ICLR

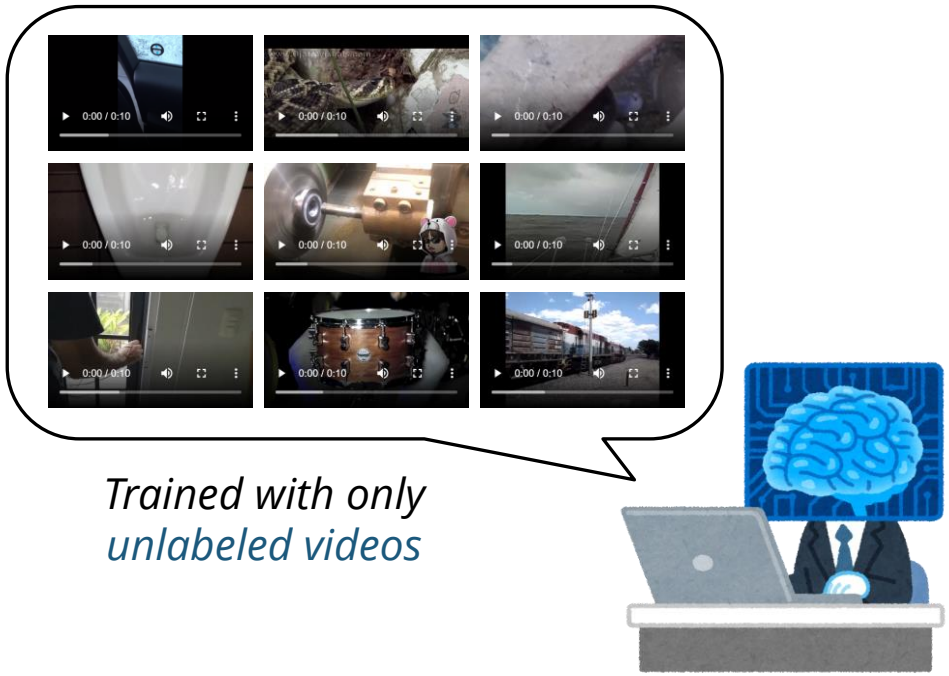
CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos

Hao-Wen Dong[1,2] * Naoya Takahashi[1] † Yuki Mitsufuji[1]
Julian McAuley[2] Taylor Berg-Kirkpatrick[2]
[1]Sony Group Corporation [2]University of California San Diego
* Work done during an internship at Sony † Corresponding author

SONY UC San Diego

Self-supervised Text-queried Sound Separation

Training



Trained with only unlabeled videos

Scalable to larger dataset

Inference

Natural text query-based interface

Data

MUSIC
(Zhao et al., 2018)

Violin

Acoustic guitar

Accordion

VGGSound
(Chen et al., 2020)

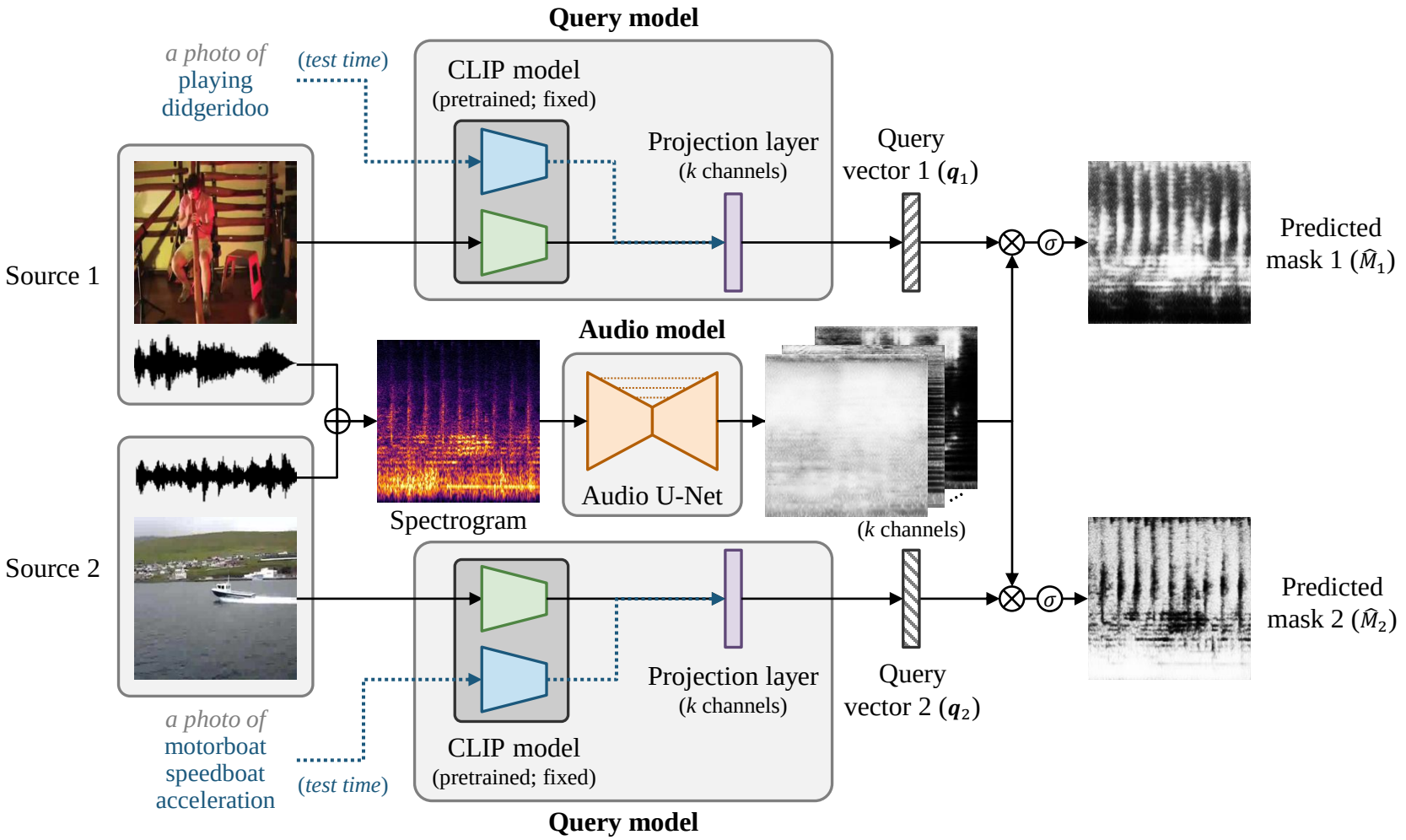
Hedge trimmer running

Dog bow-wow

Bird chirping, tweeting

Zhao et al., "The Sound of Pixels," *Proc. ECCV*, 2018.
Chen et al., "VGGSound: A Large-Scale Audio-Visual Dataset," *Proc. ICASSP*, 2020.

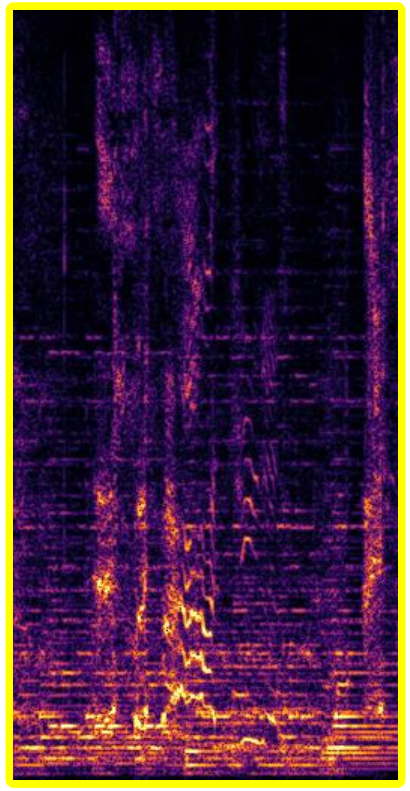
CLIPSep


Query model
a photo of
playing
didgeridoo
(test time)
CLIP model
(pretrained; fixed)
Projection layer
(k channels)
Query
vector 1 (q_1)
Source 1
Audio model
Spectrogram
Audio U-Net
(k channels)
Predicted
mask 1 ($\hat{M}_1$)
Source 2
Projection layer
(k channels)
Query
vector 2 (q_2)
Predicted
mask 2 ($\hat{M}_2$)
a photo of
motorboat
speedboat
acceleration
(test time)
CLIP model
(pretrained; fixed)
Query model

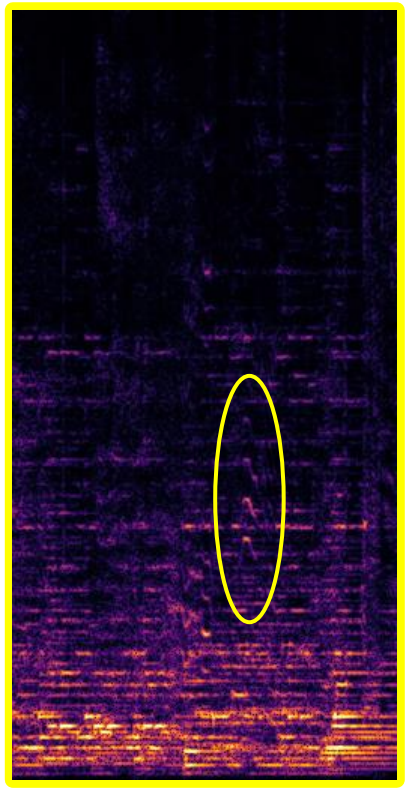
Demo – CLIPSep

Query: "*playing harpsichord*"

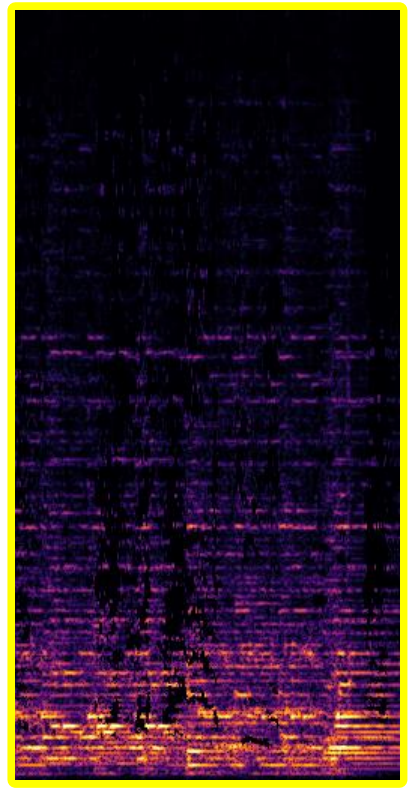
Mixture



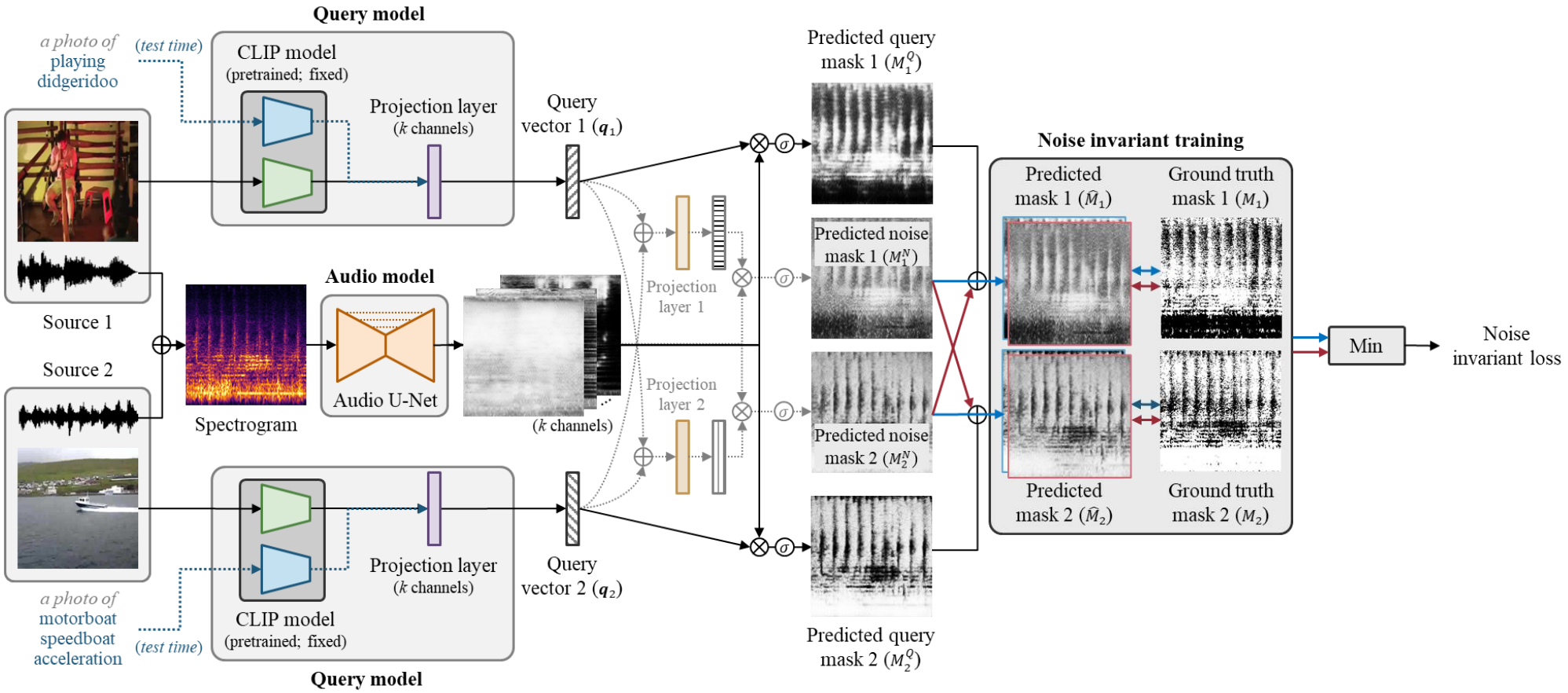
CLIPSep



Ground truth

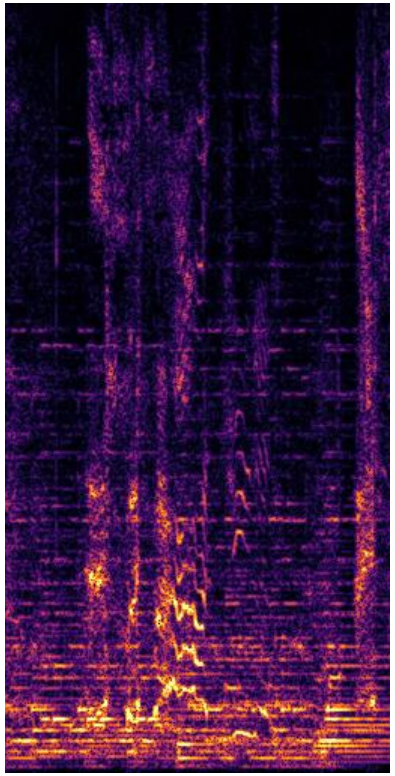
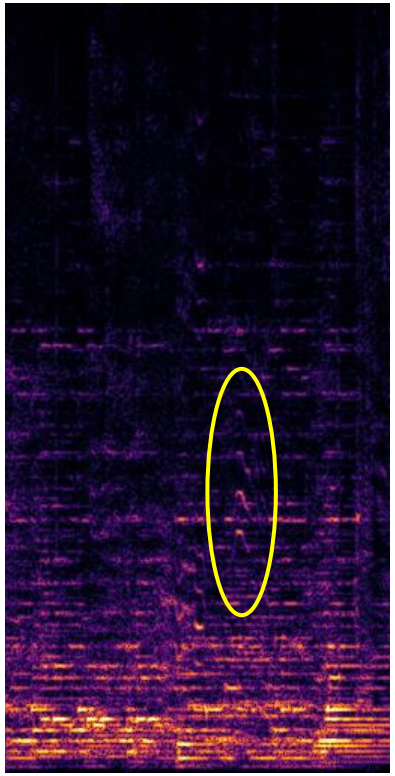
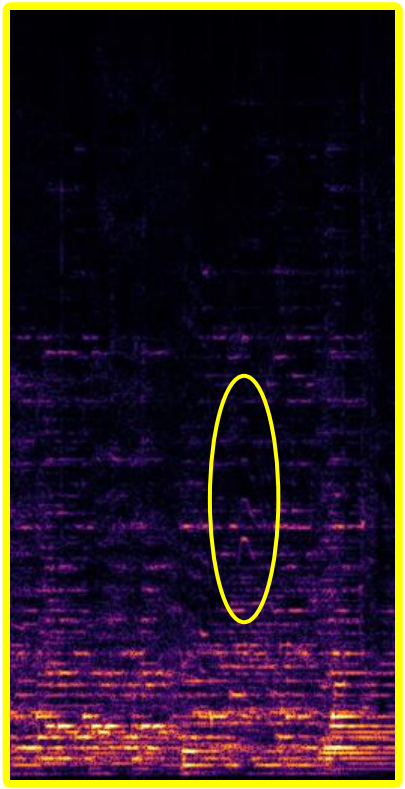
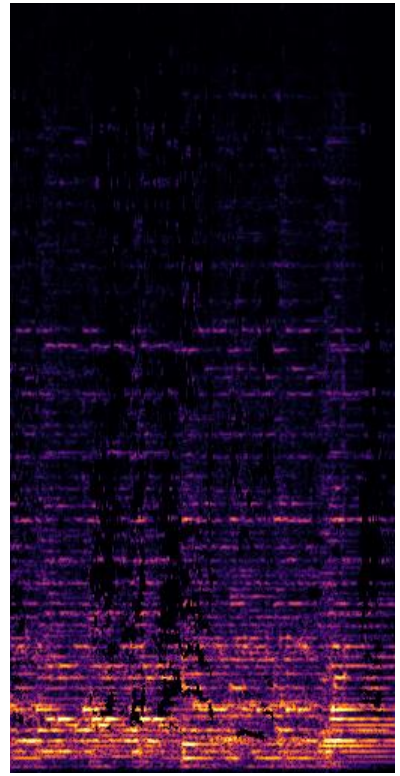


Noise Invariant Training (NIT)



Demo – CLIPSep-NIT

Query: "*playing harpsichord*"

Mixture	CLIPSep	CLIPSep-NIT	Ground truth
			

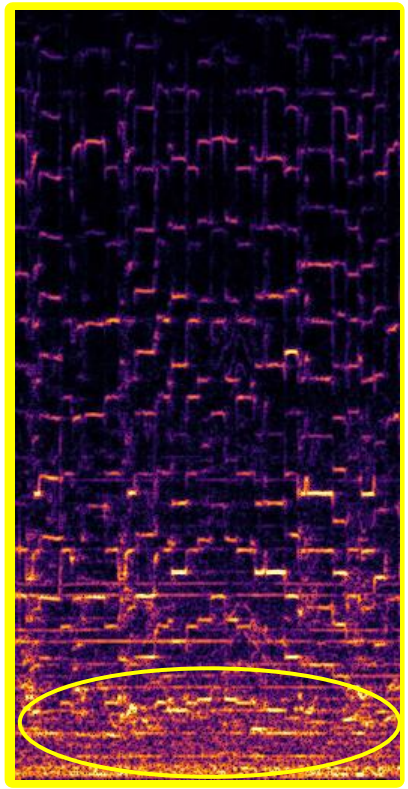
Quantitative Results

Model	Unlabeled data	Post-proc. free	MUSIC+ Mean SDR	MUSIC+ Median SDR	VGGSound-Clean+ Mean SDR	VGGSound-Clean+ Median SDR
Mixture	-	-	4.49 ± 1.41	2.04	-0.77 ± 1.31	-0.84
Text-queried models						
CLIPSep	✓	✓	9.71 ± 1.21	8.73	2.76 ± 1.00	**3.95**
CLIPSep-NIT	✓	✓	**10.27 ± 1.04**	**10.02**	**3.05 ± 0.73**	3.26
BERTSep		✓	4.67 ± 0.44	4.41	5.09 ± 0.80	5.49
CLIPSep-Text		✓	10.73 ± 0.99	9.93	5.49 ± 0.82	5.06

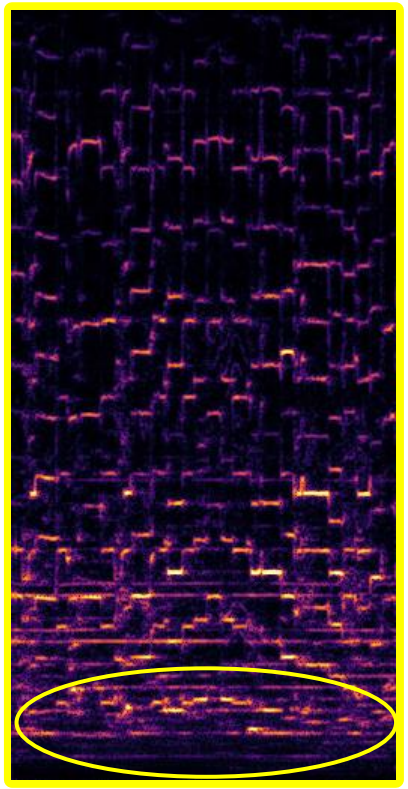
Demo – Noise Removal

Query: "*playing bagpipe*"

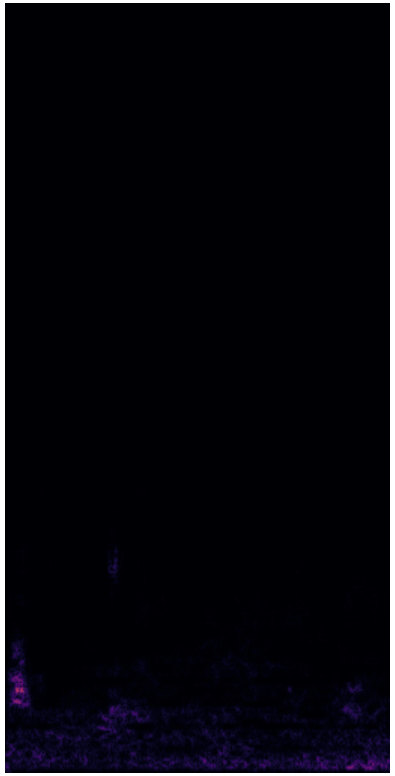
Mixture



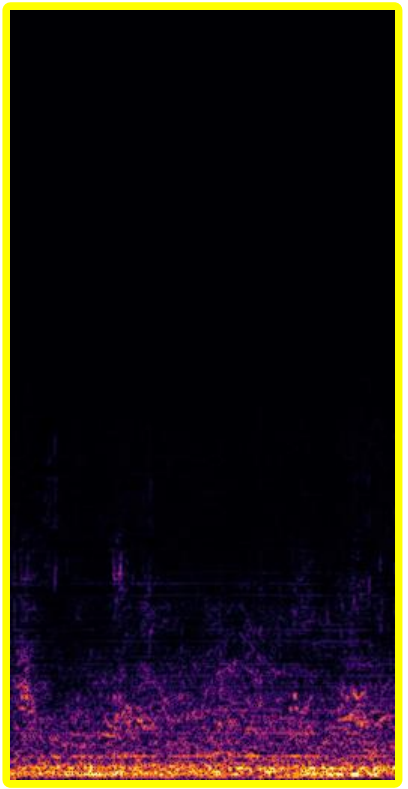
Prediction



Noise head 1



Noise head 2



Summary

CLIPSep

First text-queried universal sound separation model that can be trained **using only unlabeled videos**

Noise Invariant Training

An approach for training a query-based sound separation model with **noisy data in the wild**

Paper: arxiv.org/abs/2212.07065
Demo: sony.github.io/CLIPSep/
Code: github.com/sony/CLIPSep