Learning Sound Separation and Synthesis from Videos using Pretrained Language-vision Models

Hao-Wen (Herman) Dong

University of California San Diego

UC San Diego

About Me

Hi, I'm Herman.
I do AI x Music research.
I love music and movies!

國立臺灣大學 National Taiwan University
2013 – 2017
B.S. in Electrical Engineering

中央研究院 ACADEMIA SINICA
2017 – 2019
Research Assistant

UC San Diego
2019 – 2021
M.S. in Computer Science

UC San Diego
2019 – *present*
Ph.D. in Computer Science (expected)

Summer 2019
YAMAHA
Research Intern

Summer 2021
Dolby
Deep Learning Audio Intern

Summer 2022
SONY
Student Intern

Fall 2022
amazon
Applied Scientist Intern

Winter 2023
Dolby
Speech/Audio Deep Learning Intern

Summer 2023
Adobe
Research Scientist/Engineer Intern

Fall 2023
NVIDIA
Research Intern

My Research

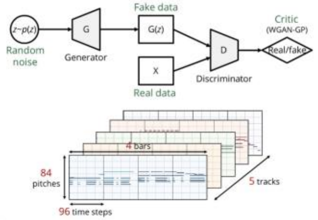
AI × Music

Multitrack Music Generation

Generating new music contents automatically

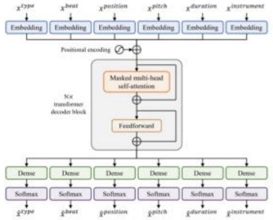
MuseGAN

(AAAI 2018)



Multitrack Music Transformer

(ICASSP 2023)

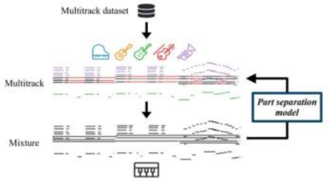

Embedding
Positional encoding
Masked multi-head self-attention
Feedforward
N× transformer decoder block
Dense
Softmax

Assistive Music Creation Tools

Assisting humans to create and perform music

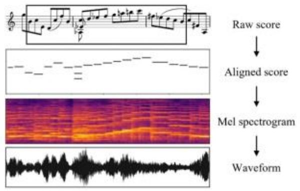
Arranger

(ISMIR 2021)



Deep Performer

(ICASSP 2022)

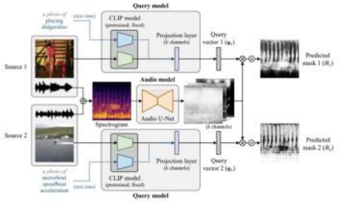


Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

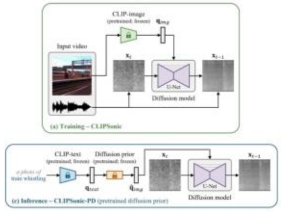
CLIPSep

(ICLR 2023)



CLIPSonic

(WASPAA 2023)



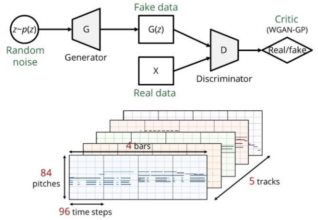
My Research

AI × Music

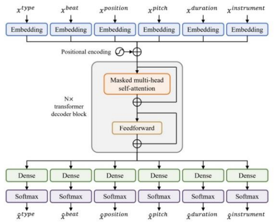
Multitrack Music Generation

Generating new music contents automatically

MuseGAN
(AAAI 2018)



Multitrack Music Transformer
(ICASSP 2023)



Assistive Music Creation Tools

Assisting humans to create and perform music

Arranger
(ISMIR 2021)

Deep Performer
(ICASSP 2022)

Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep
(ICLR 2023)

CLIPSonic
(WASPAA 2023)

Featured in Amazon AWS DeepComposer

Pop music generation

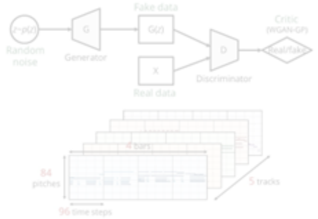
My Research

Multitrack Music Ge

Generating new music contents automatically

MuseGAN (AAAI 2018)

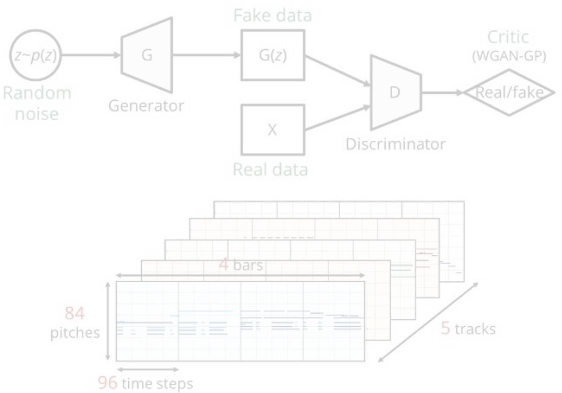
Multit



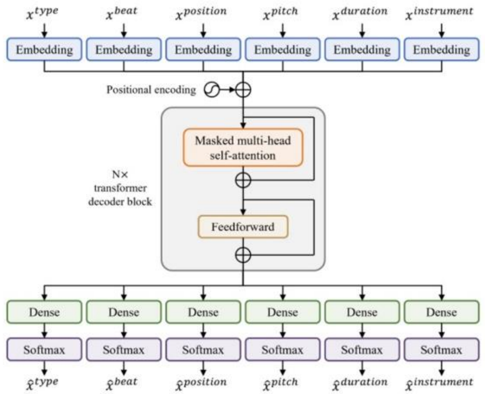
Multitrack Music Generation

Generating new music contents automatically

MuseGAN
(AAAI 2018)



Multitrack Music Transformer
(ICASSP 2023)



Orchestral music generation

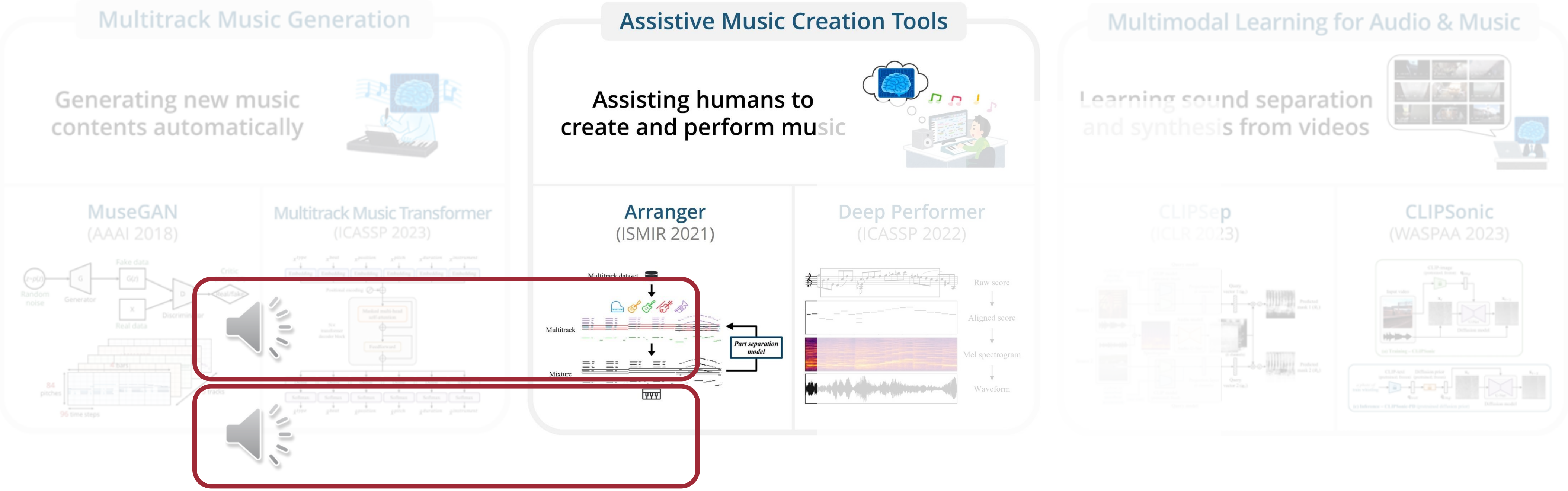
earning for Audio & Music

separation from videos

CLIPSonic (WASPAA 2023)

My Research

AI × Music



Multitrack Music Generation

Generating new music contents automatically

MuseGAN
(AAAI 2018)

Multitrack Music Transformer
(ICASSP 2023)

Assistive Music Creation Tools

Assisting humans to create and perform music

Arranger
(ISMIR 2021)

Deep Performer
(ICASSP 2022)

Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep
(ICLR 2023)

CLIPSonic
(WASPAA 2023)

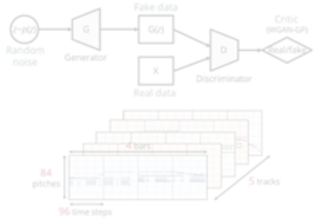
Automatic instrumentation

My Research

Multitrack Music Ge

Generating new music contents automatically

MuseGAN
(AAAI 2018)

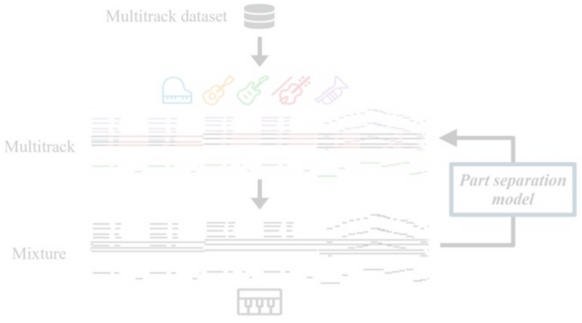


Multit

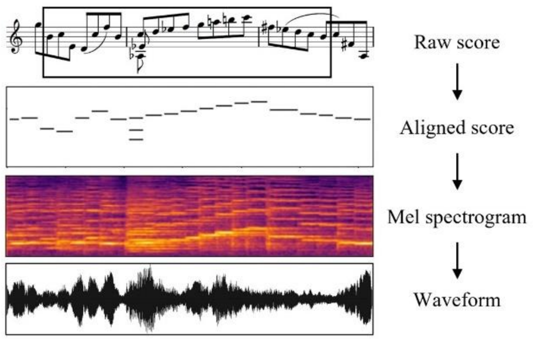
Assistive Music Creation Tools

Assisting humans to create and perform music

Arranger
(ISMIR 2021)



Deep Performer
(ICASSP 2022)

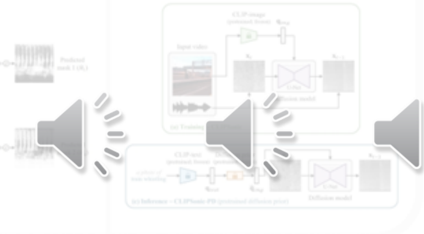


Score-to-audio synthesis

earning for Audio & Music

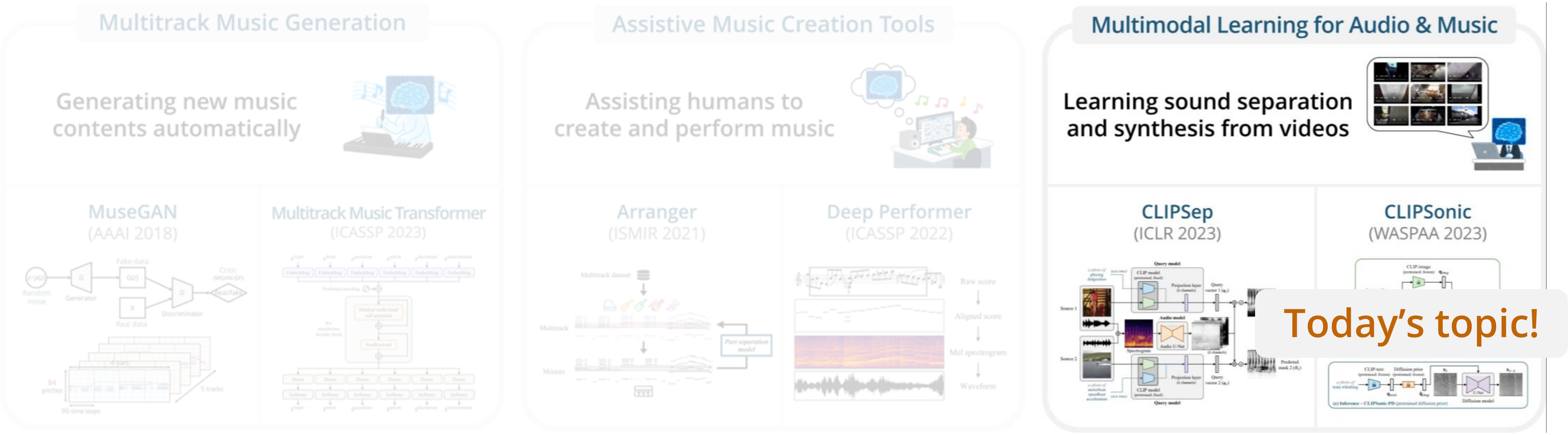
separation
om videos

CLIPSonic
(WASPAA 2023)



My Research

AI × Music



Multitrack Music Generation

Generating new music contents automatically

MuseGAN (AAAI 2018)

Multitrack Music Transformer (ICASSP 2023)

Assistive Music Creation Tools

Assisting humans to create and perform music

Arranger (ISMIR 2021)

Deep Performer (ICASSP 2022)

Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep (ICLR 2023)

CLIPSonic (WASPAA 2023)

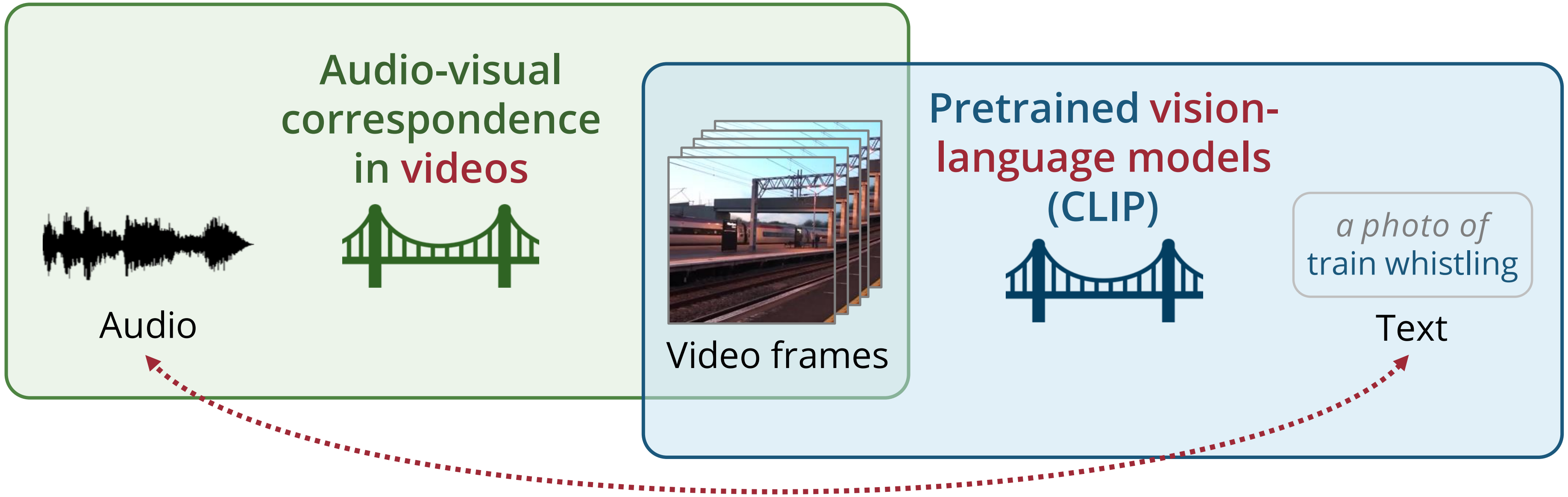
Today's topic!

Text-queried sound separation

Text-to-audio synthesis

Introduction

Leveraging the Visual Domain as a Bridge

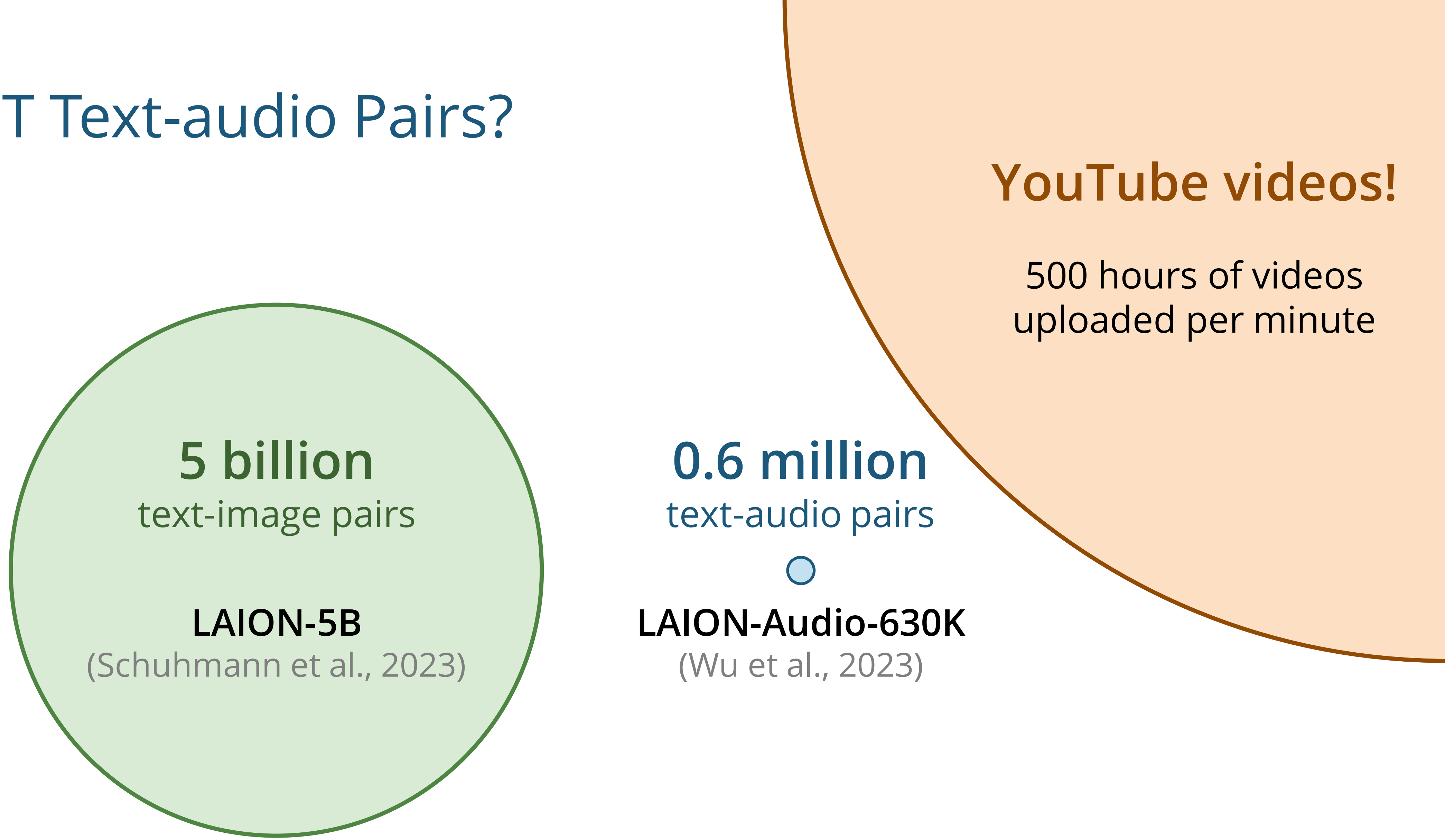

Audio-visual correspondence in videos
Audio
Video frames
Pretrained vision-language models (CLIP)
a photo of train whistling
Text

Desired audio-text correspondence

No text-audio pairs required!

Scalable to large video datasets!

Why NOT Text-audio Pairs?



YouTube videos!

500 hours of videos uploaded per minute

5 billion
text-image pairs

LAION-5B
(Schuhmann et al., 2023)

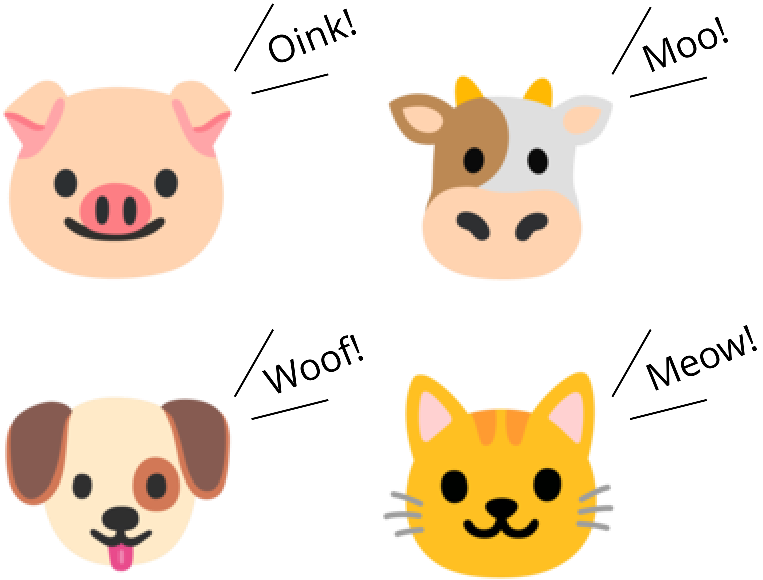
0.6 million
text-audio pairs

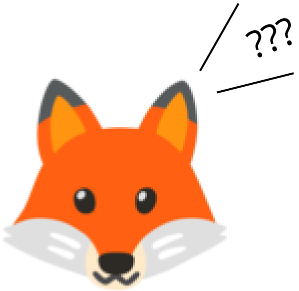
LAION-Audio-630K
(Wu et al., 2023)

Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," *NeurIPS, Datasets and Benchmarks Track*, 2023.
Wu et al., "Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," *ICASSP*, 2023.

Learning Sounds from Videos

- Watching a dog barking, humans can *associate the barking sound to the dog*
- Can machines learn to synthesize sounds from watching *noisy* videos?

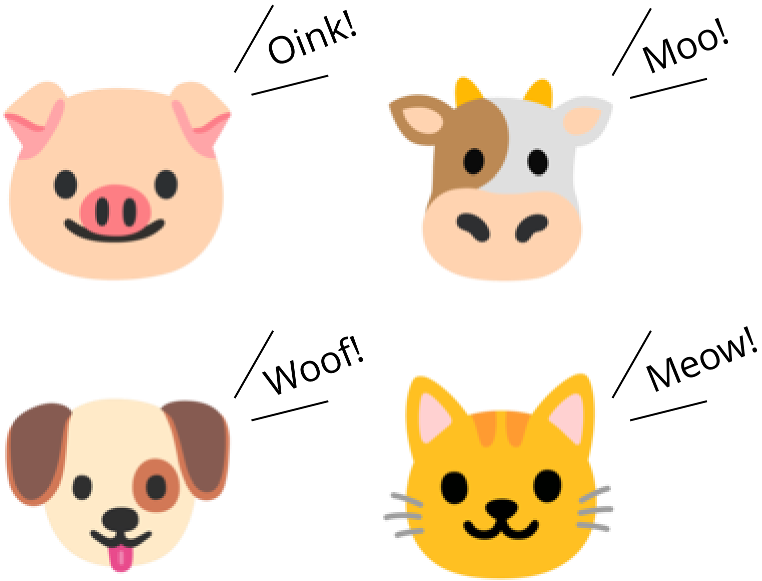




What does the fox say?

Learning Sounds from Videos

- Watching a dog barking, humans can *associate the barking sound to the dog*
- Can machines learn to synthesize sounds from watching *noisy* videos?



[redacted]

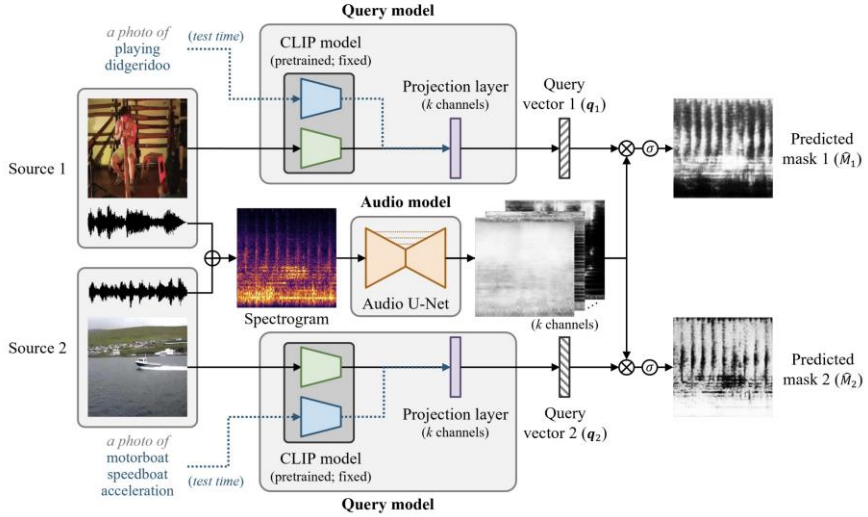
URURU TV, "Excited Fox Noises - Fox Laughing and Squeaking," *YouTube*, https://youtu.be/XiCdBFDfUTg, 2019.

Overview

CLIPSep

(Dong et al., ICLR 2023)

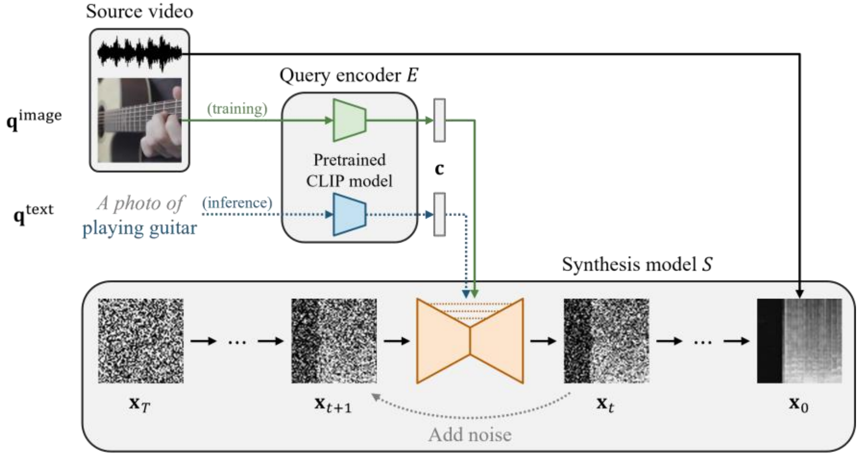
For text-queried sound separation



CLIPSonic

(Dong et al., WASPAA 2023)

For text-to-audio synthesis



Dong et al., "CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos," *ICLR*, 2023.
Dong et al., "CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models," *WASPAA*, 2023.

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

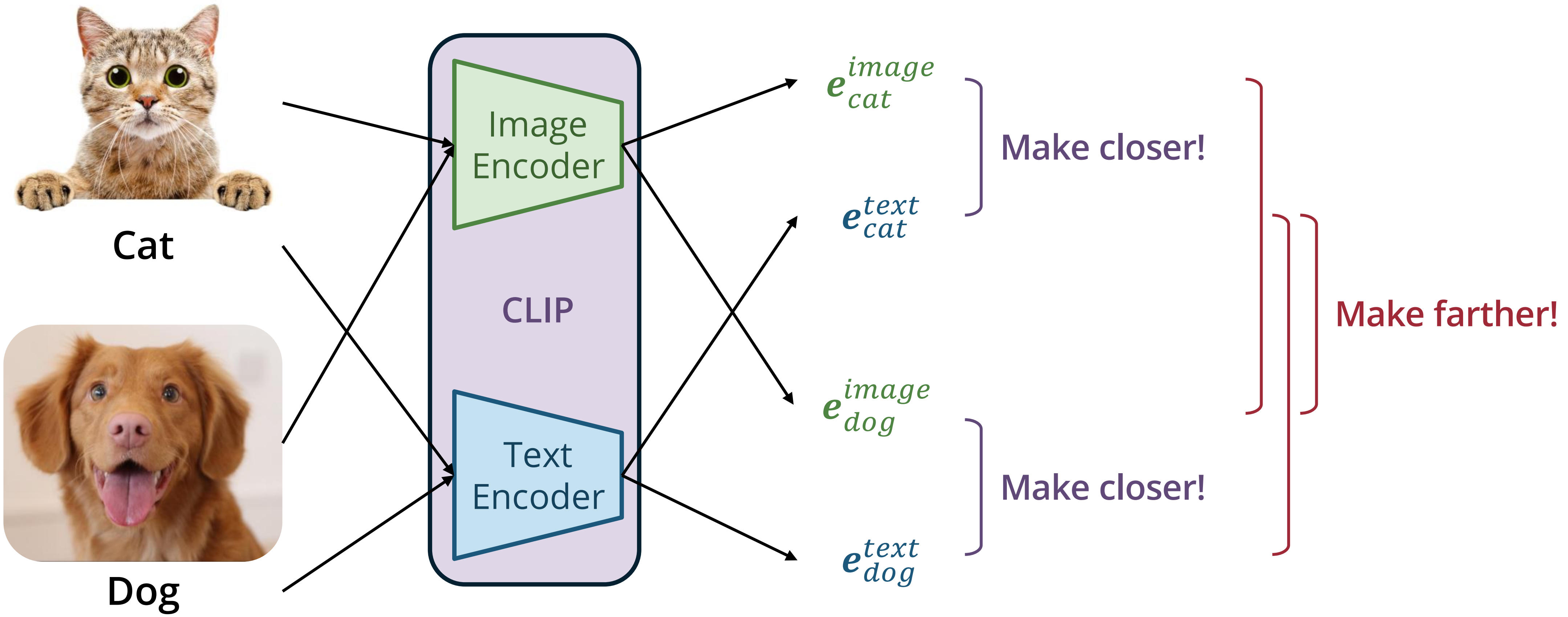
Overview – Text-queried Sound Separation

More samples

salu133445.github.io/clipsep

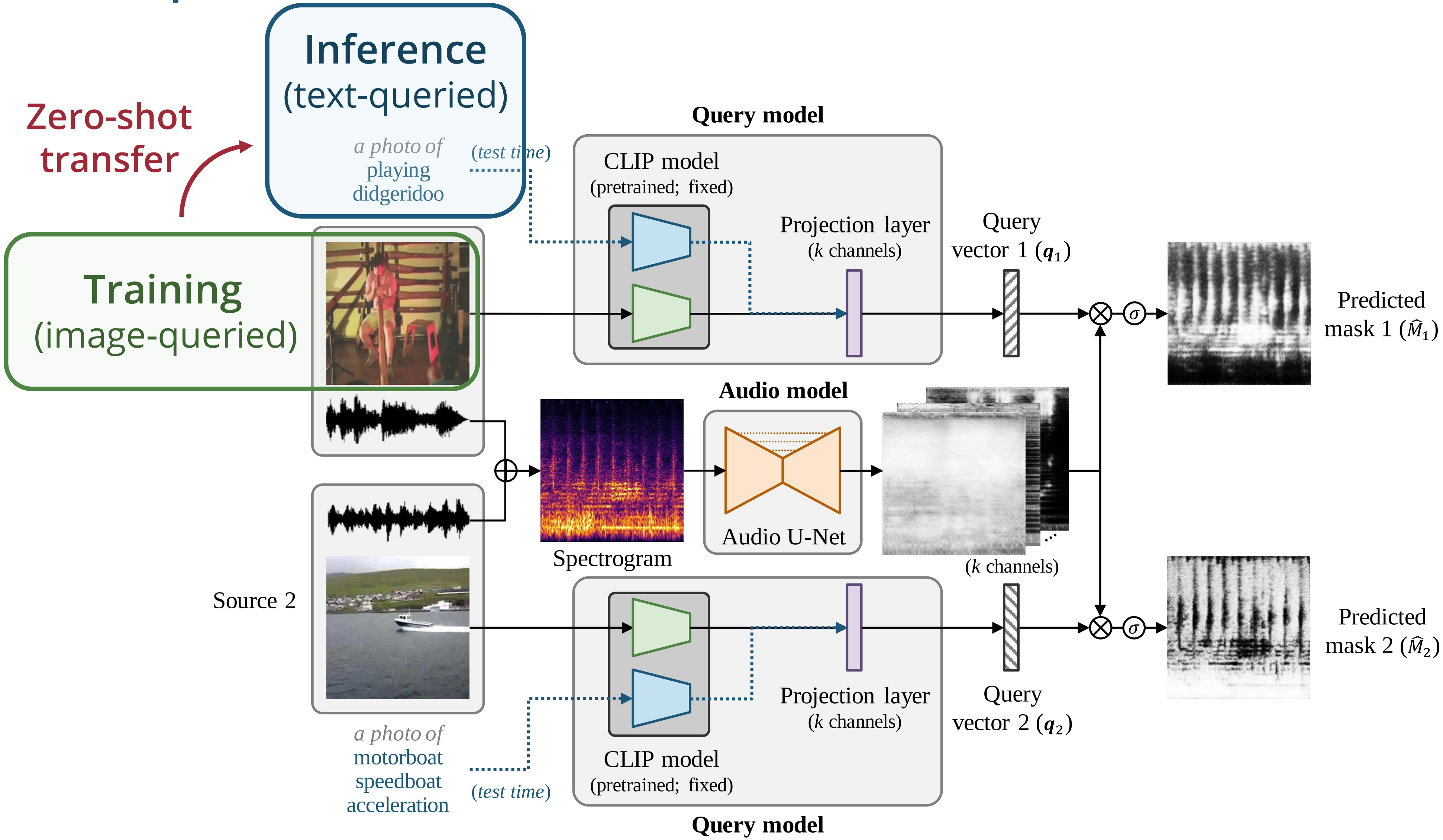
CLIP (Contrastive Language-Image Pretraining)

- Learn a shared embedding space for images and texts via *contrastive learning*



Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," *ICML*, 2021.

CLIPSep



Zero-shot transfer

Inference (text-queried)

Training (image-queried)

Data

MUSIC

(Zhao et al., 2018)

Violin

Acoustic guitar

Accordion

Music instrument playing videos

(1,055 videos, 21 instruments)

VGGSound

(Chen et al., 2020)

Hedge trimmer running

Dog bow-wow

Bird chirping, tweeting

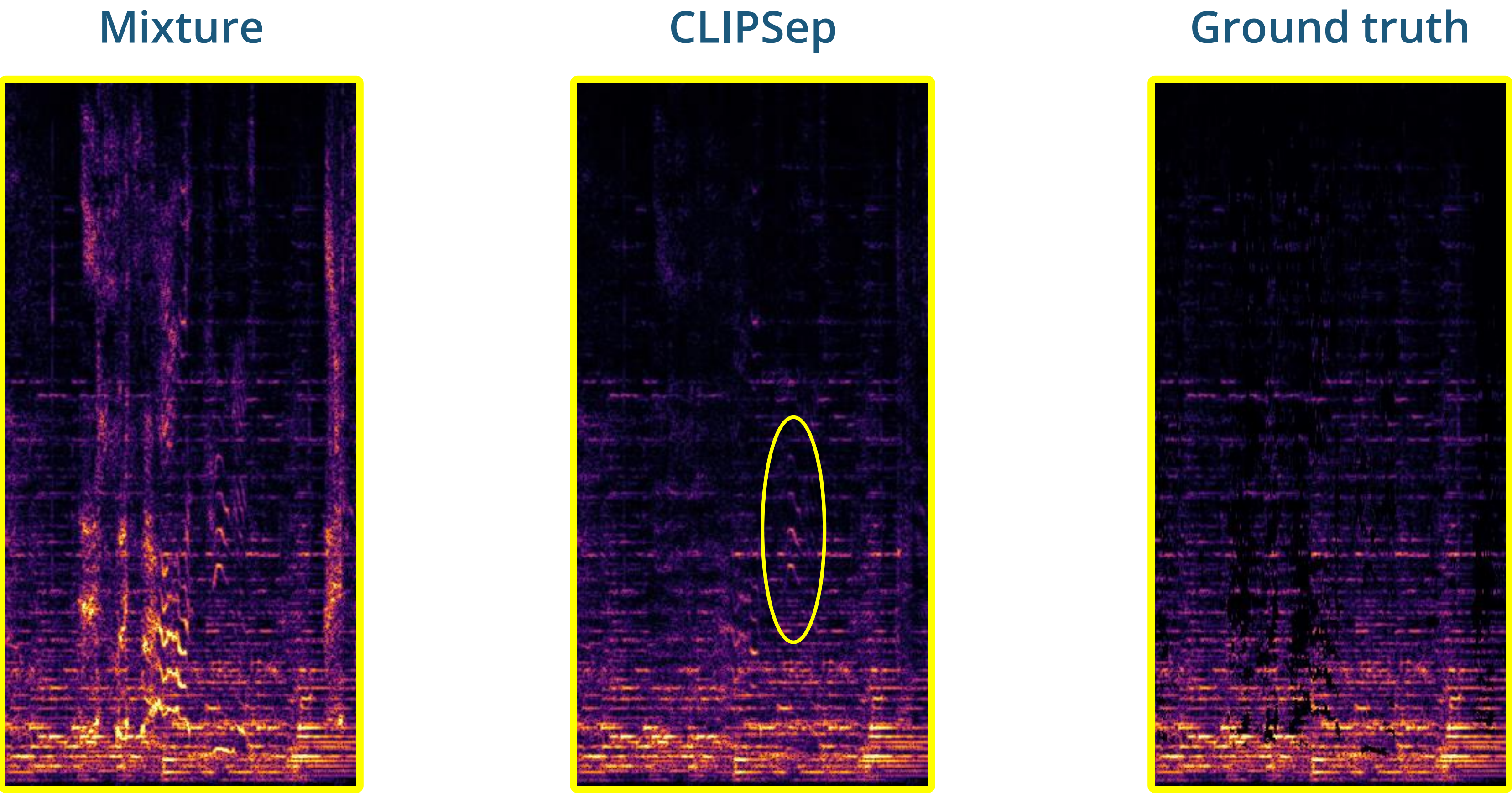
Noisy videos with diverse sounds

(172K videos, 310 classes)

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

Demo – CLIPSep

Query: “*playing harpsichord*”

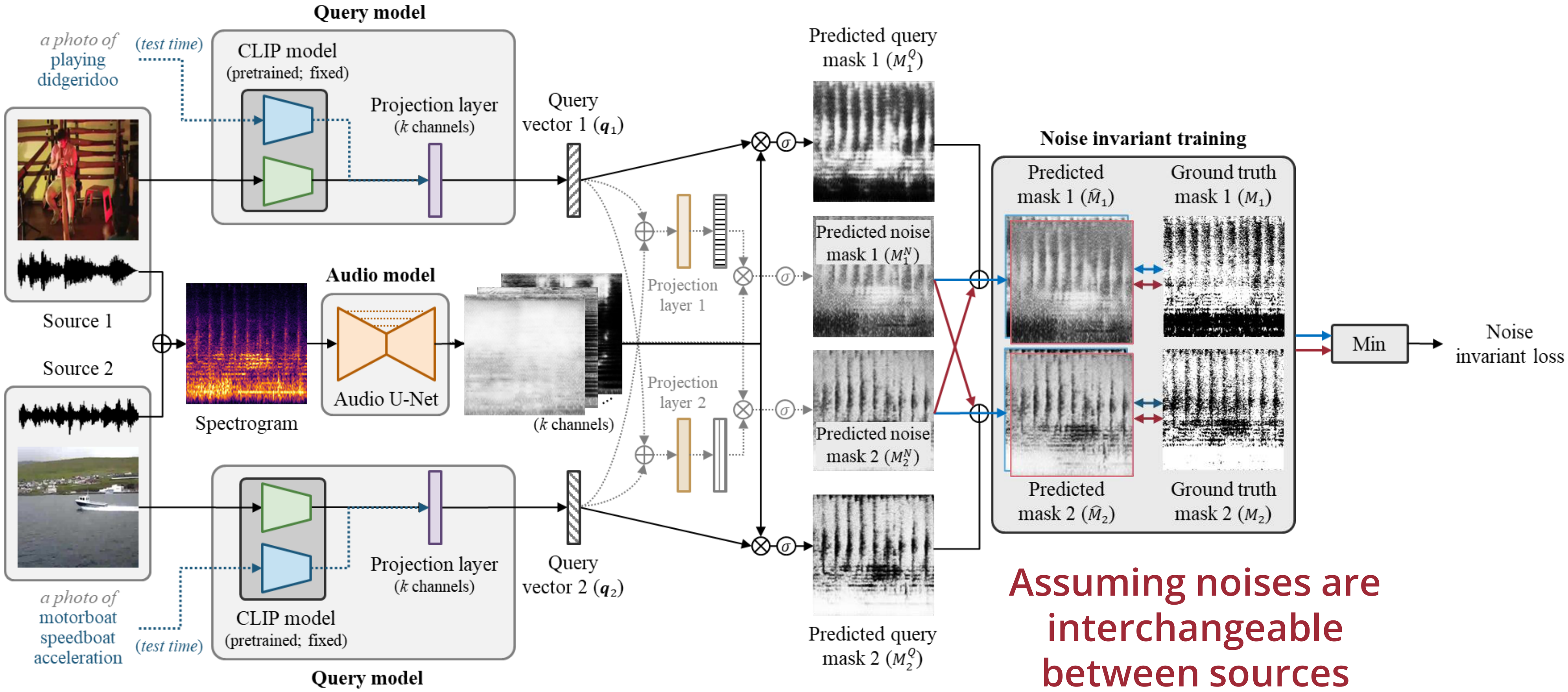


Mixture

CLIPSep

Ground truth

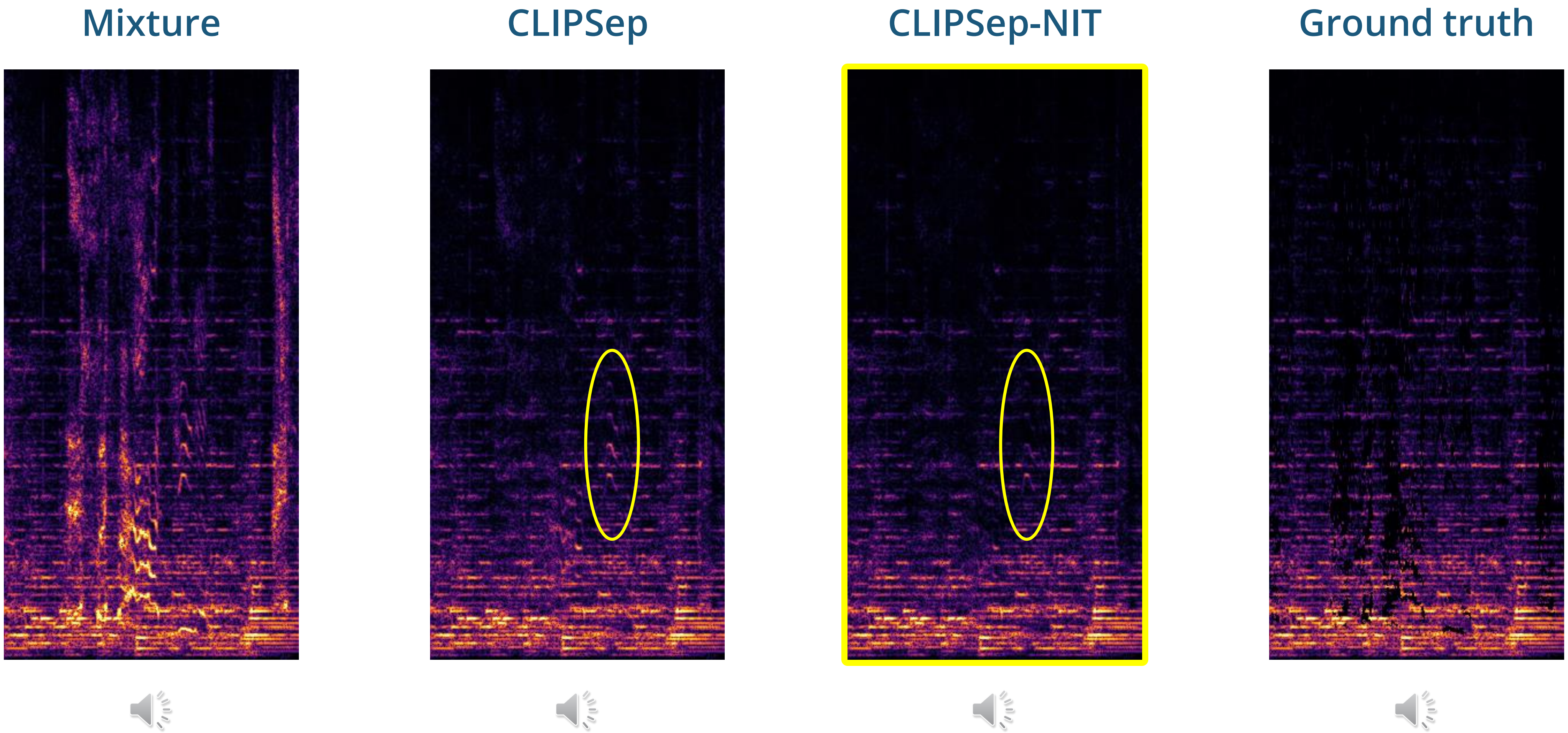
Noise Invariant Training (NIT)



Assuming noises are interchangeable between sources

Demo – CLIPSep-NIT

Query: “*playing harpsichord*”



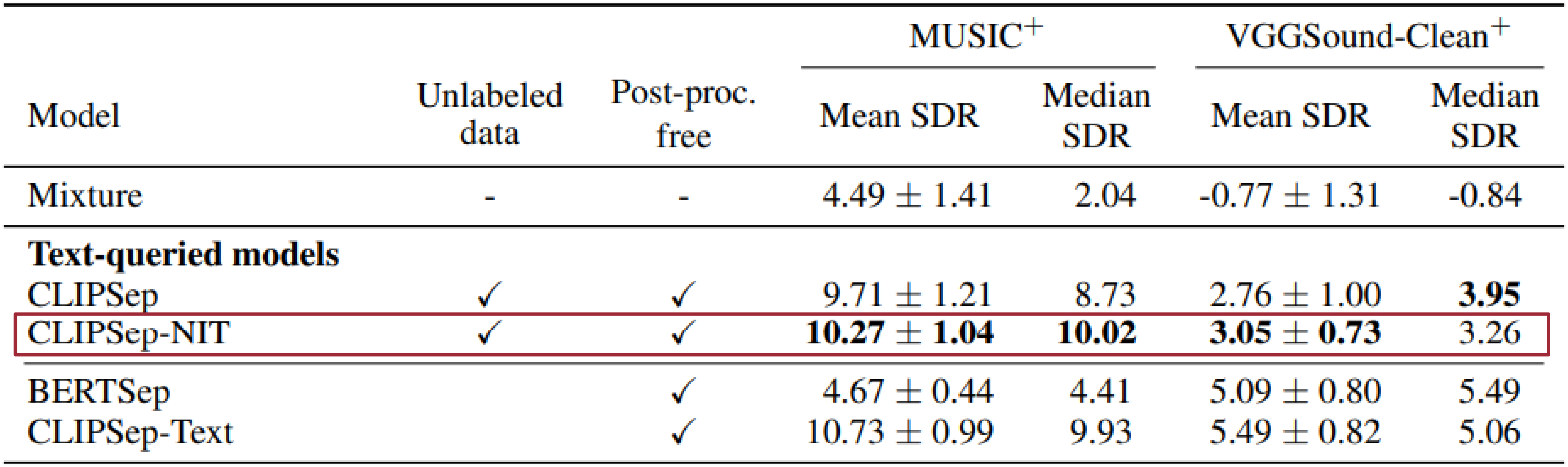
Mixture

CLIPSep

CLIPSep-NIT

Ground truth

Quantitative Results

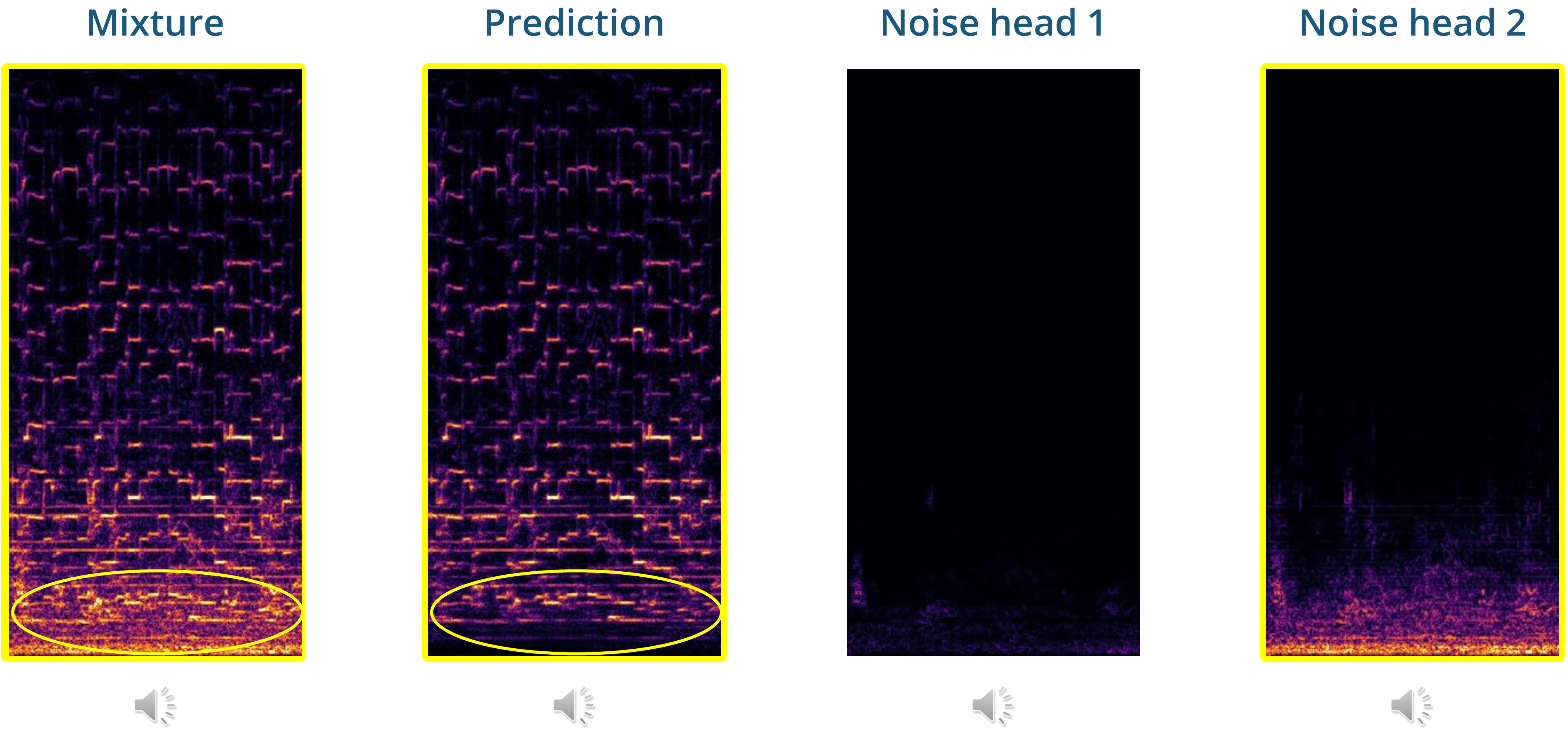


Model	Unlabeled data	Post-proc. free	MUSIC$^+$		VGGSound-Clean$^+$	
			Mean SDR	Median SDR	Mean SDR	Median SDR
Mixture	-	-	4.49 ± 1.41	2.04	-0.77 ± 1.31	-0.84
Text-queried models						
CLIPSep	✓	✓	9.71 ± 1.21	8.73	2.76 ± 1.00	**3.95**
CLIPSep-NIT	✓	✓	**10.27 ± 1.04**	**10.02**	**3.05 ± 0.73**	3.26
BERTSep		✓	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

Significant performance improvement against the baseline!

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

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

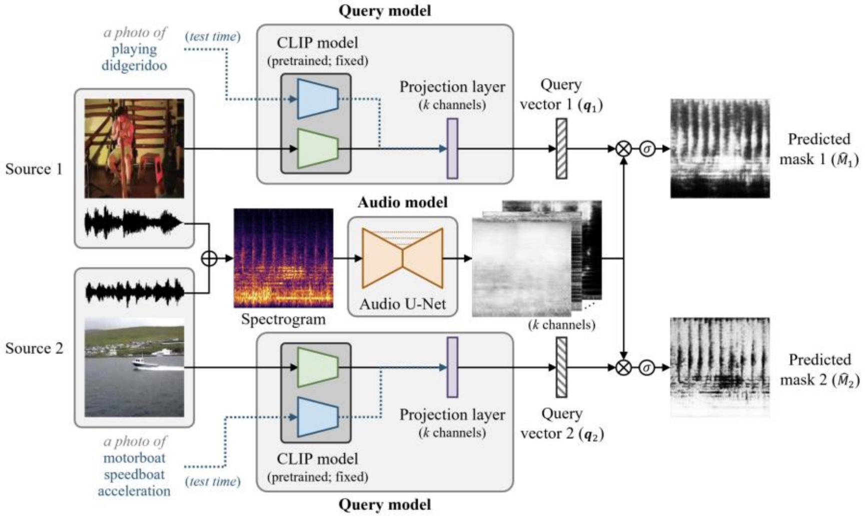
Paper: arxiv.org/abs/2212.07065
Demo: sony.github.io/CLIPSep/
Code: github.com/sony/CLIPSep

Overview

CLIPSep

(Dong et al., ICLR 2023)

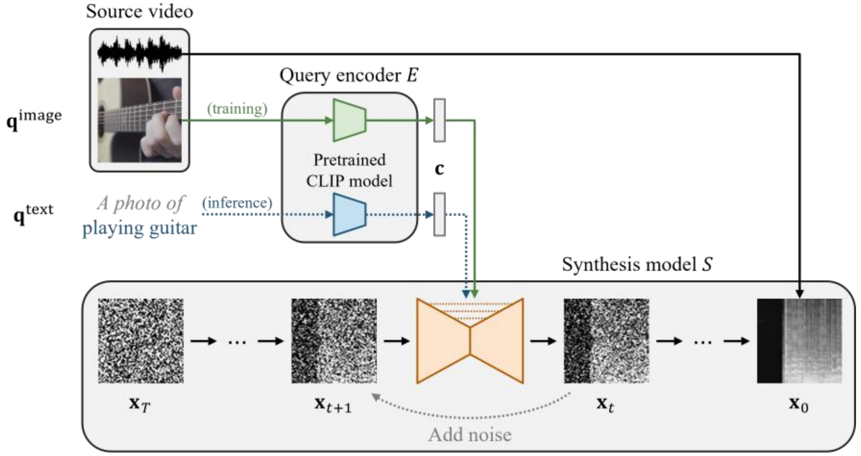
For text-queried sound separation



CLIPSonic

(Dong et al., WASPAA 2023)

For text-to-audio synthesis



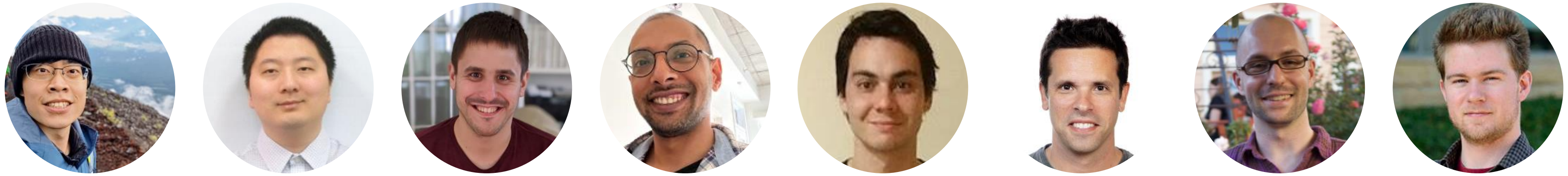
Dong et al., "CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos," *ICLR*, 2023.
Dong et al., "CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models," *WASPAA*, 2023.

IEEE
WASPAA
2023

CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models

Hao-Wen Dong[1,2]* Xiaoyu Liu[1] Jordi Pons[1] Gautam Bhattacharya[1]
Santiago Pascual[1] Joan Serrà[1] Taylor Berg-Kirkpatrick[2] Julian McAuley[2]

[1] Dolby Laboratories [2] University of California San Diego
* Work done during an internship at Dolby



Dolby UC San Diego

Overview – Text-to-Audio Synthesis

(These samples are generated by our proposed model.)

More samples

salu133445.github.io/clipsonic

Prior Work – Text-to-Audio Synthesis

- Diffsound (Yang et al., 2023)
- AudioGen (Kreuk et al., 2023)
- AudioLDM (Liu et al., 2023)
- Make-An-Audio (Huang et al., 2023)
- Noise2Music (Huang et al., 2023)
- MusicLM (Agostinelli et al., 2023)

→ All rely on large amounts of **text-audio training pairs**

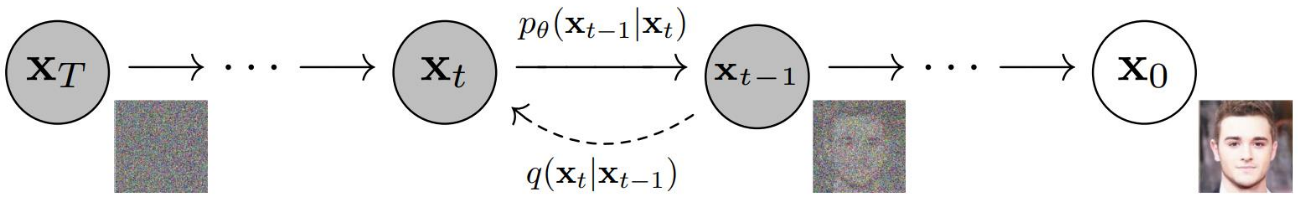
Can we learn text-to-audio synthesis *without* **using any text-audio pairs?**

Yang et al., "Diffsound: Discrete Diffusion Model for Text-to-sound Generation," *TASLP*, 2022.
Kreuk et al., "AudioGen: Textually Guided Audio Generation," *ICLR*, 2023.
Liu et al., "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models," *ICML*, 2023.
Huang et al., "Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models," *ICML*, 2023.
Huang et al., "Noise2Music: Text-conditioned Music Generation with Diffusion Models," *arXiv preprint arXiv:2302.03917*, 2023.
Agostinelli et al., "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2302.03917*, 2023.

Diffusion Model

Add noise gradually

(Forward diffusion process)

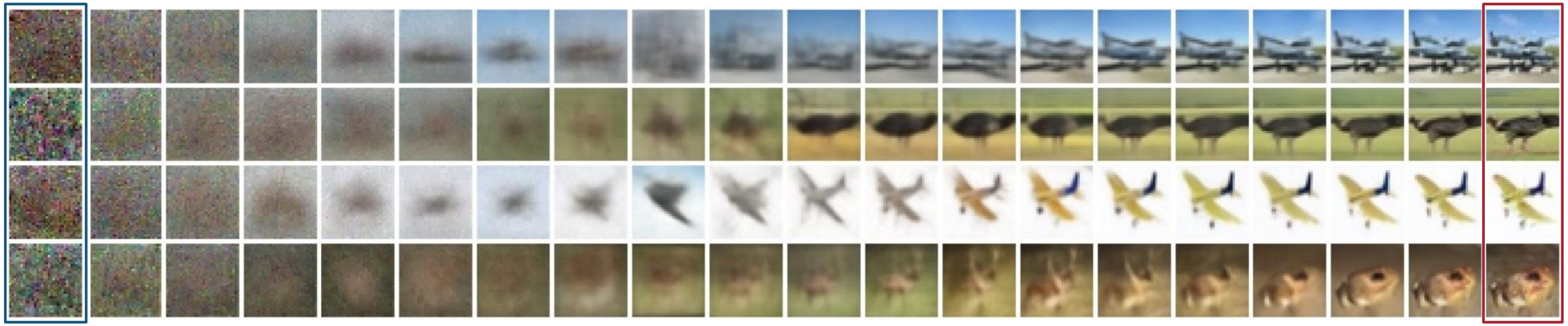

$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$\mathbf{x}_0$

Remove noise gradually

(Backward diffusion process)

Input

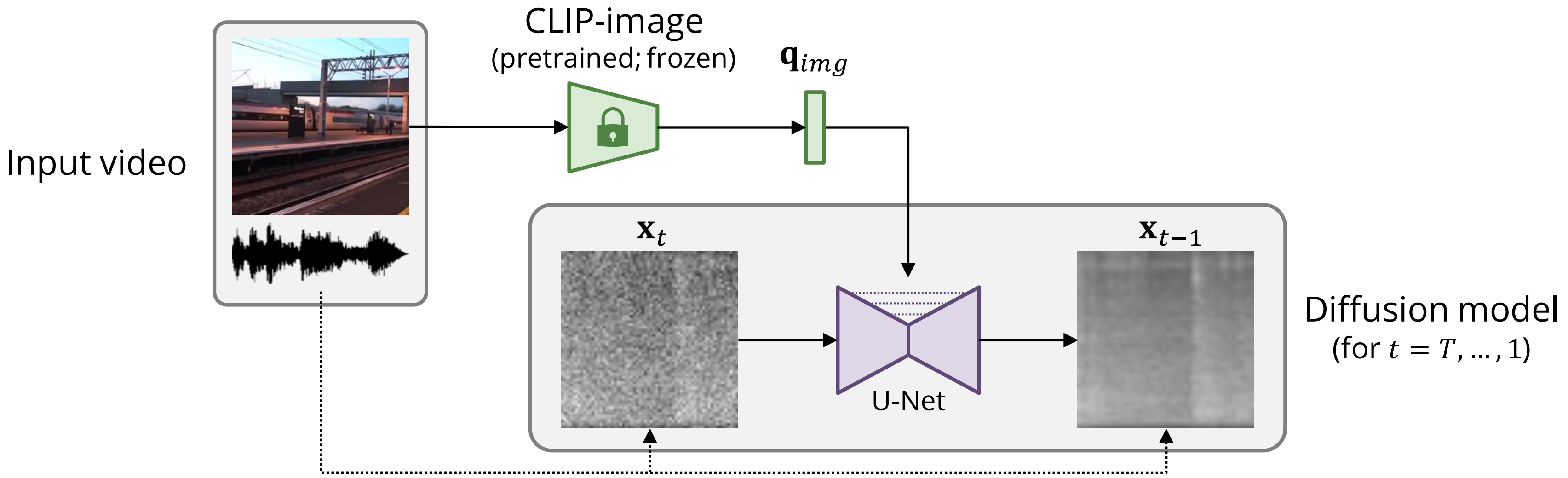
Output



Ho et al., "Denoising Diffusion Probabilistic Models," *NeurIPS*, 2020.

Training – Image-queried

- We train an image-to-audio synthesis model using a diffusion model on mel spectrograms and a pretrained CLIP-image encoder



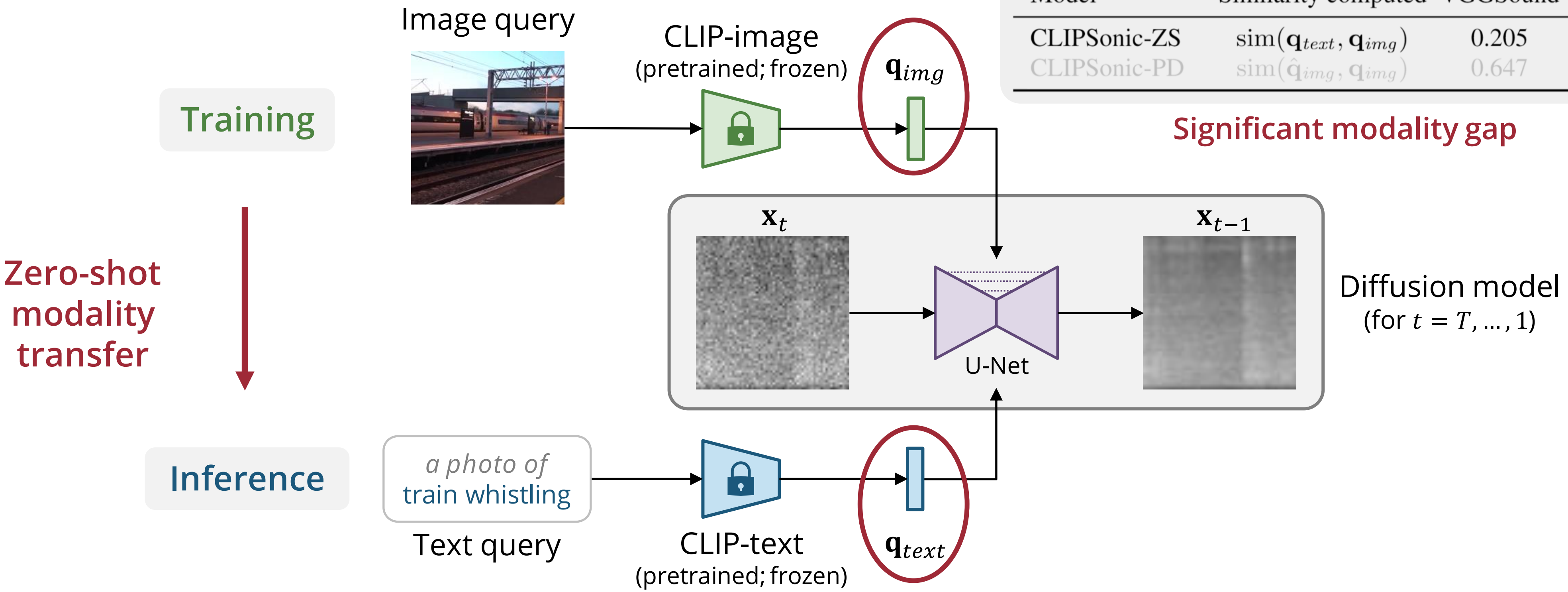
Inference – Zero-shot Modality Transfer (CLIPSonic-ZS)

- We first explore using a pretrained CLIP-text encoder directly

Table 2: Cosine similarities between various query embeddings.

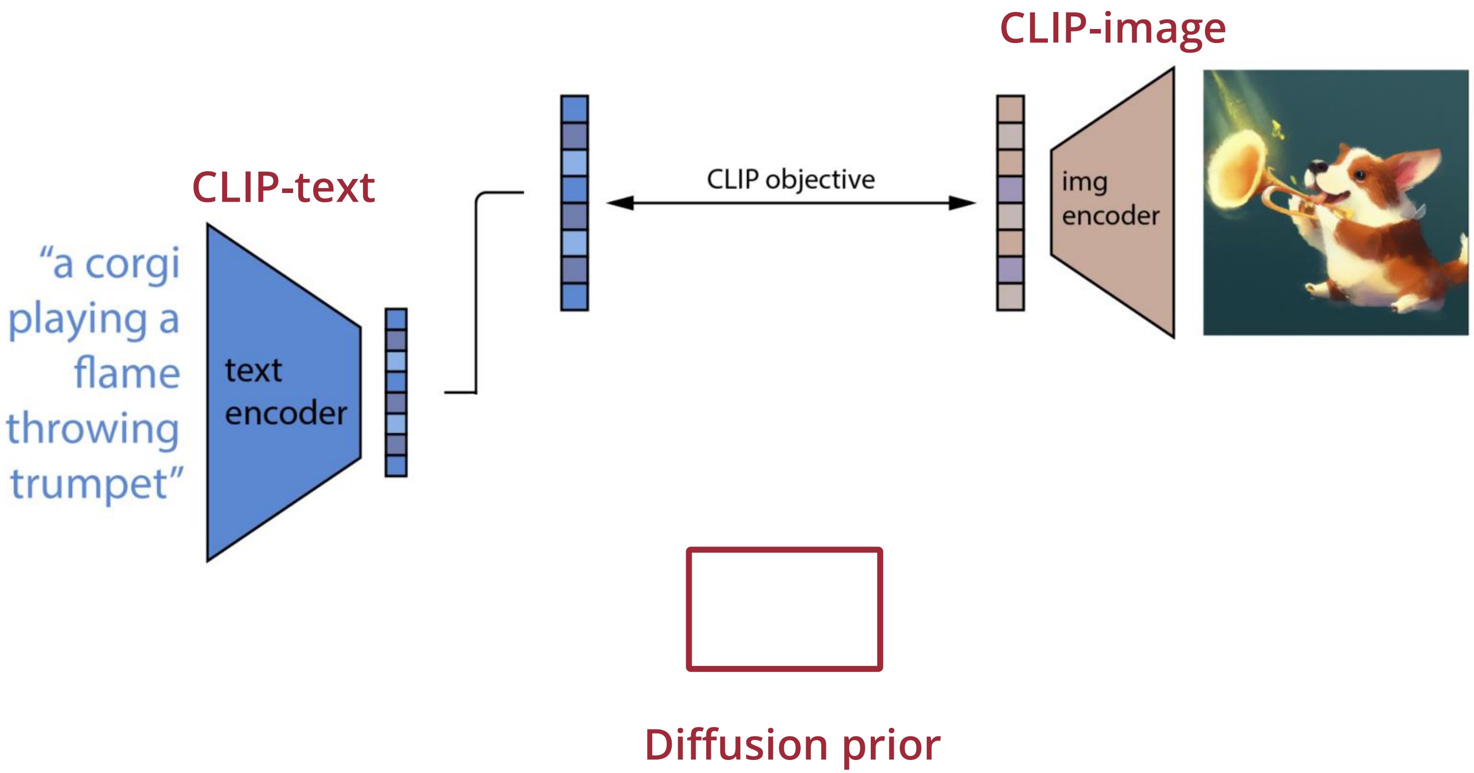
Model	Similarity computed	VGGSound	MUSIC
CLIPSonic-ZS	$\text{sim}(\mathbf{q}_{text}, \mathbf{q}_{img})$	0.205	0.245
CLIPSonic-PD	$\text{sim}(\hat{\mathbf{q}}_{img}, \mathbf{q}_{img})$	0.647	0.720

Significant modality gap


Training
Zero-shot modality transfer
Inference
Image query
CLIP-image (pretrained; frozen)
$\mathbf{q}_{img}$
$\mathbf{x}_t$
U-Net
$\mathbf{x}_{t-1}$
Diffusion model (for $t = T, \ldots, 1$)
a photo of train whistling
Text query
CLIP-text (pretrained; frozen)
$\mathbf{q}_{text}$

How to overcome this modality gap?

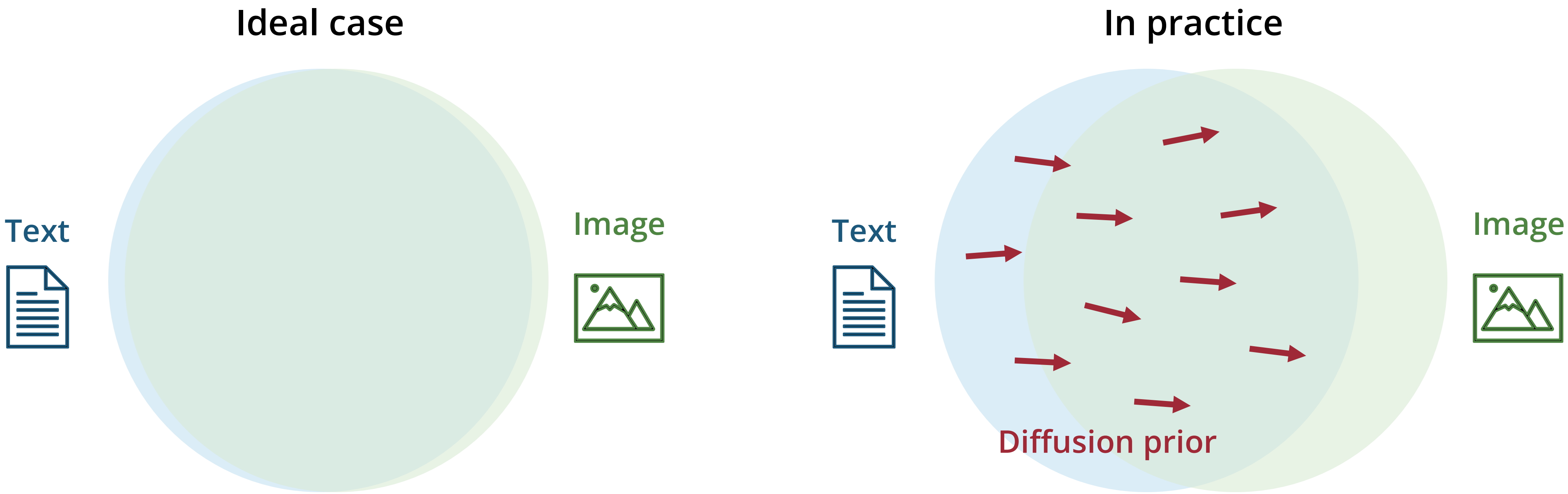
- We leverage a pretrained diffusion prior model (Ramesh et al., 2022)



Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint arXiv:2204.06125*, 2022.

Diffusion Prior (Ramesh et al., 2022)

CLIP embedding spaces



Ideal case

Text

Image

In practice

Text

Image

Diffusion prior

Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint arXiv:2204.06125*, 2022.

Inference – Pretrained Diffusion Prior (CLIPSonic-PD)

- We then explore using a pretrained diffusion prior model (Ramesh et al., 2022)

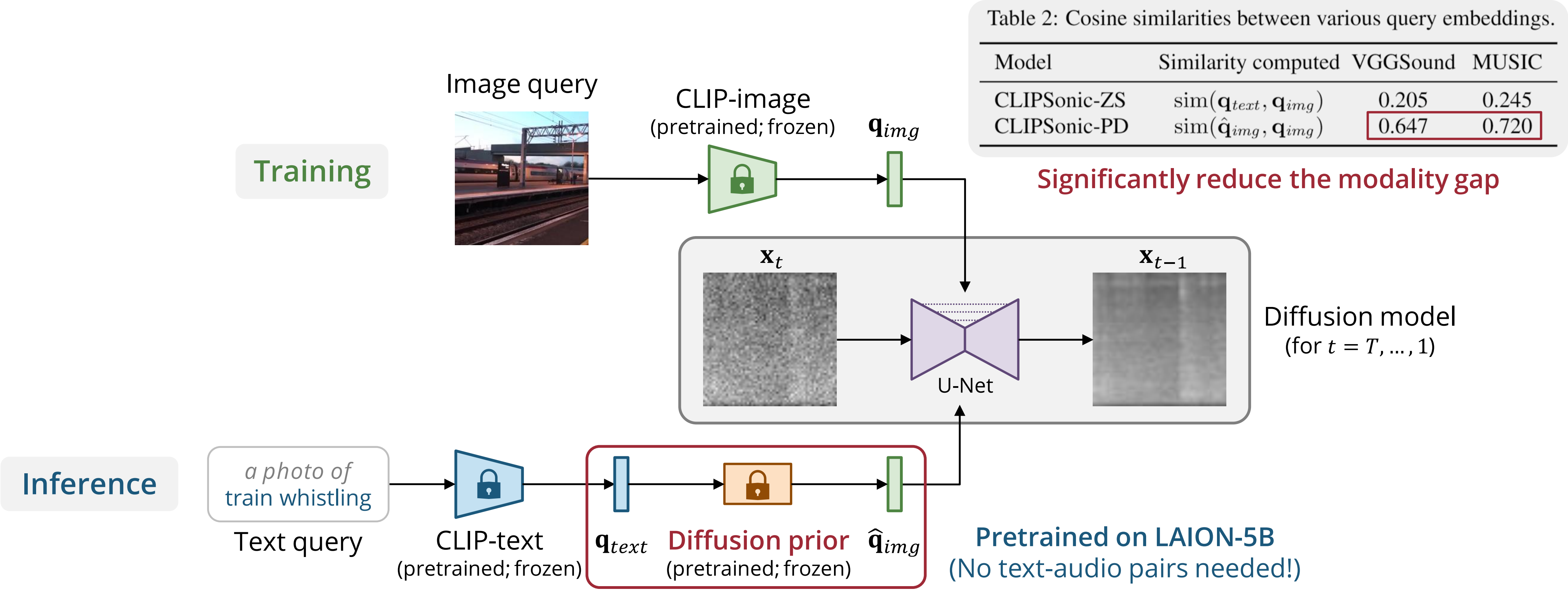


Table 2: Cosine similarities between various query embeddings.

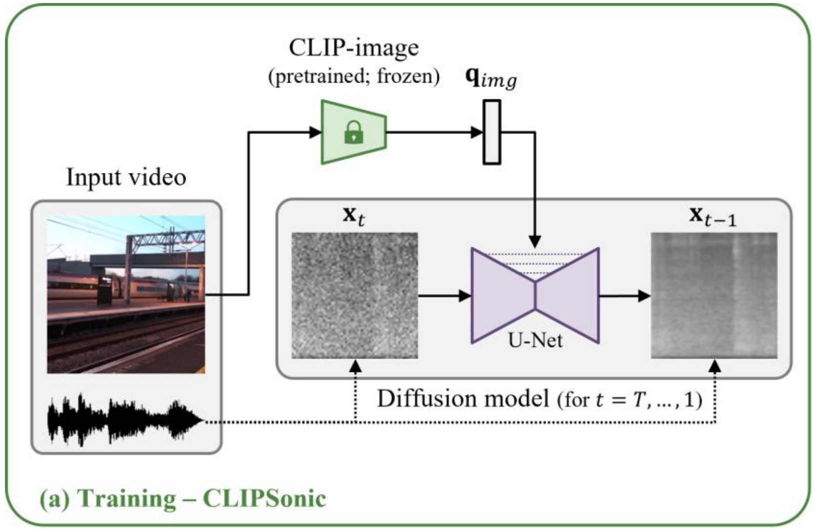
Model	Similarity computed	VGGSound	MUSIC
CLIPSonic-ZS	sim($\mathbf{q}_{text}, \mathbf{q}_{img}$)	0.205	0.245
CLIPSonic-PD	sim($\hat{\mathbf{q}}_{img}, \mathbf{q}_{img}$)	0.647	0.720

Significantly reduce the modality gap

Pretrained on LAION-5B
(No text-audio pairs needed!)

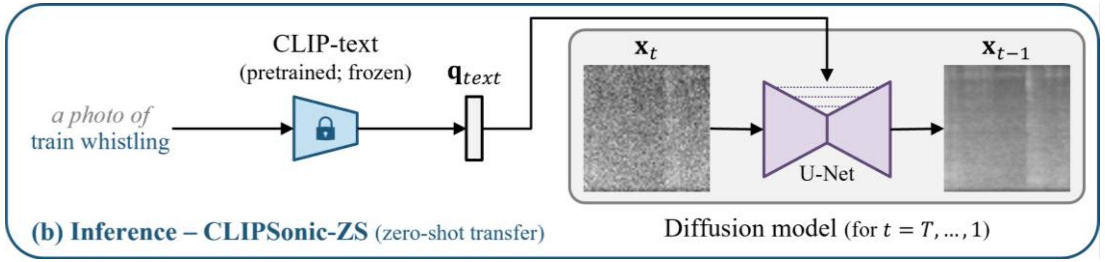
Recap

Training

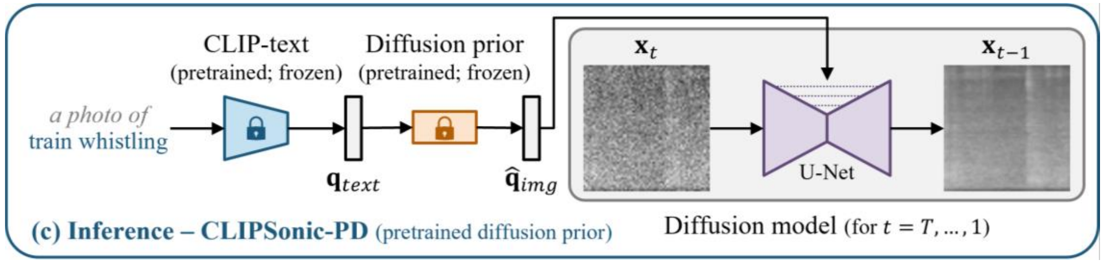


CLIPSonic-IQ
(image-queried)

Inference



CLIPSonic-ZS
(zero-shot transfer)



CLIPSonic-PD
(pretrained diffusion prior)

Data

MUSIC

(Zhao et al., 2018)

Violin

Acoustic guitar

Accordion

Music instrument playing videos

(1,055 videos, 21 instruments)

VGGSound

(Chen et al., 2020)

Hedge trimmer running

Dog bow-wow

Bird chirping, tweeting

Noisy videos with diverse sounds

(172K videos, 310 classes)

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

Examples of VGGSound

pheasant crowing

railroad car, train wagon

Chen et al., "VGGSound: A Large-Scale Audio-Visual Dataset," *ICASSP*, 2020.

Implementation Details

Mel spectrogram configuration

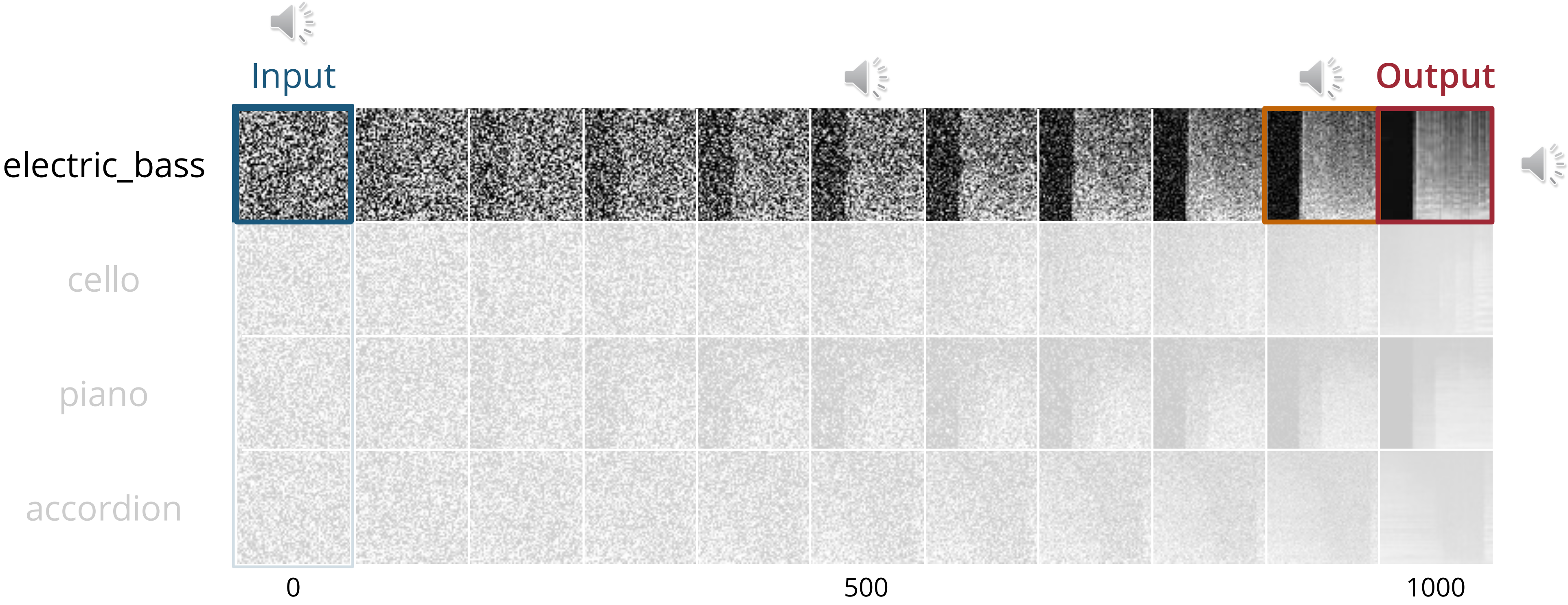
- Sampling rate: 16 kHz
- Hop size: 512
- FFT filter size: 2048
- 64 mel bands
- Inverted back to waveforms using BigVGAN (Lee et al., 2023)

Diffusion model

- Based on Improved DDPM (Nichol and Dhariwal, 2019)
- Diffusion steps:
  - Training: 4000
  - Inference: 1000
- Training iterations
  - MUSIC: 200K (1 day on 2 RTX 2080 Tis)
  - VGGSound: 500K (2 days on 2 RTX 2080 Tis)

Nichol and Dhariwal, "Improved Denoising Diffusion Probabilistic Models," *ICML*, 2019.
Lee et al., "BigVGAN: A Universal Neural Vocoder with Large-Scale Training," *ICLR*, 2023.

Inference – Examples



Text-to-Audio Synthesis – Demo

Rapping

Sea waves

Thunder

Smoke detector beeping

Playing table tennis

Playing violin fiddle

Text-to-Audio Synthesis – Demo

	Rapping	Sea waves	Playing table tennis
CLIPSonic-ZS (zero-shot modality transfer)	🔊	🔊	🔊
CLIPSonic-PD (pretrained diffusion prior)	🔊	🔊	🔊

The pretrained diffusion prior model improves the text-audio relevance.

Text-to-Audio Synthesis – Listening Test

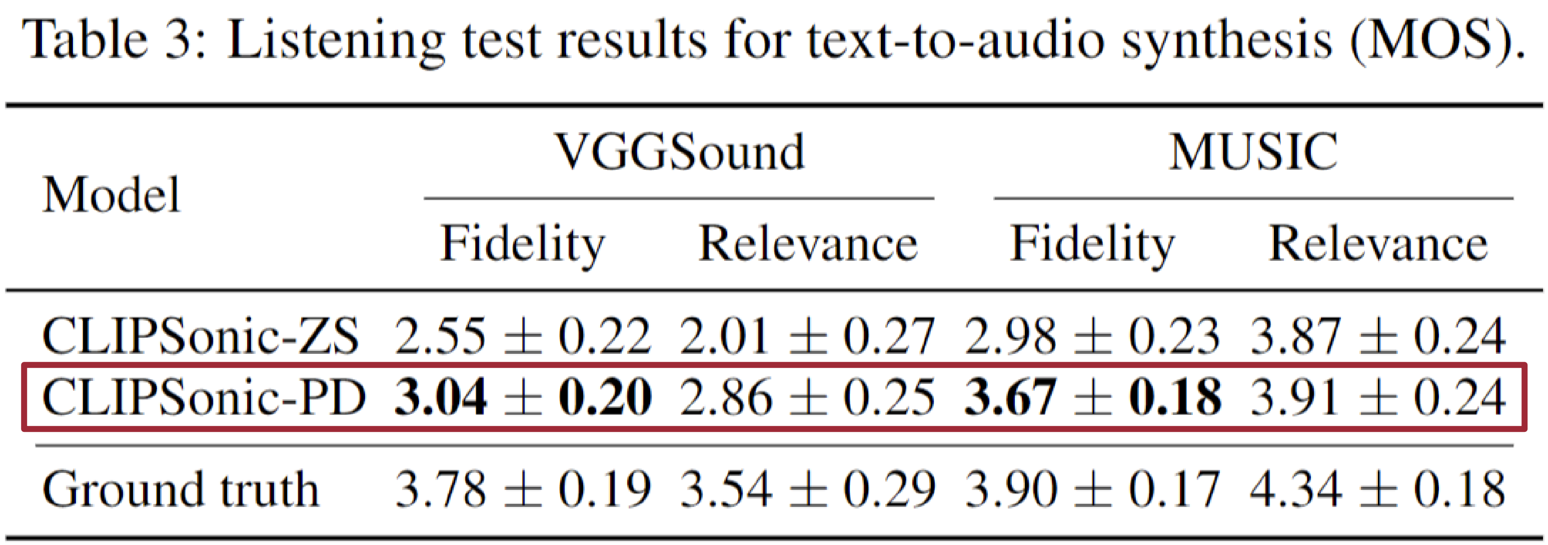


Table 3: Listening test results for text-to-audio synthesis (MOS).

Model	VGGSound		MUSIC	
	Fidelity	Relevance	Fidelity	Relevance
CLIPSonic-ZS	2.55 ± 0.22	2.01 ± 0.27	2.98 ± 0.23	3.87 ± 0.24
CLIPSonic-PD	**3.04** ± **0.20**	2.86 ± 0.25	**3.67** ± **0.18**	3.91 ± 0.24
Ground truth	3.78 ± 0.19	3.54 ± 0.29	3.90 ± 0.17	4.34 ± 0.18

Significant performance improvement against the baseline!

Image-to-Audio Synthesis – Demo (Out-of-distribution)

Image-to-Audio Synthesis – Demo (Out-of-distribution)

CLIPSonic-IQ (ours)	🔊	🔊	🔊	🔊
Im2wav (Sheffer & Adi, 2023)	🔊	🔊	🔊	🔊
SpecVQGAN (Iashin & Rahtu, 2021)	🔊	🔊	🔊	🔊

Our proposed method generates clearer audio than two existing models!

Sheffer and Adi, "I Hear Your True Colors: Image Guided Audio Generation," *ICASSP*, 2023.
Iashin and Rahtu, "Taming Visually Guided Sound Generation," *BMVC*, 2021.

Image-to-Audio Synthesis – Listening Test

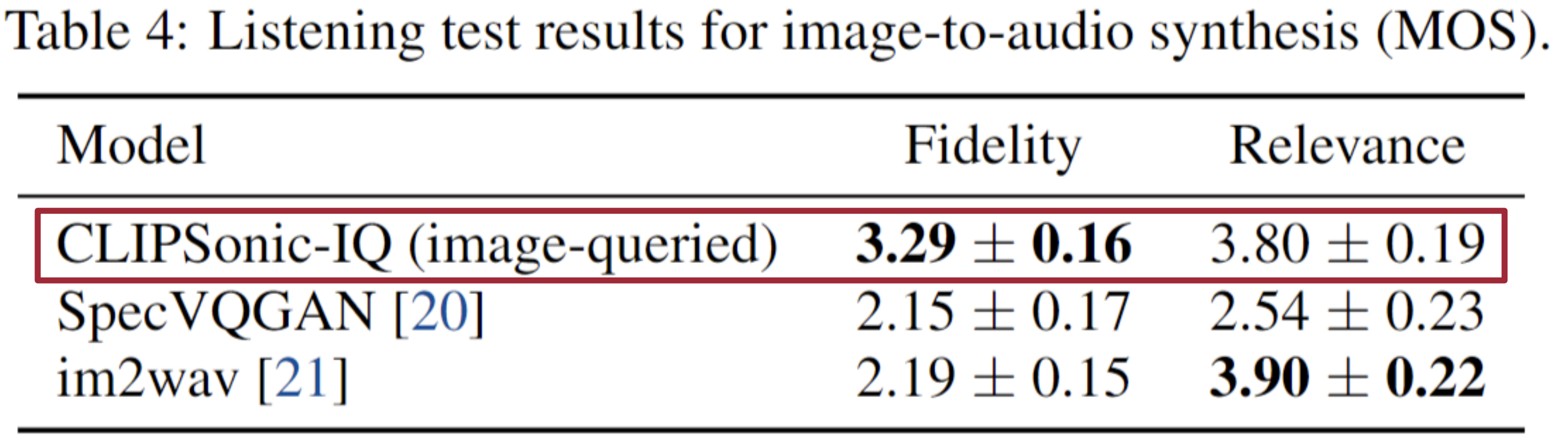


Table 4: Listening test results for image-to-audio synthesis (MOS).

Model	Fidelity	Relevance
CLIPSonic-IQ (image-queried)	**3.29 ± 0.16**	3.80 ± 0.19
SpecVQGAN [20]	2.15 ± 0.17	2.54 ± 0.23
im2wav [21]	2.19 ± 0.15	**3.90 ± 0.22**

State-of-the-art image-to-audio performance!

Sheffer and Adi, “I Hear Your True Colors: Image Guided Audio Generation,” *ICASSP*, 2023.
Iashin and Rahtu, “Taming Visually Guided Sound Generation,” *BMVC*, 2021.

Objective Evaluation Metrics

- Evaluated with Fréchet audio distance (FAD) and CLAP score

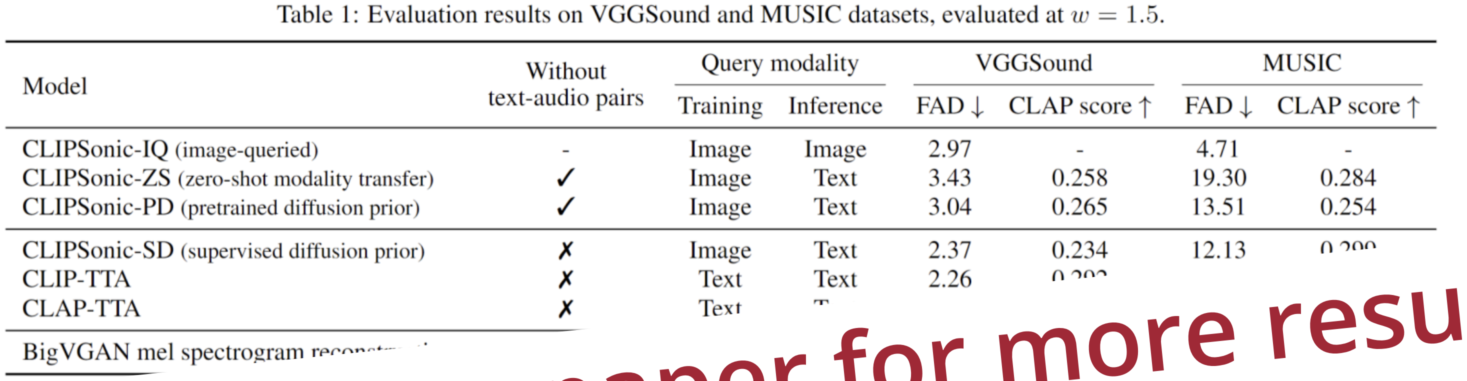
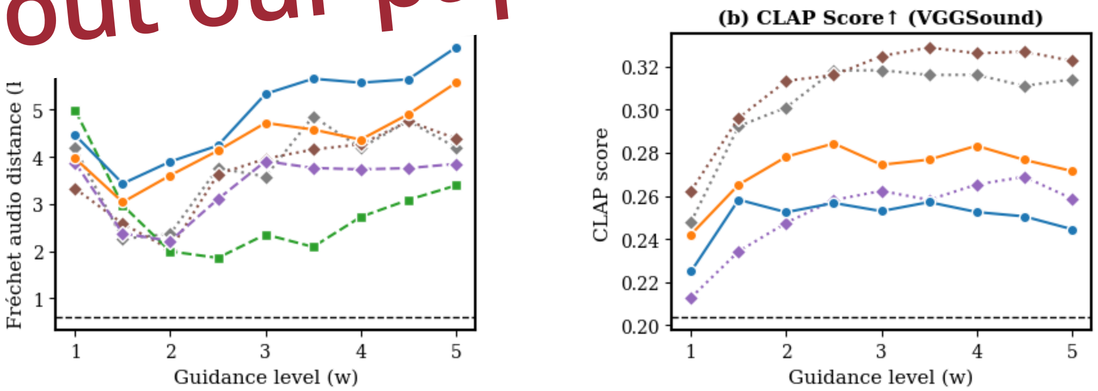


Table 1: Evaluation results on VGGSound and MUSIC datasets, evaluated at $w = 1.5$.

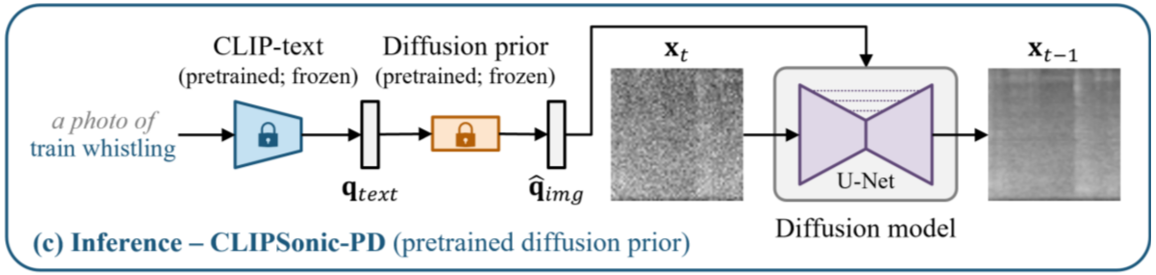
Model	Without text-audio pairs	Query modality		VGGSound		MUSIC	
		Training	Inference	FAD ↓	CLAP score ↑	FAD ↓	CLAP score ↑
CLIPSonic-IQ (image-queried)	-	Image	Image	2.97	-	4.71	-
CLIPSonic-ZS (zero-shot modality transfer)	✓	Image	Text	3.43	0.258	19.30	0.284
CLIPSonic-PD (pretrained diffusion prior)	✓	Image	Text	3.04	0.265	13.51	0.254
CLIPSonic-SD (supervised diffusion prior)	✗	Image	Text	2.37	0.234	12.13	[illegible]
CLIP-TTA	✗	Text	Text	2.26	[illegible]		
CLAP-TTA	✗	Text	[illegible]				
BigVGAN mel spectrogram [illegible]							

Check out our paper for more results!



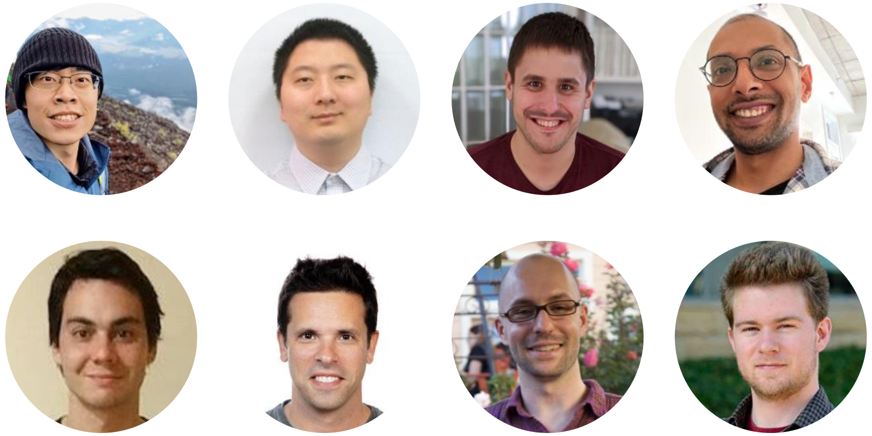
Summary

- Proposed a text-to-audio synthesis model that requires *no* text-audio pairs
- Achieves strong performance in objective and subjective evaluations
- Achieves state-of-the-art performance in image-to-audio synthesis



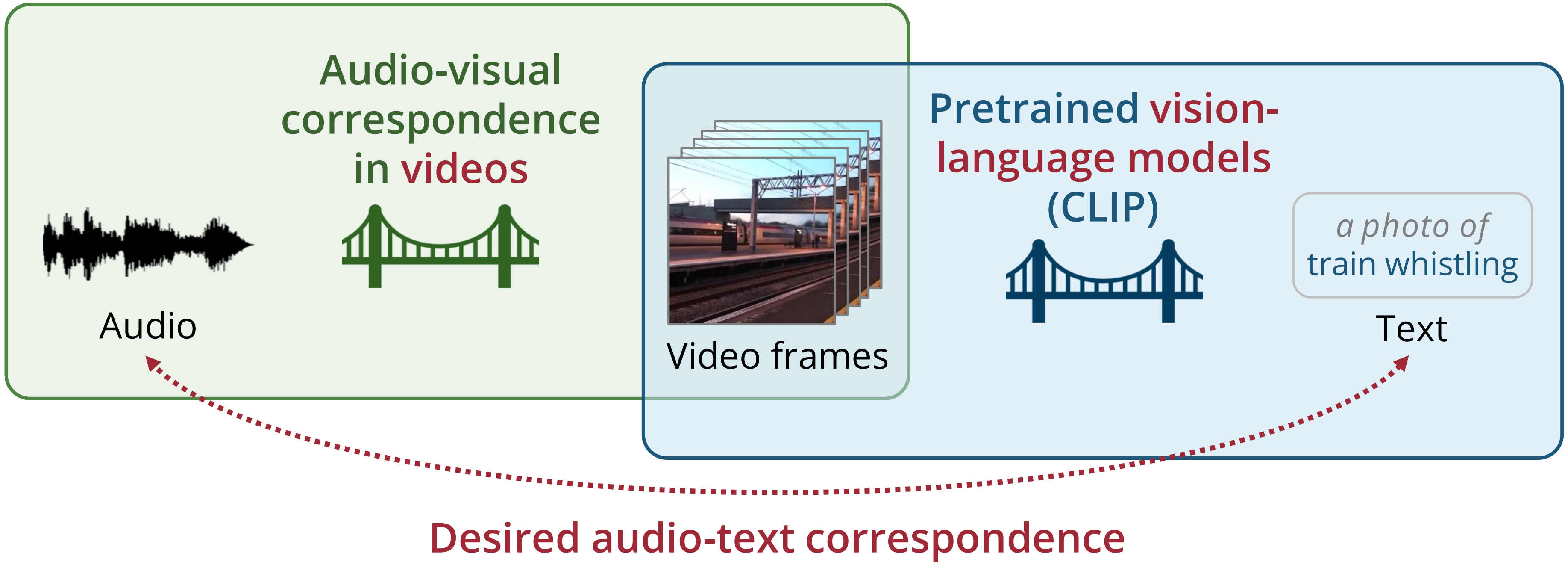
(c) Inference – CLIPSonic-PD (pretrained diffusion prior)

Paper: arxiv.org/abs/2306.09635
Demo: salu133445.github.io/clipsonic



Conclusion

Leveraging the Visual Domain as a Bridge


Audio-visual correspondence in videos
Audio
Video frames
Pretrained vision-language models (CLIP)
a photo of train whistling
Text
Desired audio-text correspondence

No text-audio pairs required!

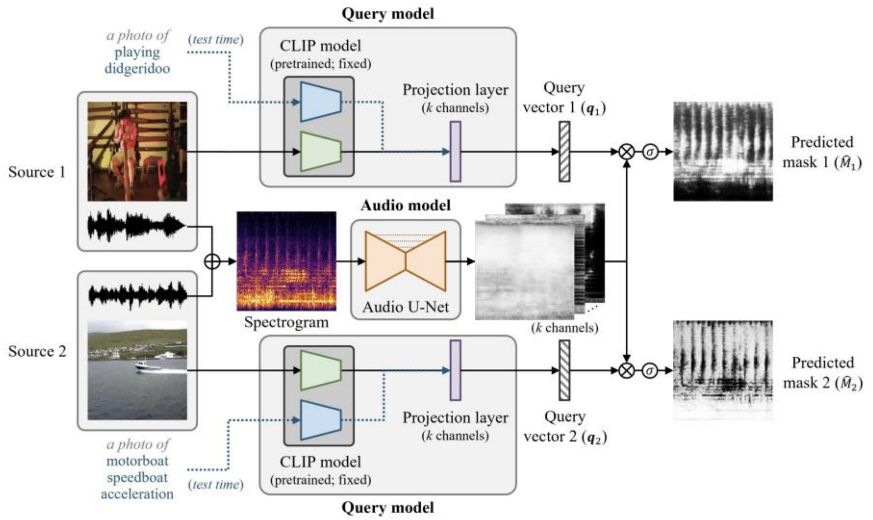
Scalable to large video datasets!

Overview

CLIPSep

(Dong et al., ICLR 2023)

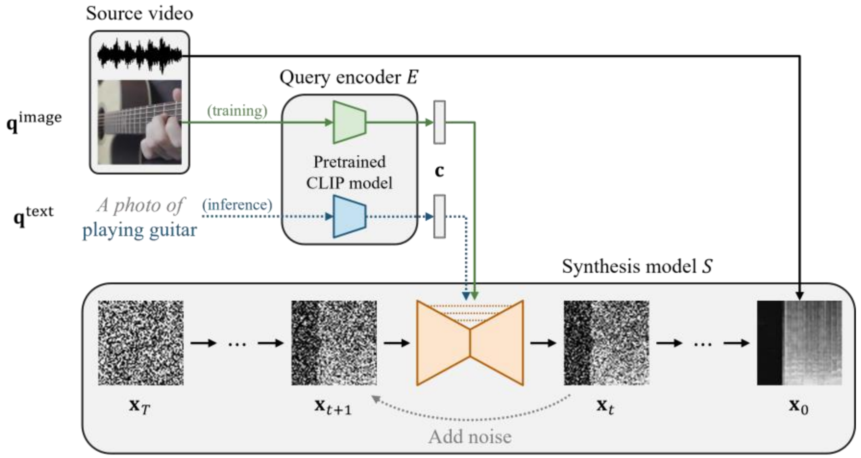
For text-queried sound separation



CLIPSonic

(Dong et al., WASPAA 2023)

For text-to-audio synthesis



Dong et al., "CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos," *ICLR*, 2023.
Dong et al., "CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models," *WASPAA*, 2023.

Limitations & Future Work

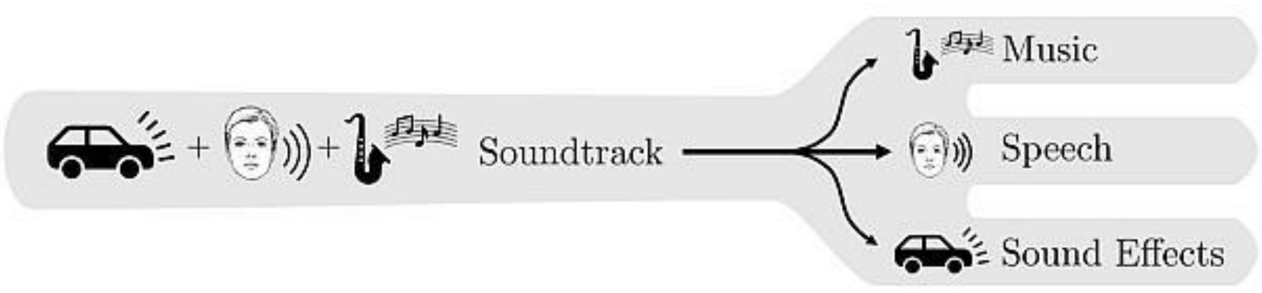
- Off-screen sounds occur frequently in videos
- Cannot handle purely audio-specific queries
- Can we enable compositional prompts?
- Scale up to larger video datasets!

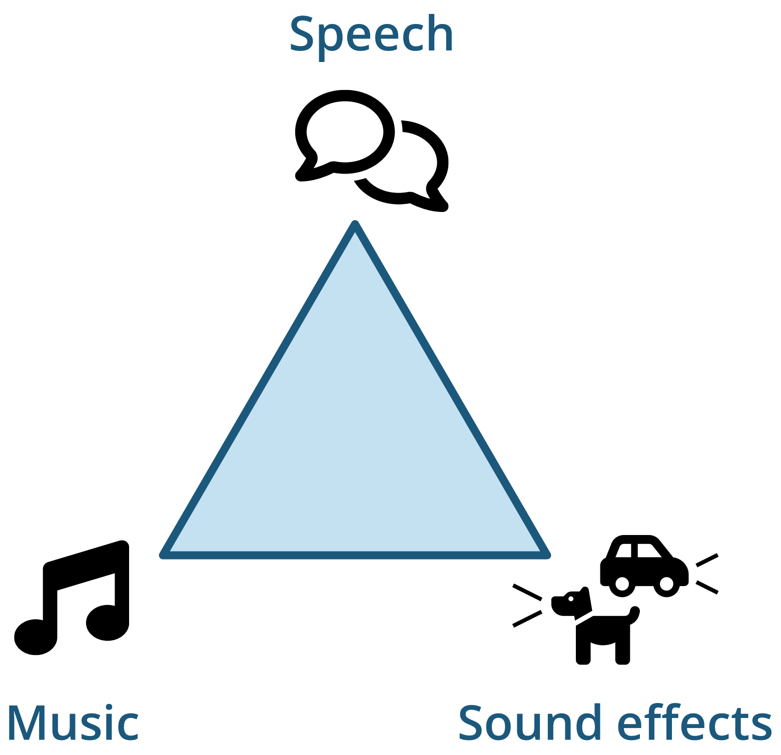
Future Directions

Future Directions

- Audio-visual sound separation
- Multimodal generative AI

Cocktail Fork Problem




Speech

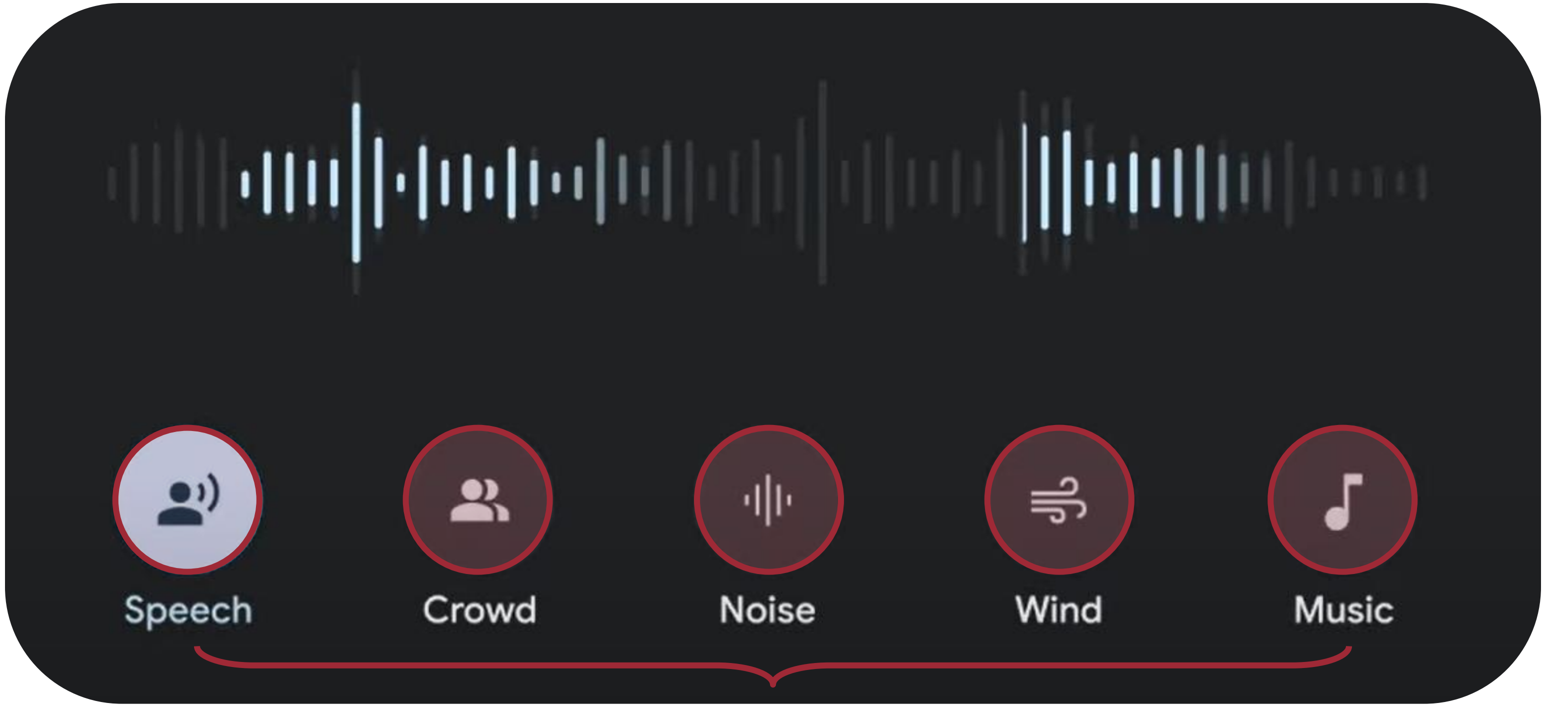
Music

Sound effects

Petermann et al., "The Cocktail Fork Problem: Three-Stem Audio Separation for Real-World Soundtracks," *ICASSP*, 2022.

Sound Separation in Practice

Google's Audio Magic Eraser



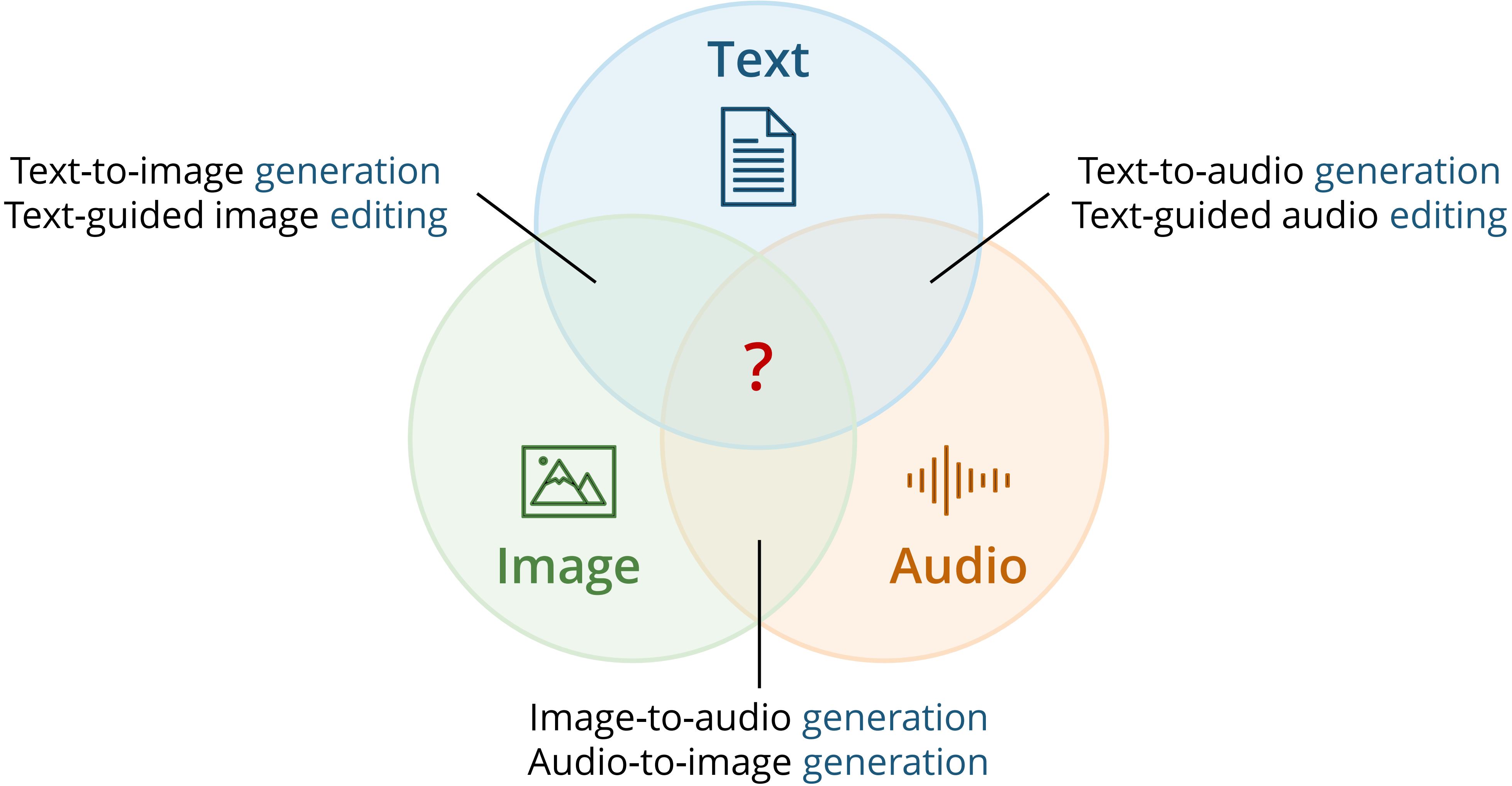
Predefined categories

Dina Berrada, "4 new Google Photos features on Pixel 8 and Pixel 8 Pro," *The Keyword*, https://blog.google/products/photos/google-photos-features-pixel-8-pro/, 2023.

Audio-visual Sound Separation for Audio Remixing

Kirillov et al., "Segment Anything," *ICCV*, 2023.

Multimodal Generative AI



Text-to-image generation
Text-guided image editing

Text-to-audio generation
Text-guided audio editing

Image-to-audio generation
Audio-to-image generation

Mumbai, the city of dreams.

Multimodal Generative AI for Films

Visuals	**Midjourney**
Video	**Runway**
Narration (script)	**ChatGPT**
Narration (voice)	**ElevenLabs**
Sound effects	**Audiocraft**

Sagar Amruth, *Twitter*, https://twitter.com/sagartheamruth/status/1688530899271364608, August 7, 2023.

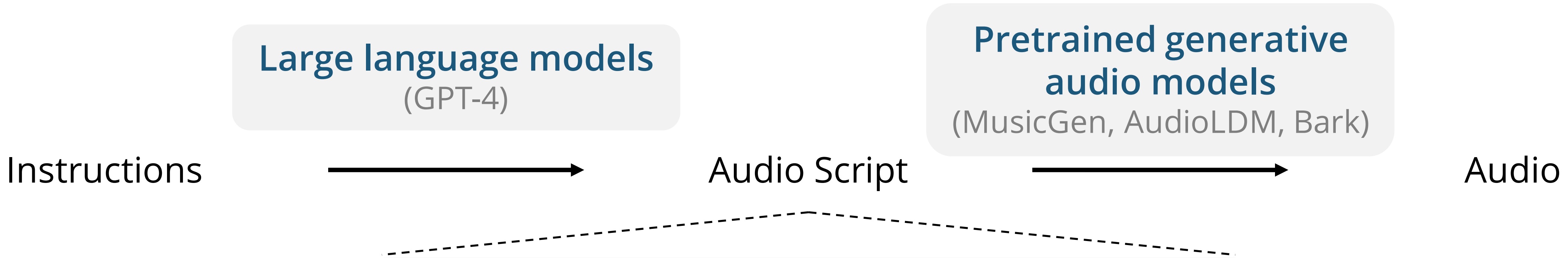
Multimodal Generative AI for News

Generate an audio in Science Fiction theme: Mars News reporting that Humans send light-speed probe to Alpha Centauri. *Start with news anchor, followed by a reporter interviewing a chief engineer from an organization that built this probe, founded by United Earth and Mars Government, and end with the news anchor again.*

Script	GPT-4
Music	MusicGen
Narration	Bark
Sound effects	AudioLDM

Liu et al., "WavJourney: Compositional Audio Creation with Large Language Models," *arXiv preprint arXiv:2307.14335*, 2023.

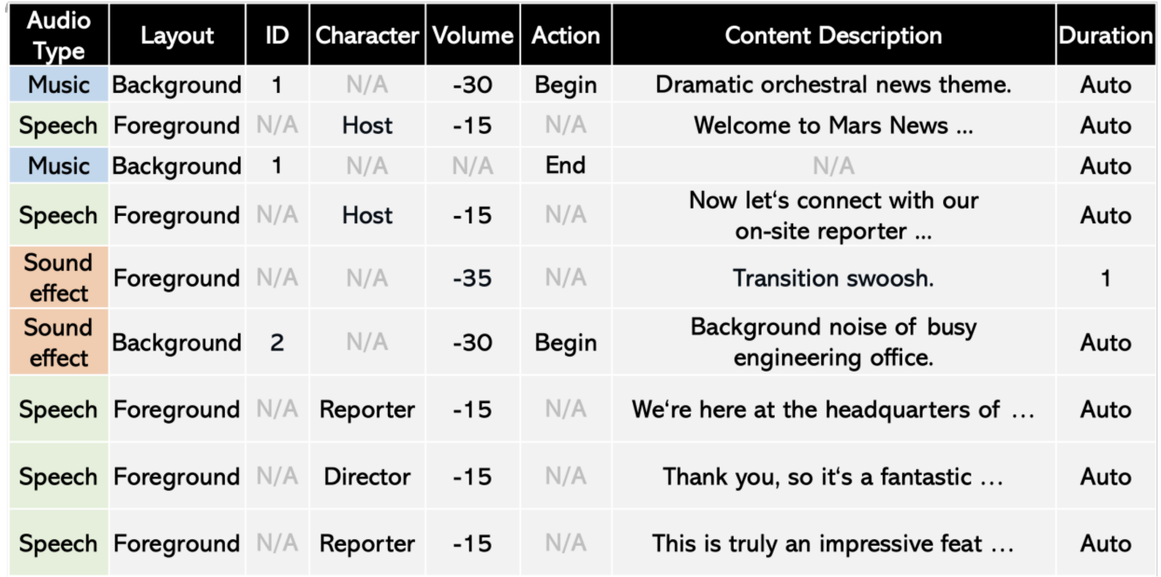
Controllable Multimodal Generative AI



Large language models
(GPT-4)

Pretrained generative audio models
(MusicGen, AudioLDM, Bark)

Instructions → Audio Script → Audio

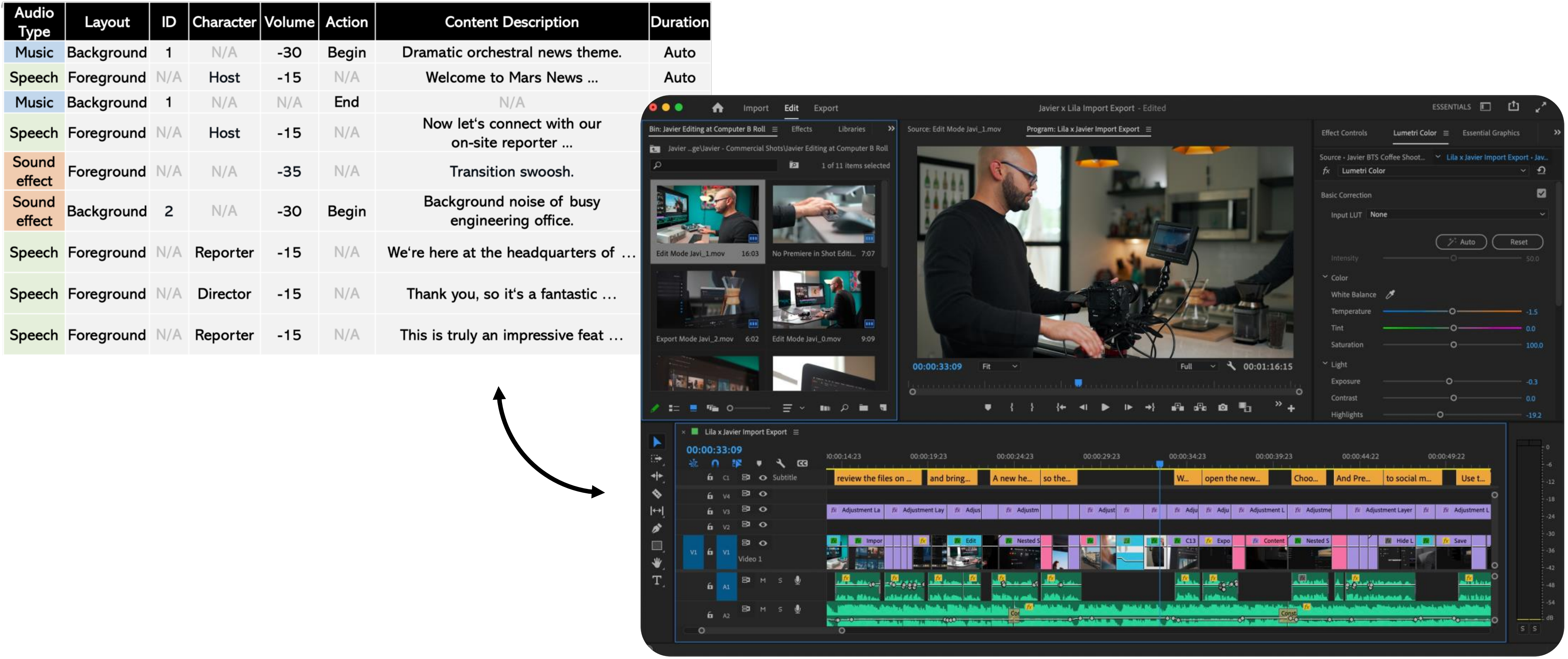


Audio Type	Layout	ID	Character	Volume	Action	Content Description	Duration
Music	Background	1	N/A	-30	Begin	Dramatic orchestral news theme.	Auto
Speech	Foreground	N/A	Host	-15	N/A	Welcome to Mars News ...	Auto
Music	Background	1	N/A	N/A	End	N/A	Auto
Speech	Foreground	N/A	Host	-15	N/A	Now let's connect with our on-site reporter ...	Auto
Sound effect	Foreground	N/A	N/A	-35	N/A	Transition swoosh.	1
Sound effect	Background	2	N/A	-30	Begin	Background noise of busy engineering office.	Auto
Speech	Foreground	N/A	Reporter	-15	N/A	We're here at the headquarters of …	Auto
Speech	Foreground	N/A	Director	-15	N/A	Thank you, so it's a fantastic …	Auto
Speech	Foreground	N/A	Reporter	-15	N/A	This is truly an impressive feat …	Auto

Interactable intermediate outputs

Liu et al., "WavJourney: Compositional Audio Creation with Large Language Models," *arXiv preprint arXiv:2307.14335*, 2023.

Controllable Multimodal Generative AI



Audio Type	Layout	ID	Character	Volume	Action	Content Description	Duration
Music	Background	1	N/A	-30	Begin	Dramatic orchestral news theme.	Auto
Speech	Foreground	N/A	Host	-15	N/A	Welcome to Mars News …	Auto
Music	Background	1	N/A	N/A	End	N/A	
Speech	Foreground	N/A	Host	-15	N/A	Now let's connect with our on-site reporter …	
Sound effect	Foreground	N/A	N/A	-35	N/A	Transition swoosh.	
Sound effect	Background	2	N/A	-30	Begin	Background noise of busy engineering office.	
Speech	Foreground	N/A	Reporter	-15	N/A	We're here at the headquarters of …	
Speech	Foreground	N/A	Director	-15	N/A	Thank you, so it's a fantastic …	
Speech	Foreground	N/A	Reporter	-15	N/A	This is truly an impressive feat …	

Integration into professional creative workflow

Acknowledgements

UC San Diego

中央研究院 ACADEMIA SINICA

Dolby

SONY

amazon

MINISTRY OF EDUCATION 教育部

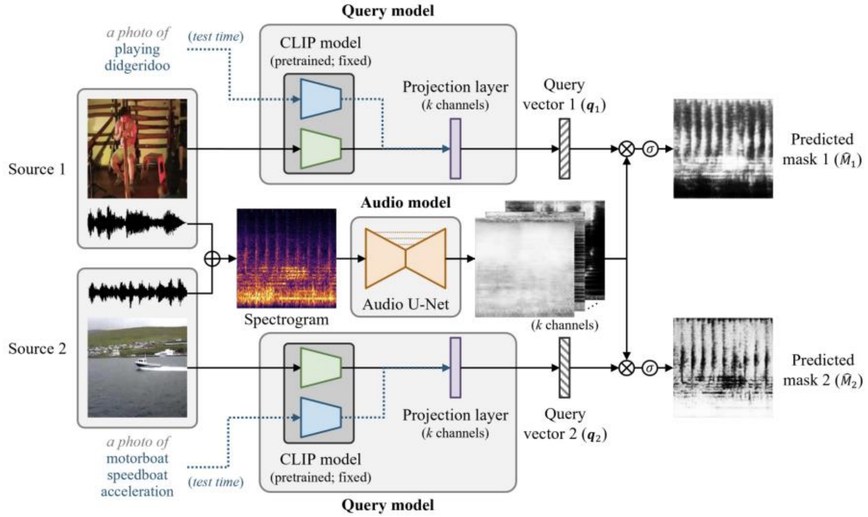
erc

Thank you!

CLIPSep

(Dong et al., ICLR 2023)

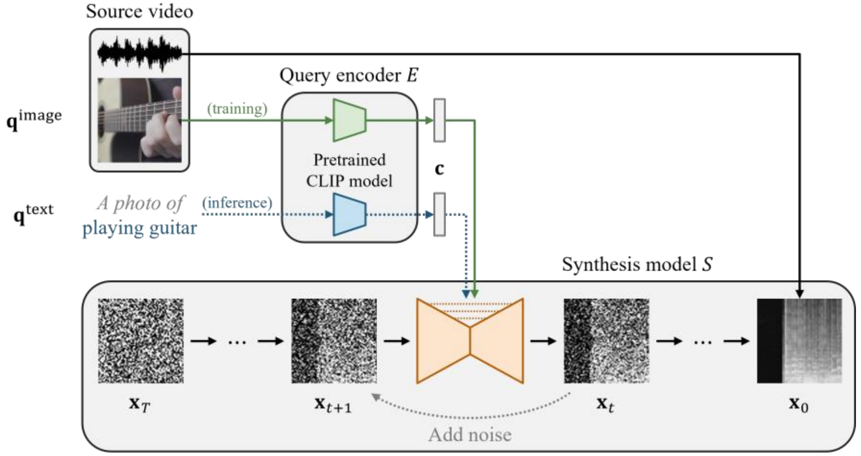
For text-queried sound separation



CLIPSonic

(Dong et al., WASPAA 2023)

For text-to-audio synthesis



UC San Diego