR'21), CoSep(ICCV'19), SOP(ECCV'18)

Limitations of prior works

- **Unconditional Mix-and-Separate**
  - It’s the primary baseline for unsupervised sound separation
  - The method works well if we consider <u>mixtures as a single sounding source</u>
  - With increasing the number of sounding sources in the mixtures, the method’s performance significantly drops
    - The training objective becomes more challenging to discover clean sounds from complex mixtures

- **Vision-Conditional Sound Separation**
  - Conditioning with videos suffer another challenge of computational complexity and extracting sounding sources
    - Sounds may appear from non-visible sources

Related Works (CLIPSep, ICLR-2023)



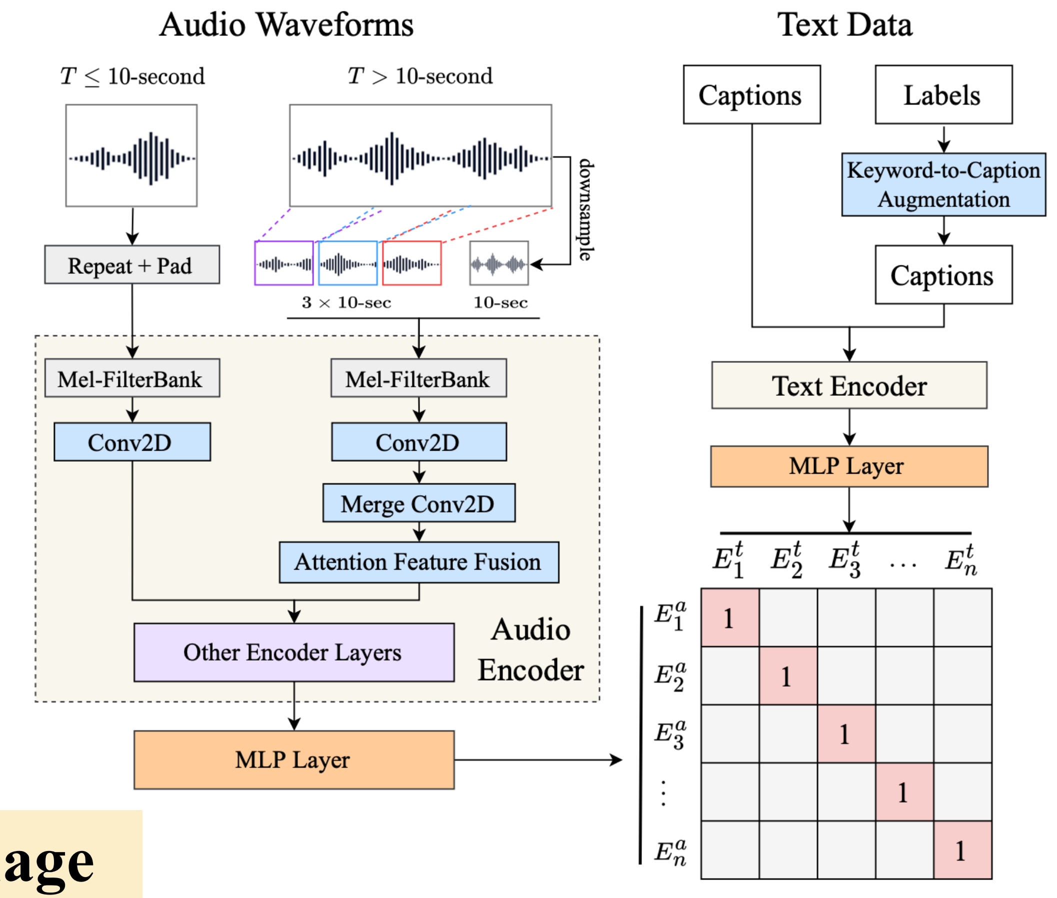
- **A mix-and-separate framework**
- **Key contribution:**
  - Modality inversion of conditioning
  - Directly source video can be used for training without captions
  - Test scenarios can be either from visual or text conditions
- **Limitations:**
  - Limited to single source data
  - Multi-source videos can have silent sources, background objects, etc.
  - Performance drops largely on multi-source only training

Dong, H. W., Takahashi, N., Mitsufuji, Y., McAuley, J., & Berg-Kirkpatrick, T. (2022, September). CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos. In *The Eleventh International Conference on Learning Representations*.

Related Works: CLAP (ICASSP '23)

Data Statistics

Dataset	Pairs	Audio Durations (hrs)
Clotho [15]	5,929	37.00
SoundDescs [16]	32,979	1060.40
AudioCaps [17]	52,904	144.94
LAION-Audio-630K	633,526	4325.39



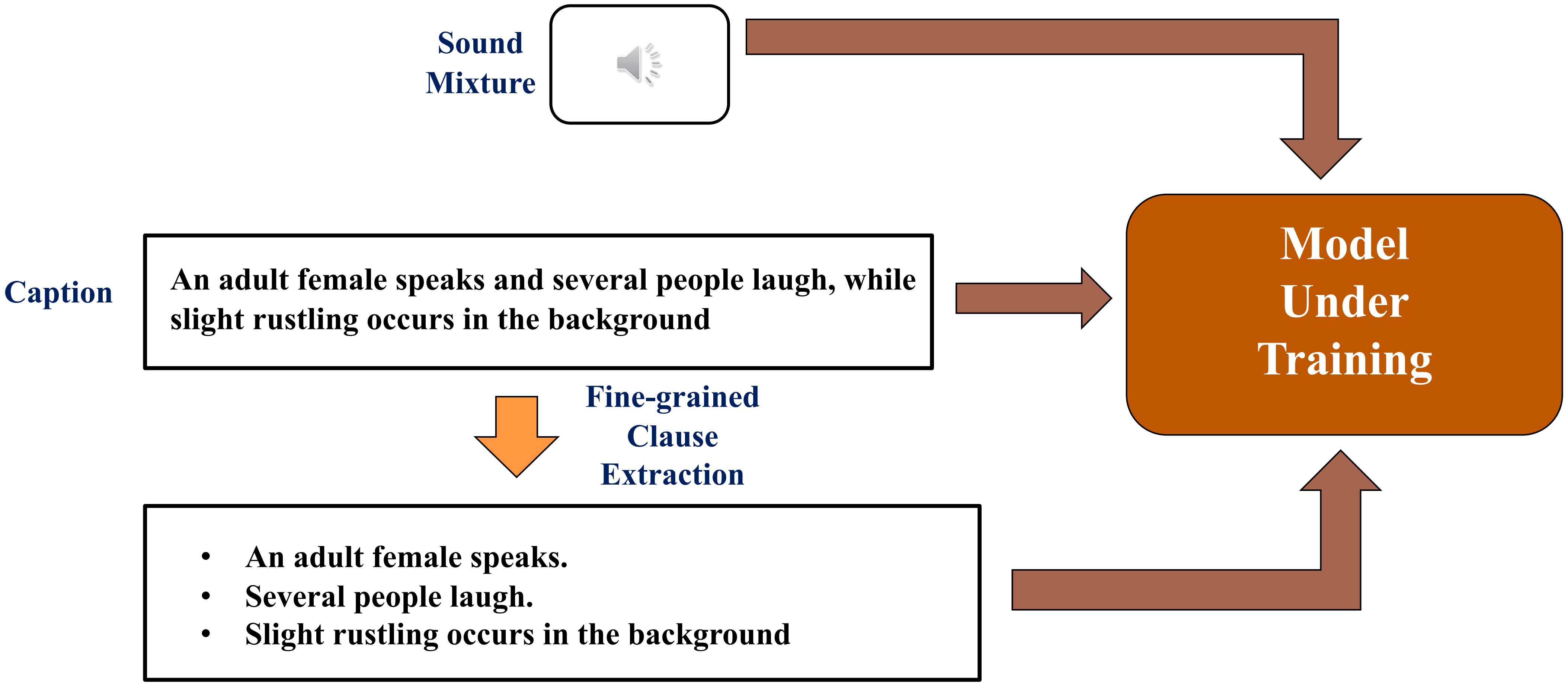
Grounds Audios and representative Language captions through large-scale pretraining

An Idea

- **Text can represent fine-grained details of the audio mixtures**

Is it possible to extract fine-grained details of sounding sources from text, and improve unsupervised sound separation from natural mixtures?

The Main Hypothesis

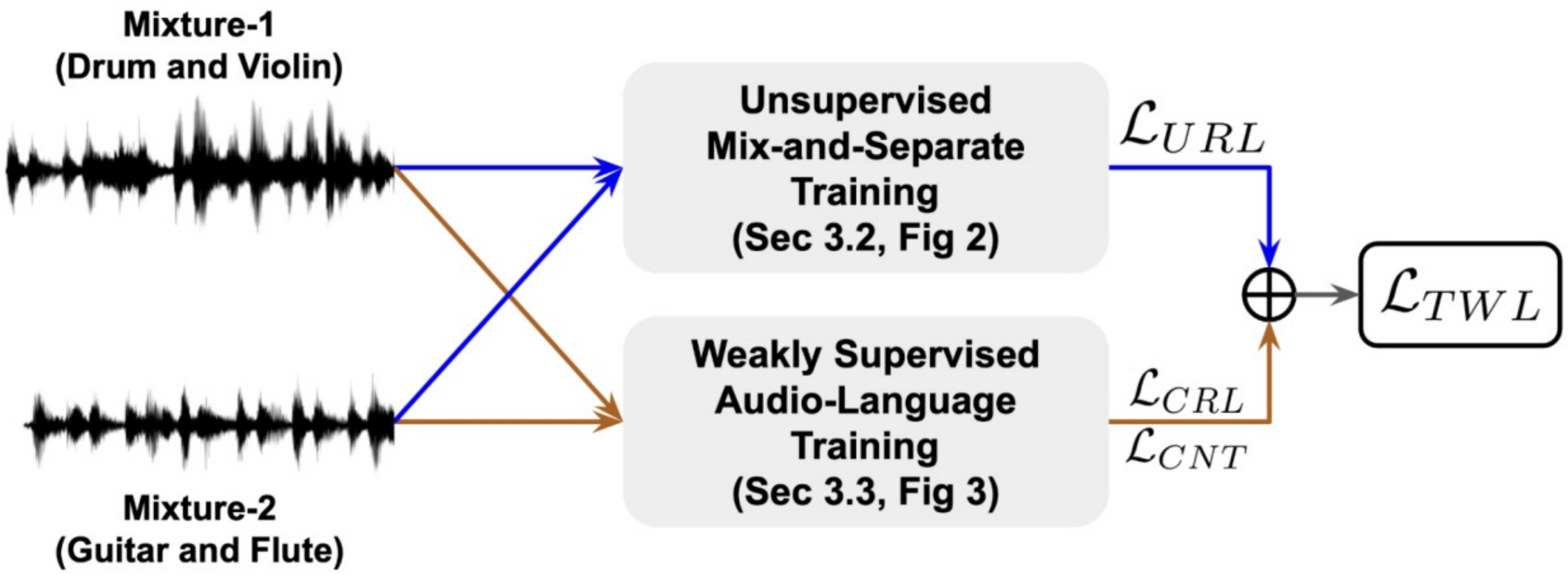


In the absence of clean training audio data, can we use fine-grained semantic text-clauses of different sound sources as a form of supervision to train a conditional sound separation model?

Problem Statement

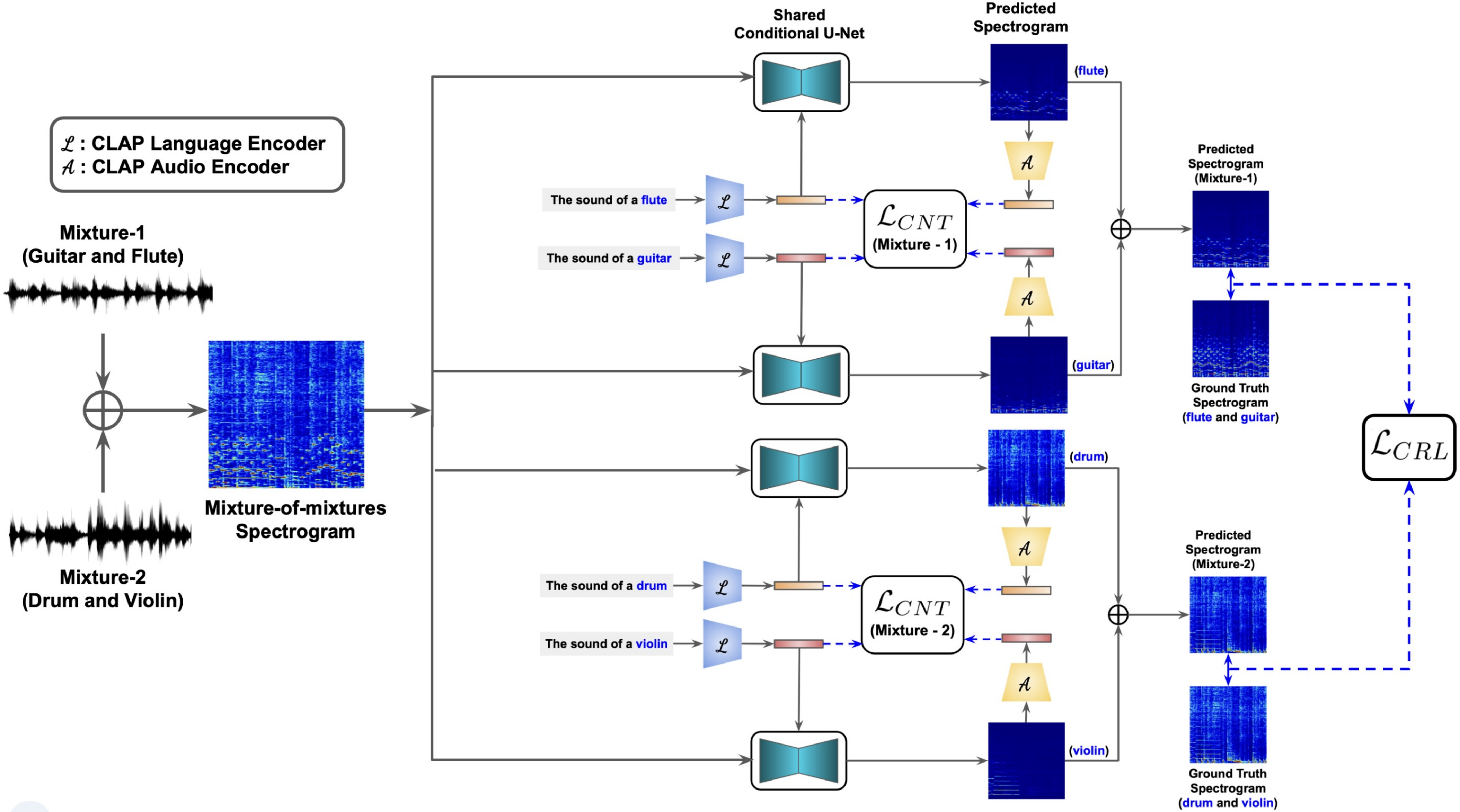
How to leverage natural language caption of a sound mixture, to train a conditional sound separation, without having access to single-source audio data during training?

Proposed Framework

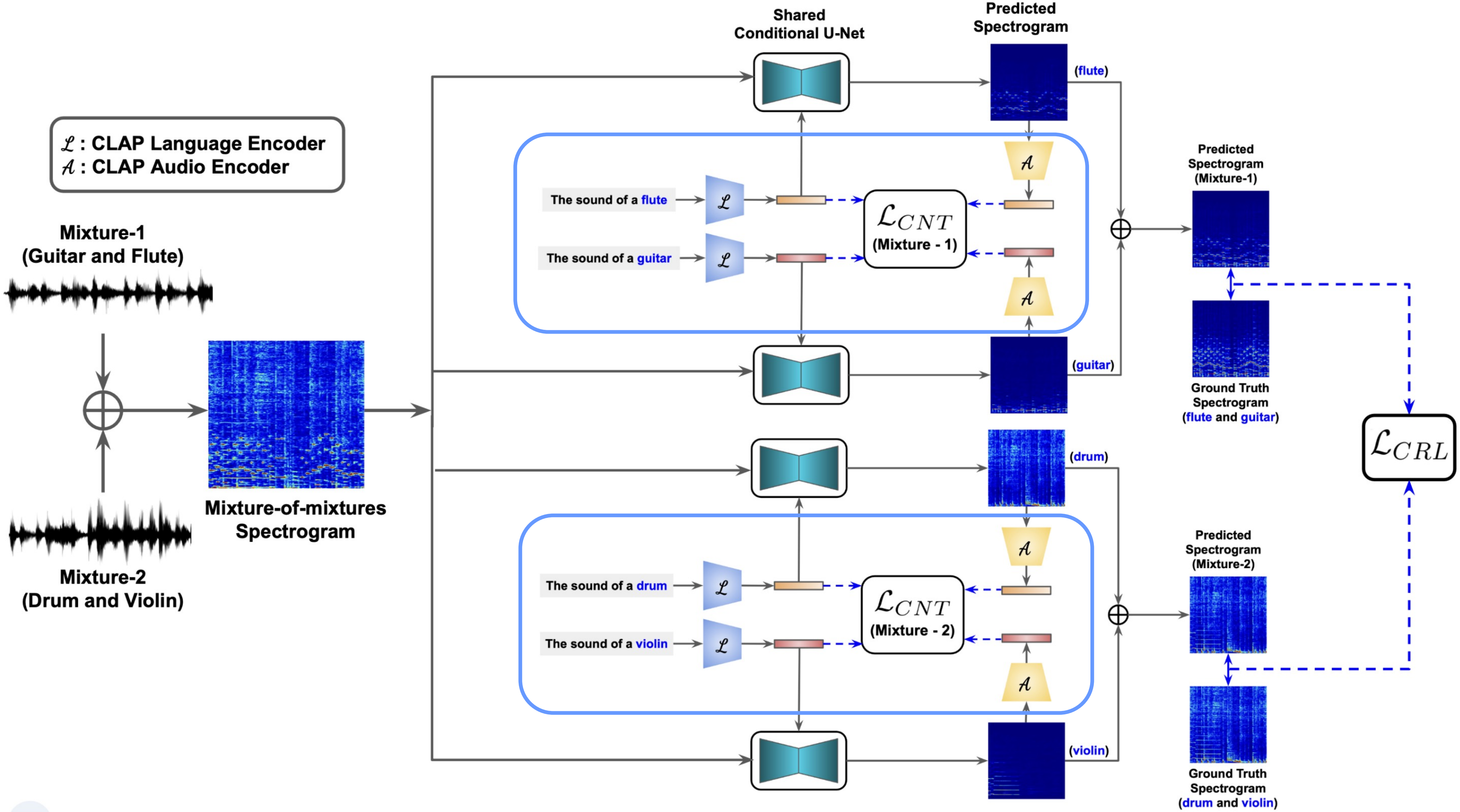


We propose an weakly supervised audio-language training method, to overcome limitations of multi-source natural mixtures

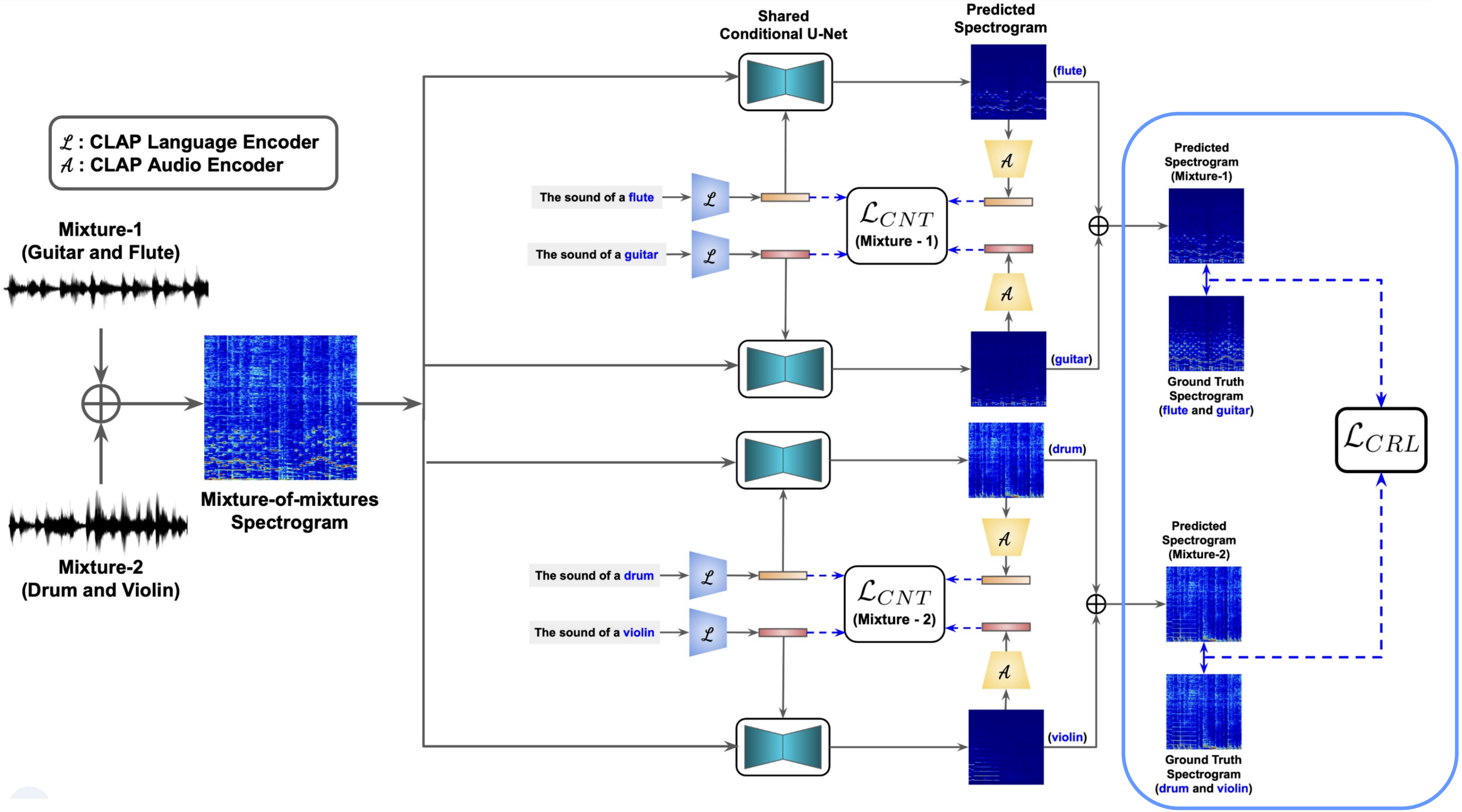
Proposed Weakly Supervised Audio-Language Training



Proposed Weakly Supervised Audio-Language Training



Proposed Weakly Supervised Audio-Language Training

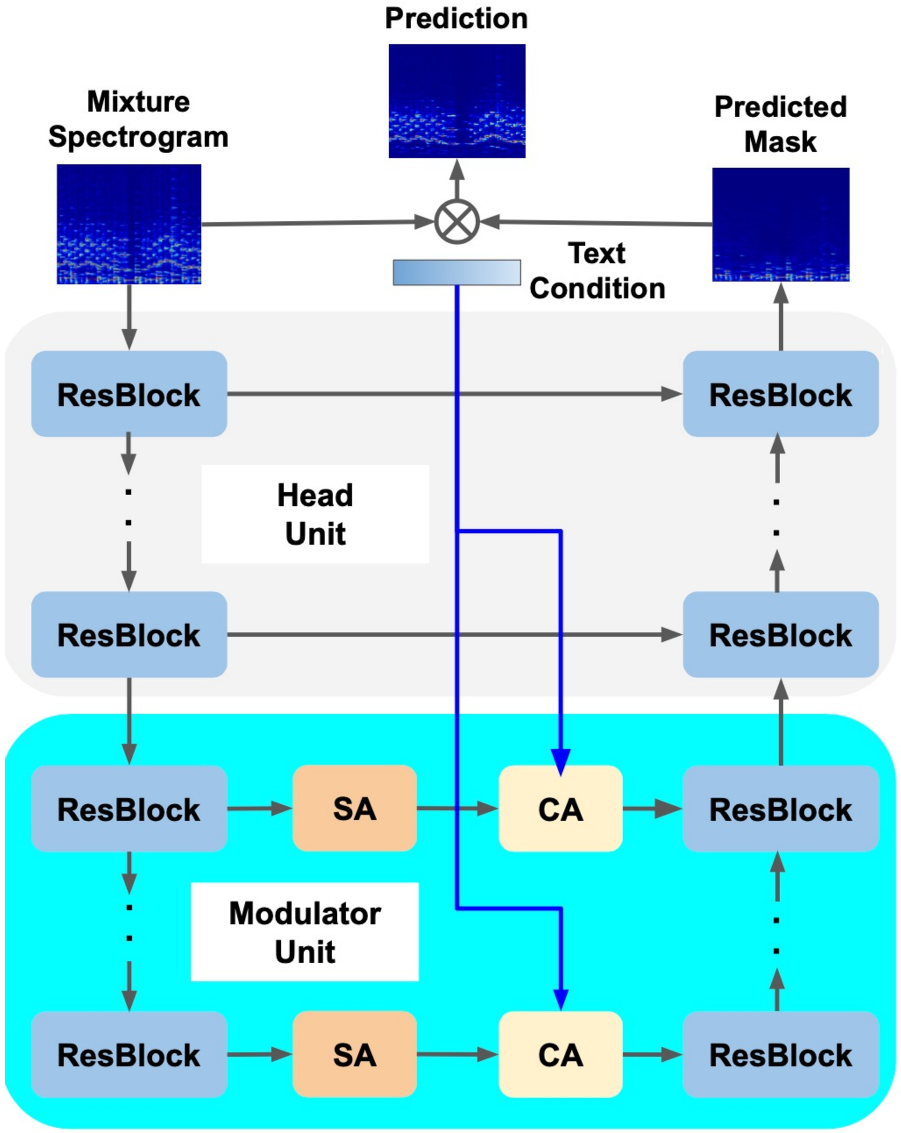


Proposed Semi-Supervised Learning

- **Combines learning with supervised (clean sounds) and unsupervised (mixture sounds)**
- **Only mix-and-separate is used for clean sound learning**
- **Proposed framework is used for learning on mixtures:**
  - Combining mix-and-separate with proposed weakly supervised method

$$\mathcal{L}_{SSL}(\mathcal{B}' \cup \mathcal{S}', \theta) = \lambda_s \cdot \mathcal{L}_{URL}(\mathcal{S}', \theta) + \lambda_u \cdot \mathcal{L}_{TWL}(\mathcal{B}', \theta)$$

Modifications of Conditional U-Net Architecture



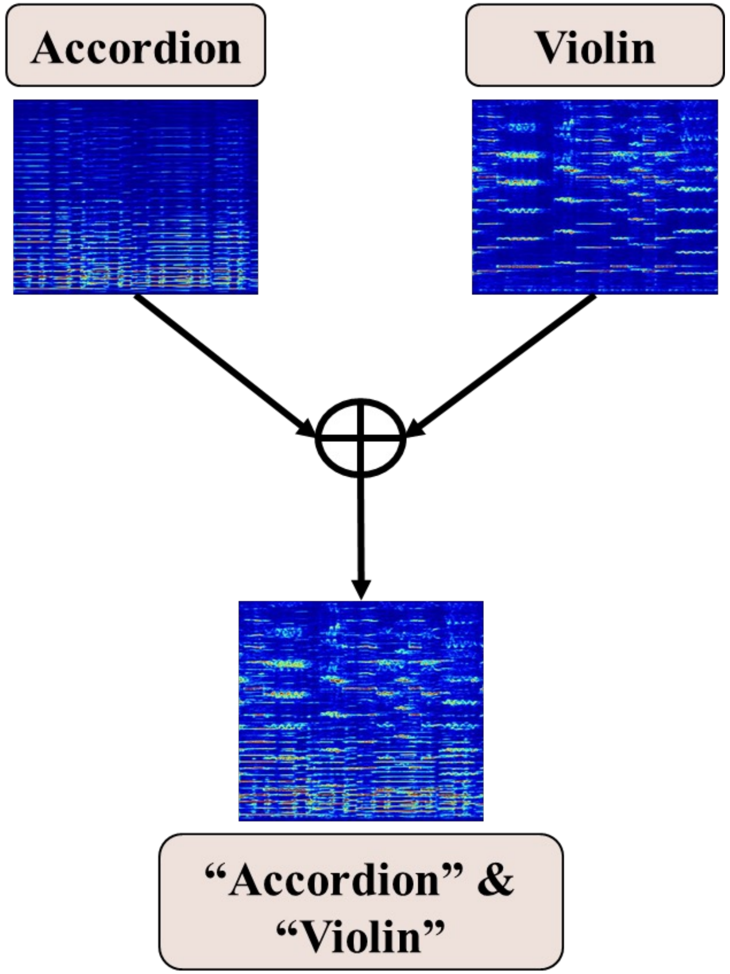
- Prior works rely on unconditional U-Net architecture with late-conditioning
- Shallow architecture is used in general
- For focusing on supervised learning with clean sounds, shallow network performed well
- We modify the architecture for enhanced feature extraction with deeper conditioning

Experimental Dataset

- **MUSIC Dataset (Used for Synthetic Mixtures Training):**
  - Contains 823 audios of single sources
  - Contains 17 classes of sounds
  - Each video contains 1~4 minutes of sounds
- **VGGSound Dataset (Used for Synthetic Mixtures Training):**
  - Contains nearly 180k videos of 10s duration
  - Contains 309 classes
- **AudioCaps Dataset (Used for Natural Mixtures Training) :**
  - Contains ~50k audios of 10s duration
  - Contains natural captions
  - Diverse sounding sources with variable number of sources

Experimental Setups (Synthetic Training and Eval)

- **Synthetic Training:**
  - Every Training Mixture contains **2** sounds
  - Every Training Mixture contains **3** sounds
  - Every Training Mixture contains **4** sounds
- **Synthetic Testing:**
  - Every Test Mixture contains **2** sounds
  - Every Test Mixture contains **3** sounds
  - Every Test Mixture contains **4** sounds
- **Synthetic Training demonstrates the real-scenario of complex environmental mixtures with increasing complexity**
- **Carried out with MUSIC and VggSound datasets**



Experimental Setups (Real-world Training and Eval)

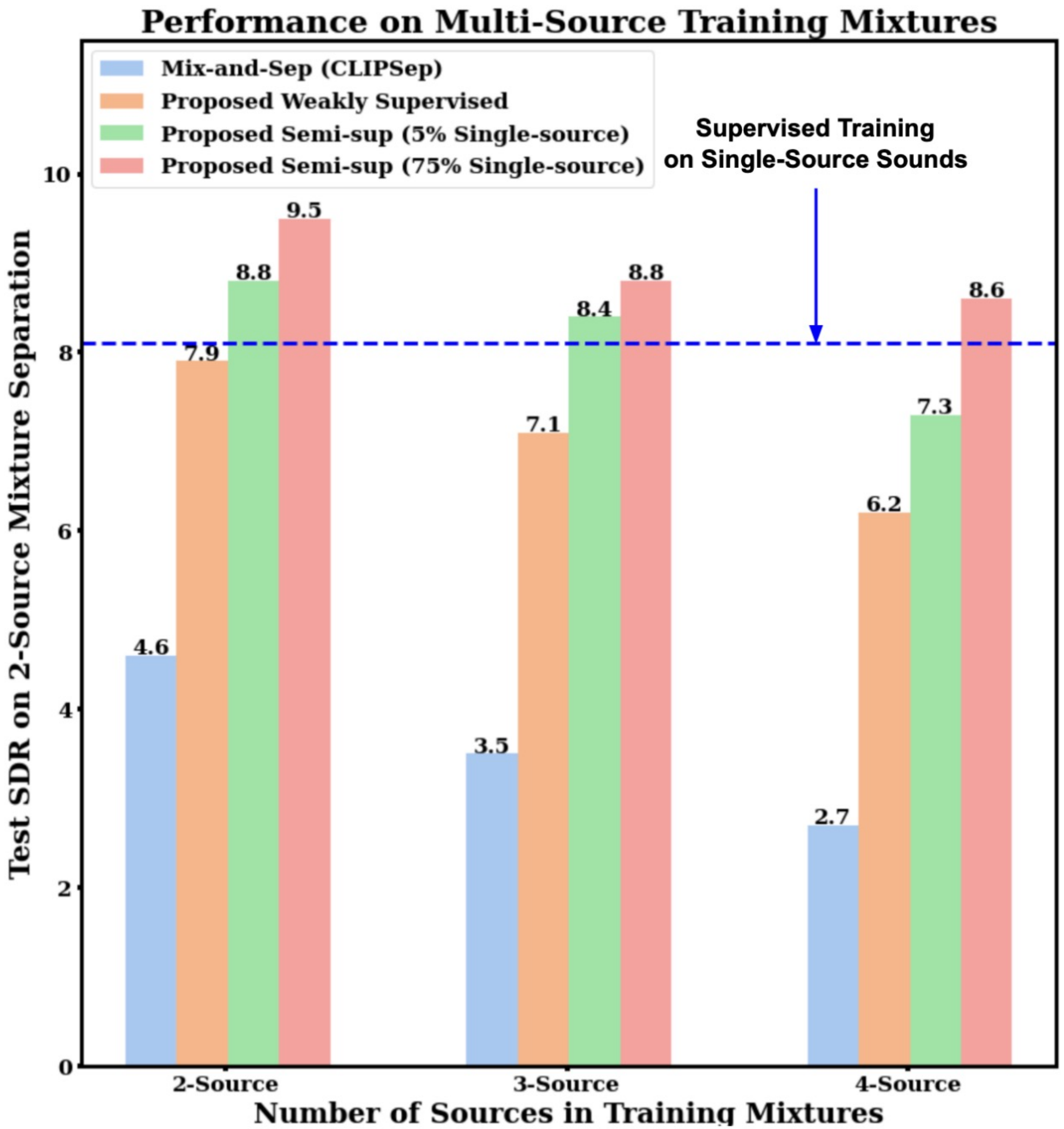
- **Training:**
  - Contains the available environmental mixtures of sounds
  - 1~6 for AudioCaps
- **Synthetic Testing:**
  - Every Test Mixture contains **2** mixture of sounds
  - Evaluation is carried on each mixture
- **Synthetic Training demonstrates the real-scenario of complex environmental mixtures with increasing complexity**
- **Carried out with large-scale AudioCaps dataset**

Caption:
An adult female speaks and several people laugh, while slight rustling occurs in the background

Evaluation Metrics

- **SDR (Source-to-Distortion Ratio):**
  - SDR is usually considered to be an overall measure of how good a source sounds
  - If a paper only reports one number for estimated quality, it is usually SDR
- **SIR (Source-to-Interference Ratio):**
  - This is usually interpreted as the number of other sources that can be heard in a source estimate
  - This is most close to the concept of “bleed”, or “leakage”
- **SAR (Source-to-Artifact Ratio):**
  - This is usually interpreted as the amount of unwanted artifacts a source estimate has with relation to the true source.

Performance on Higher Order Mixture Training



- Mix-and-Separate significantly loses performance on higher mixtures
- Proposed framework largely recovers performance loss on higher mixtures
- Learning with 5% clean sounds surpass the supervised training with 100% clean sounds in Mix-and-Separate
- This experiment is conducted on MUSIC dataset

Quantitative Results

Table 1: Comparison on MUSIC Dataset under the unsupervised setup. The supervised column is also provided as an upperbound. SDR on 2-Source separation test set is reported for all cases. All methods are reproduced under the same setting. * denotes implementation with our improved U-Net model. **Bold** and blue represents the best and second best performance in each group, respectively.

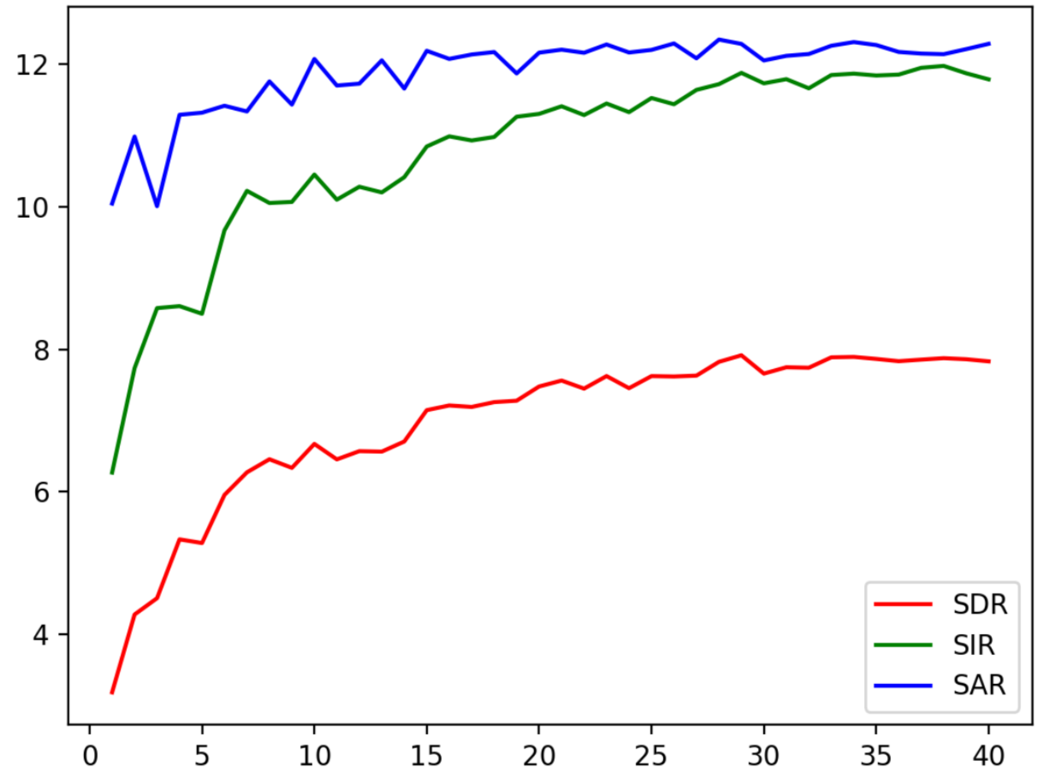
Method	Single-Source (Supervised)	Multi-Source (Unsupervised)		
		2-Source	3-Source	4-Source
Unconditional				
PIT* (Yu et al., 2017)	8.0 ± 0.26	-	-	-
MixIT (Wisdom et al., 2020)	-	3.2 ± 0.34	2.3 ± 0.57	1.4 ± 0.35
MixPIT (Karamatlı & Kırbız, 2022)	-	3.6 ± 0.46	2.1 ± 0.41	1.7 ± 0.35
Image Conditional				
CLIPSep-Img (Dong et al., 2022)	6.8 ± 0.25	3.8 ± 0.27	2.9 ± 0.35	2.1 ± 0.32
CLIPSep-Img* (Dong et al., 2022)	7.4 ± 0.22	4.6 ± 0.31	3.8 ± 0.28	2.9 ± 0.43
CoSep* (Gao & Grauman, 2019)	7.9 ± 0.28	4.9 ± 0.37	4.0 ± 0.29	3.1 ± 0.36
SOP* (Zhao et al., 2018)	6.5 ± 0.23	4.1 ± 0.41	3.5 ± 0.26	2.7 ± 0.42
Language Conditional				
CLIPSep-Text (Dong et al., 2022)	7.7 ± 0.21	4.6 ± 0.35	3.5 ± 0.27	2.7 ± 0.45
CLIPSep-Text* (Dong et al., 2022)	**8.3** ± 0.27	5.4 ± 0.41	4.7 ± 0.32	3.8 ± 0.28
BertSep*	7.9 ± 0.27	5.3 ± 0.31	4.0 ± 0.22	3.1 ± 0.27
CLAPSep*	8.1 ± 0.31	5.5 ± 0.36	4.3 ± 0.28	3.5 ± 0.33
LASS-Net (Liu et al., 2022)	7.8 ± 0.25	5.2 ± 0.26	4.2 ± 0.29	3.6 ± 0.36
Weak-Sup (Pishdadian et al., 2020)	-	3.1 ± 0.47	2.2 ± 0.38	1.9 ± 0.33
Proposed (w/ Timbre Classifier - concurrent training)	-	5.0 ± 0.29	4.5 ± 0.32	3.5 ± 0.27
Proposed (w/ Timbre Classifier - pretrained)	-	6.1 ± 0.33	5.2 ± 0.37	4.1 ± 0.35
Proposed (w/ Bi-modal CLAP)	-	**7.9** ± 0.35	**7.1** ± 0.42	**6.2** ± 0.38

Quantitative Results

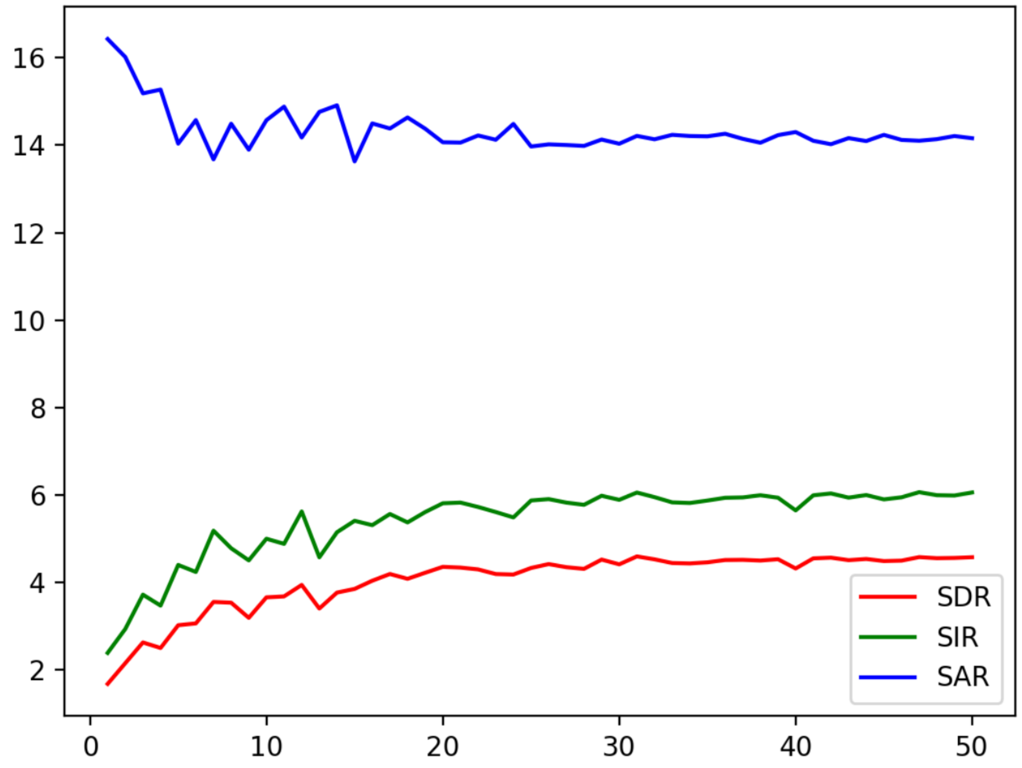
Table 2: Comparisons of the proposed semi-supervised learning with different portions of single-source and multi-source subsets. **Bold** and blue represents the best and second best performance.

Training Method	Test Set Mixture	Single-source Data: Dataset	Single-source Data: Fraction	Multi-source Mixture Data: Dataset	Multi-source Mixture Data: Fraction	Multi-source Mixture Data: #Source	Performance (SDR)
Supervised	MUSIC-2Mix	MUSIC	100%	-	-	-	8.1 ± 0.31
Supervised	MUSIC-2Mix	MUSIC	5%	-	-	-	2.6 ± 0.33
Unsupervised	MUSIC-2Mix	-	-	MUSIC	100%	2	7.9 ± 0.35
Semi-Supervised	MUSIC-2Mix	MUSIC	5%	MUSIC	95%	2	8.8 ± 0.28
Semi-Supervised	MUSIC-2Mix	MUSIC	5%	MUSIC	95%	3	8.2 ± 0.22
Semi-Supervised	MUSIC-2Mix	MUSIC	5%	MUSIC	95%	4	7.4 ± 0.31
Semi-Supervised	MUSIC-2Mix	MUSIC	10%	MUSIC	90%	2	8.9 ± 0.26
Semi-Supervised	MUSIC-2Mix	MUSIC	25%	MUSIC	75%	2	9.2 ± 0.24
Semi-Supervised	MUSIC-2Mix	MUSIC	75%	MUSIC	25%	2	9.5 ± 0.29
Semi-Supervised	MUSIC-2Mix	MUSIC	100%	VGGSound	100%	2	**9.9** ± 0.35
Semi-Supervised	MUSIC-2Mix	VGGSound	100%	MUSIC	100%	2	9.7 ± 0.35
Semi-Supervised	MUSIC-2Mix	VGGSound	100%	MUSIC	100%	3	9.2 ± 0.31
Semi-Supervised	MUSIC-2Mix	VGGSound	100%	MUSIC	100%	4	8.9 ± 0.42
Supervised	VGGSound-2Mix	VGGSound	100%	-	-	-	2.3 ± 0.23
Supervised	VGGSound-2Mix	VGGSound	5%	-	-	-	0.4 ± 0.35
Unsupervised	VGGSound-2Mix	-	-	VGGSound	100%	2	2.2 ± 0.29
Semi-Supervised	VGGSound-2Mix	VGGSound	5%	VGGSound	95%	2	3.1 ± 0.31
Semi-Supervised	VGGSound-2Mix	VGGSound	75%	VGGSound	25%	2	**3.4** ± 0.26
Unsupervised	AudioCaps-2Mix	-	-	AudioCaps	100%	1∼6	2.9 ± 0.23
Semi-Supervised	AudioCaps-2Mix	VGGSound	100%	AudioCaps	100%	1∼6	**4.3** ± 0.34

Test Metric Plot Over training Iterations



Proposed (ICLR '24)

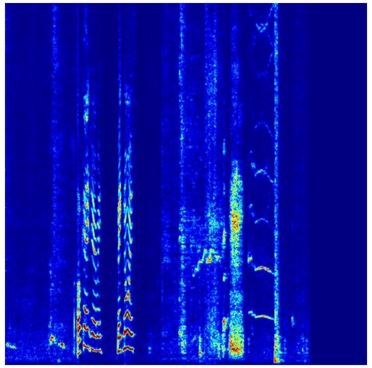


CLIPSep (ICLR '23)

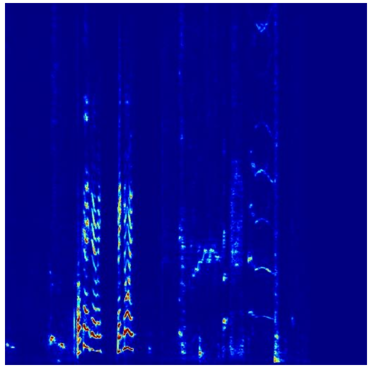
Large increase of SDR and SIR denote better separation quality with significant reduction of interference noises from other sources

Qualitative Results (Natural Mixtures)

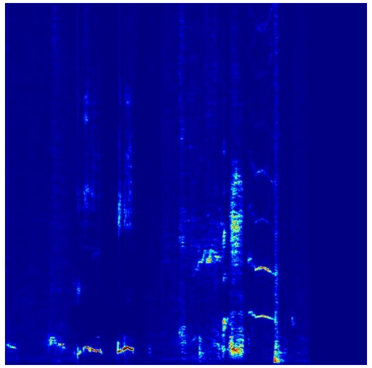
Mixture

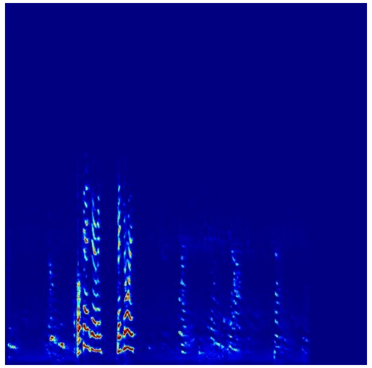


"A woman speaks" | "A cat crying"

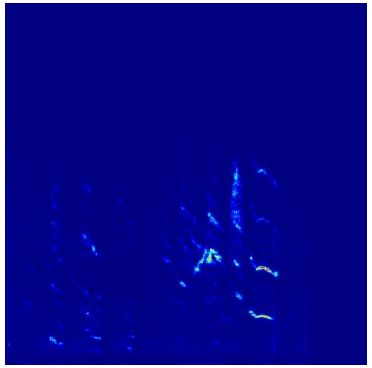


CLIPSep (ICLR '23)



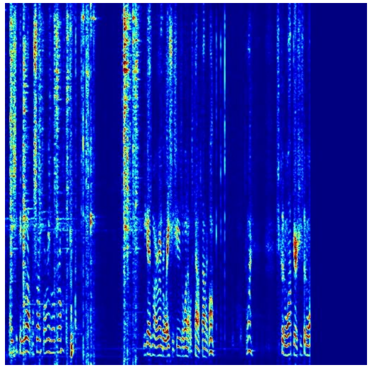


Ours (ICLR '24)



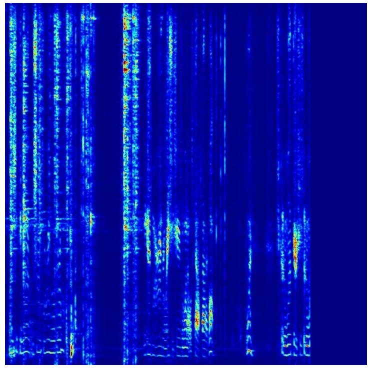
Qualitative Results (Natural Mixtures)

Mixture

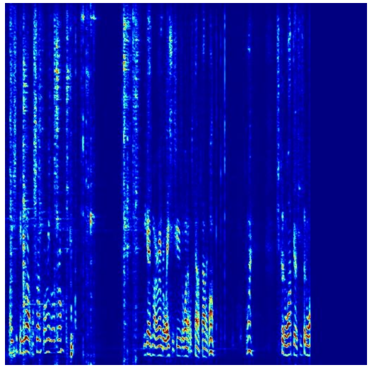


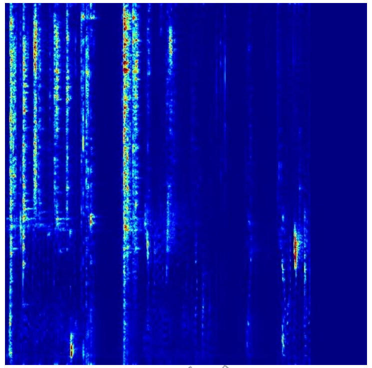
“Metal Clashses”

“A man speaks”

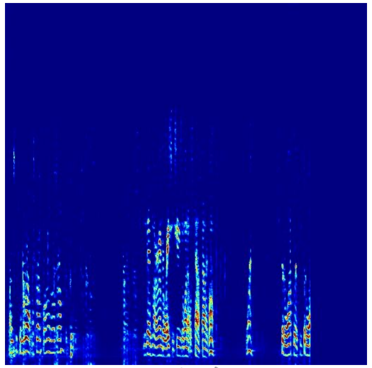


CLIPSep (ICLR ‘23)



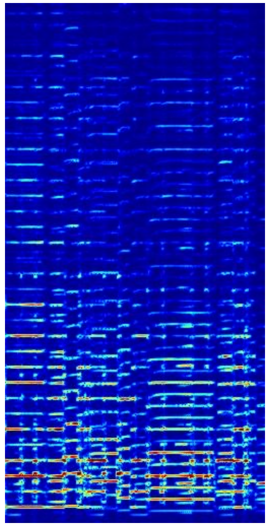


Ours (ICLR ‘24)

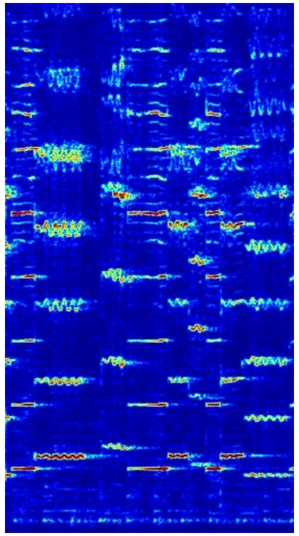


Qualitative Results (Synthetic Mixtures)

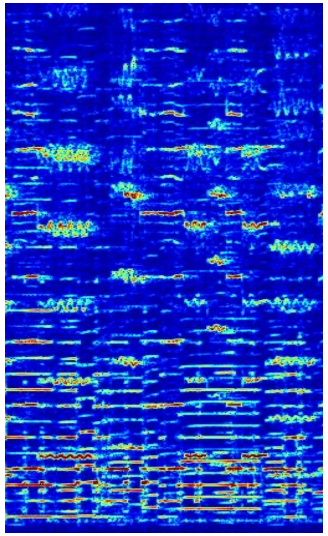
“Accordion”



“Violin”

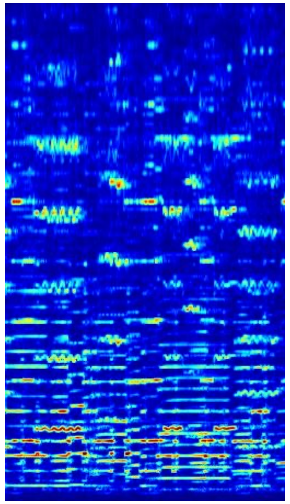


“Mixture”

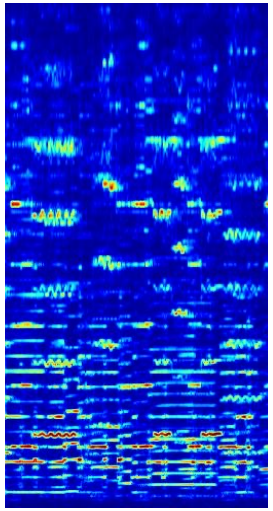


Input Data

“Accordion”

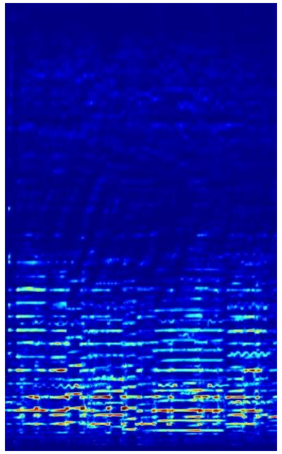


“Violin”

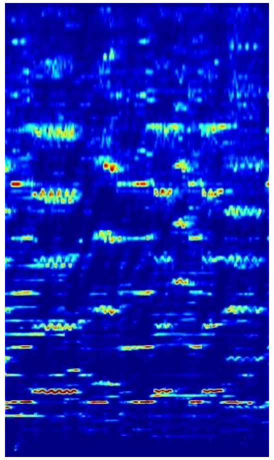


“Baseline”

“Accordion”



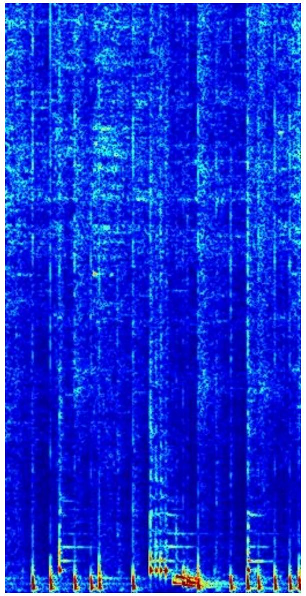
“Violin”



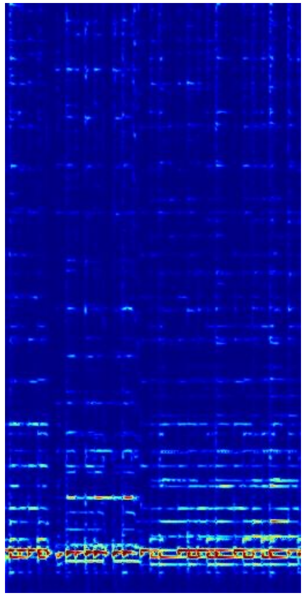
“Proposed”

Qualitative Results (Synthetic Mixtures)

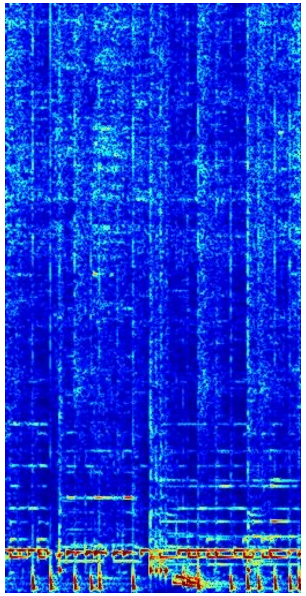
“Drum”



“Ukulele”



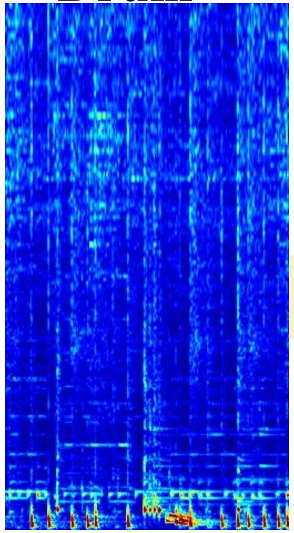
“Mixture”



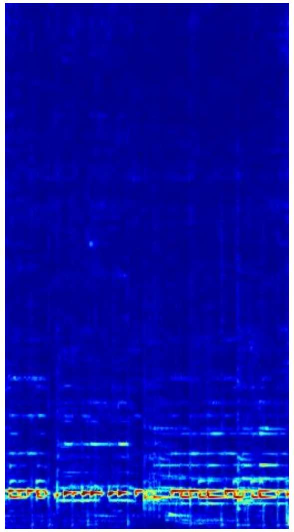
Input Data

“Baseline”

“Drum”

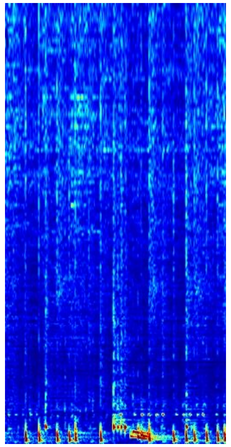


“Ukulele”

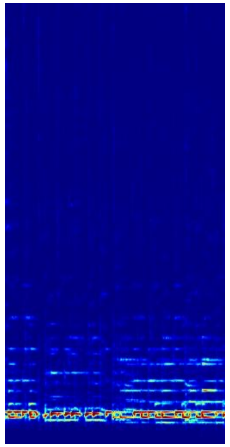


“Proposed”

“Drum”



“Ukulele”



Future Works

- **Unconditional source separation**
  - With no external text inputs
  - With new unseen audio classes
- **Joint editing and audio generation**
  - Leverage generative models for joint audio generation and editing
  - Training-free/with minimal training
- **Multi-modal fine-grained conditioning with videos in natural mixtures**
  - Automatic separation of sounds from videos

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Energy Aware Computing Research Group

Thank you!

Questions

Microsoft