Generative AI for Music and Audio

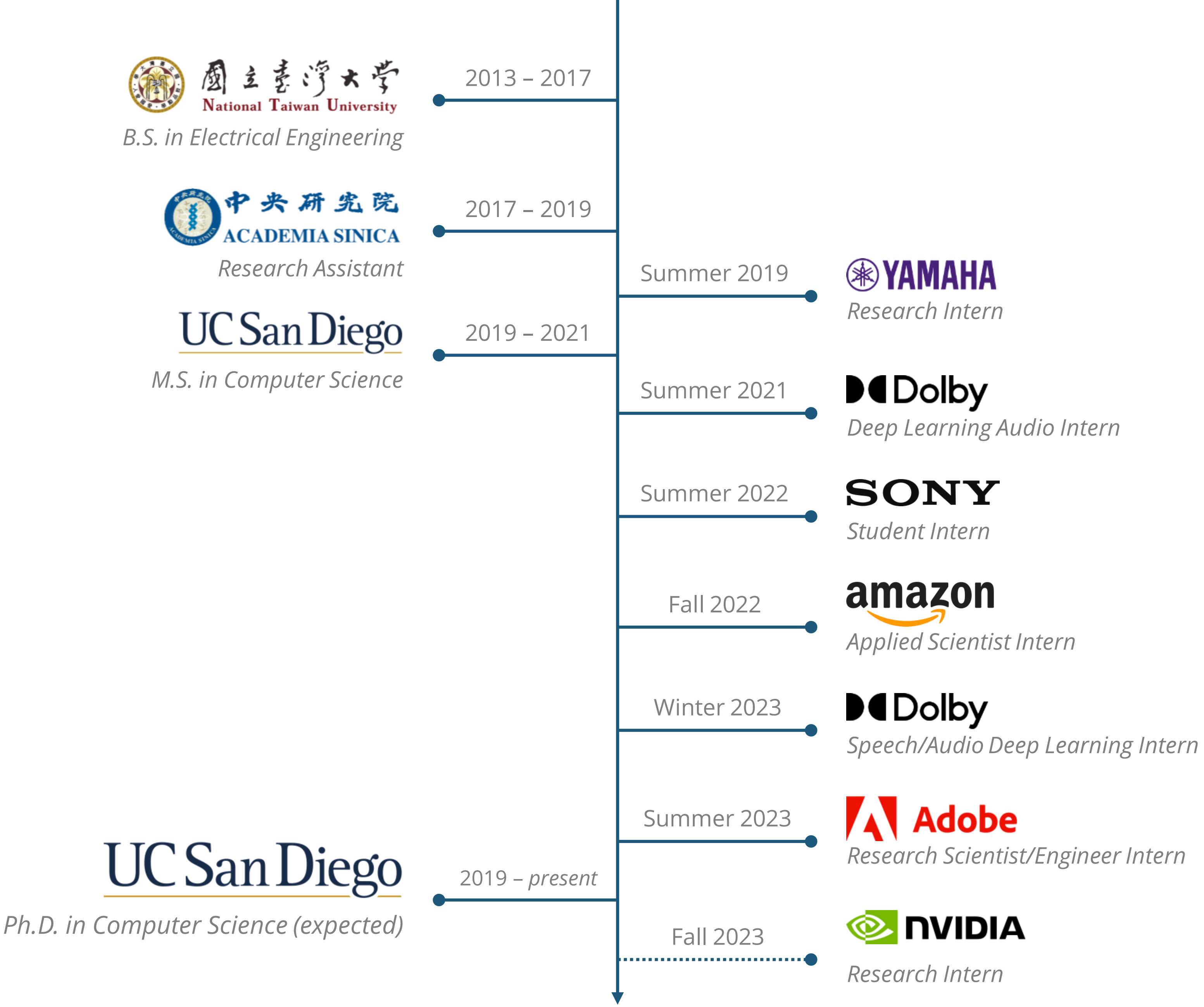
Hao-Wen (Herman) Dong

董皓文

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
B.S. in Electrical Engineering
2013 – 2017

中央研究院 ACADEMIA SINICA
Research Assistant
2017 – 2019

Summer 2019
YAMAHA
Research Intern

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

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

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

Fall 2023
NVIDIA
Research Intern

About me

EE

a female cat engineer making
an electric chip in a classroom

Music

a cat playing heavy metal

CS

a cat engineer debugging on laptop

Made with Adobe Firefly (firefly.adobe.com) and Facebook MusicGen (huggingface.co/spaces/facebook/MusicGen)

Introduction

Mumbai, the city of dreams.

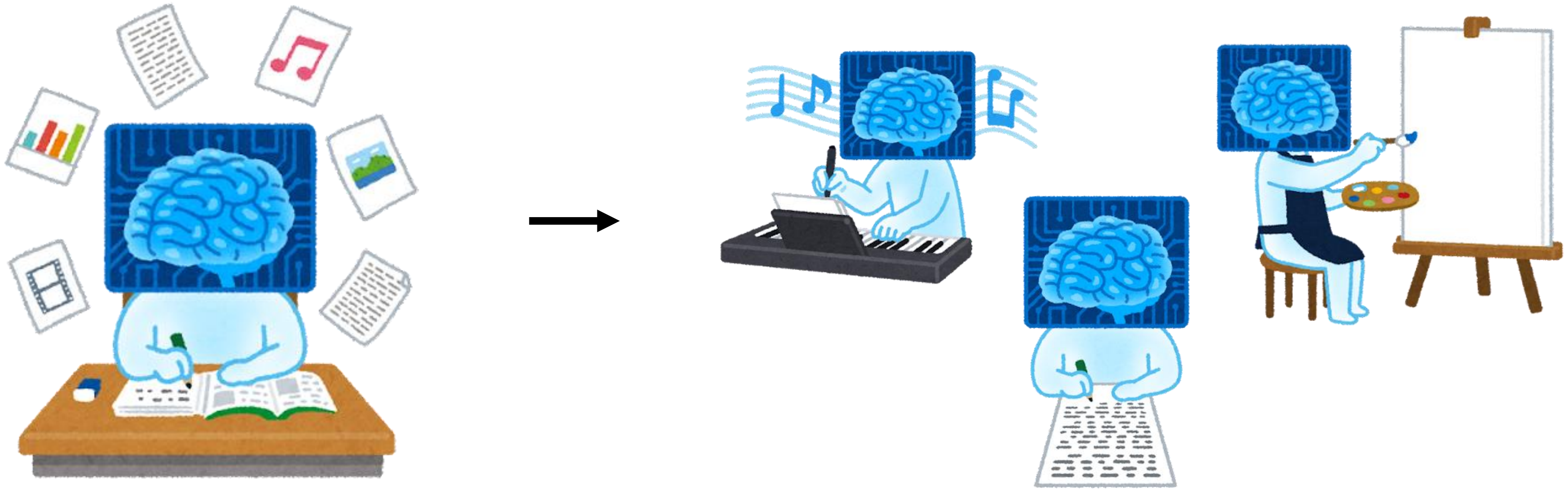
Multimodal generative AI for Films

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.

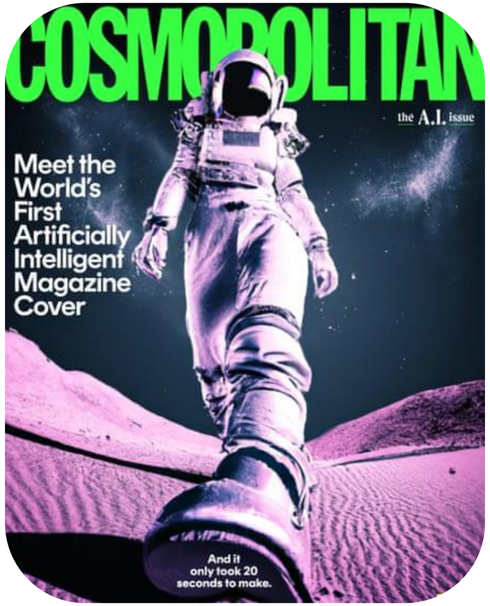
What is Generative AI?

- Generative AI is **AI capable of generating text, images, or other media**.



Generative AI for Visual Arts

AI made a magazine cover


COSMOPOLITAN
the A.I. issue
Meet the World's First Artificially Intelligent Magazine Cover
And it only took 20 seconds to make.

(Source: Cosmopolitan)

AI won an art contest

(Source: CNN Business)

AI won a photography contest

(Source: CNN)

Gloria Liu, "The World's Smartest Artificial Intelligence Just Made Its First Magazine Cover," *Cosmopolitan*, June 21, 2022.
Rachel Metz, "AI won an art contest, and artists are furious," *CNN Business*, September 3, 2022.
Lianne Kolirin, "Artist rejects photo prize after AI-generated image wins award," *CNN*, April 18, 2023.

Types of Audio

Speech

(Source: Wikimedia Commons)

Music

(Source: Wikimedia Commons)

Sound effects

(Source: Wikimedia Commons)

BPJ Media Inc, CC BY-SA 3.0, via Wikimedia Commons.
Vancouver Film SchoolRetouched version by User:Quenhitran., CC BY 2.0, via Wikimedia Commons.
The Blackbird Academy, CC BY-SA 2.0, via Wikimedia Commons.
One Man Films, "One Shot - WAR ACTION SHORT FILM," *YouTube*, September 11, 2022.

Generative AI for Music

Prompt: relaxing and smooth jazz played in a stylish cafe

Prompt: delightful country music with acoustic guitars

Prompt: cinematic and suspenseful orchestral music

Made with Facebook MusicGen (huggingface.co/spaces/facebook/MusicGen)

Generative AI for Sound Effects

Text-to-audio Synthesis

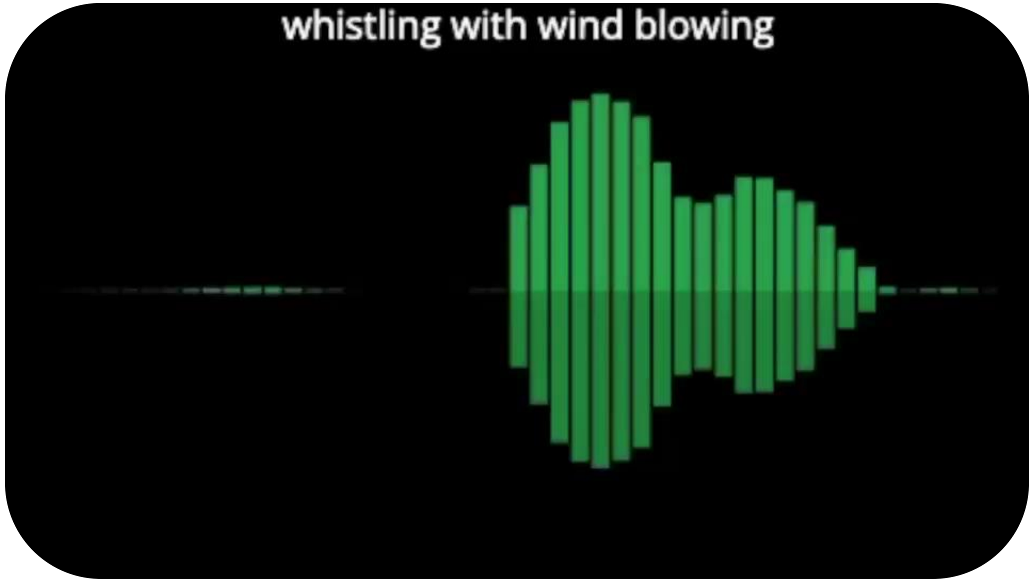
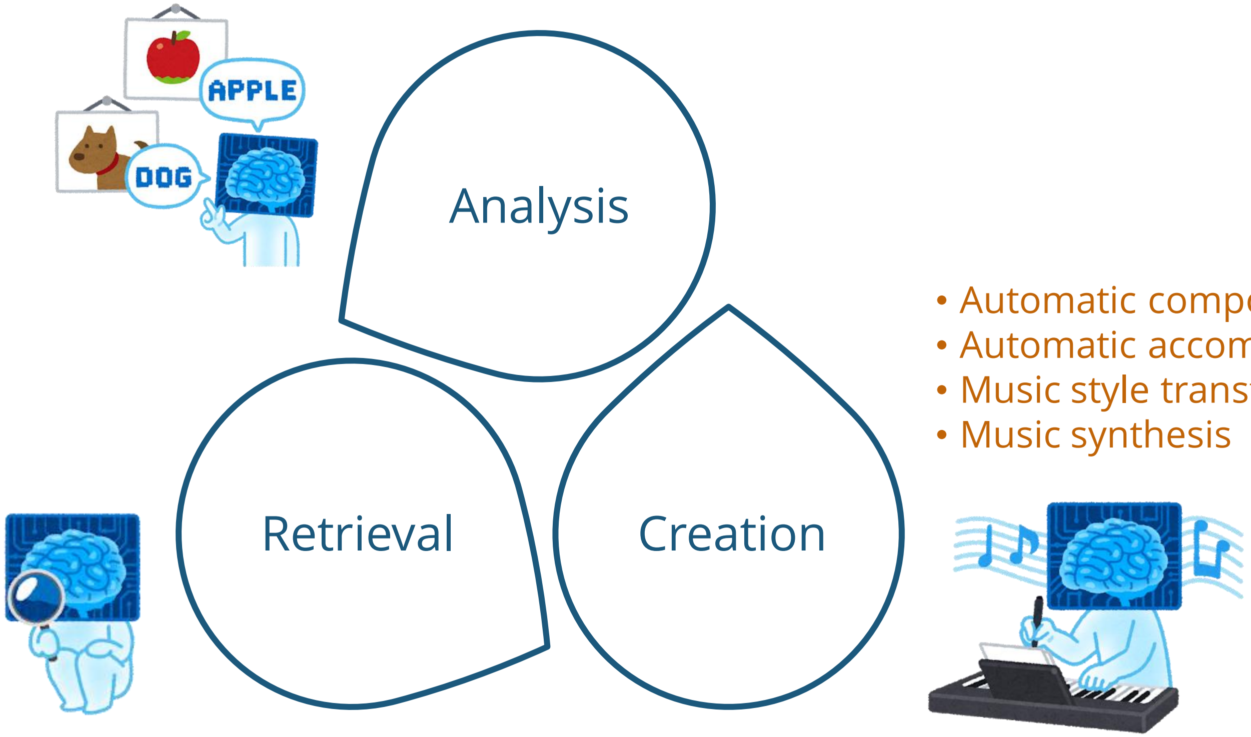


Image-to-audio Synthesis

Liu et al., “AudioLDM: Text-to-Audio Generation with Latent Diffusion Models,” *ICML*, 2023.
Dong et al., “CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models,” *WASPAA*, 2023.

Music Information Research (MIR)

- *"Intelligent ways to analyze, retrieve and create music"* (Yang 2018)



- Automatic composition
- Automatic accompaniment
- Music style transfer
- Music synthesis

Yang, "Music Information Research," *SNHCC*, TIGP, lecture notes, April 2018.

My Research

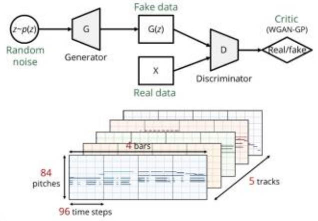
AI × Music

Multitrack Music Generation

Generating new music contents automatically

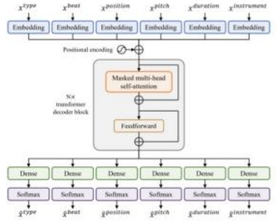
MuseGAN

(AAAI 2018)



Multitrack Music Transformer

(ICASSP 2023)

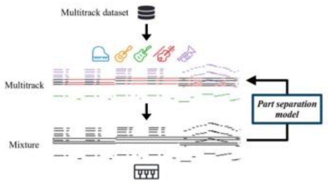

Embedding
Positional encoding
Masked multi-head self-attention
Feedforward
N× transformer decoder block
Dense
Softmax

Assistive Music Creation Tools

Assisting humans to create and perform music

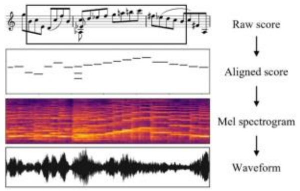
Arranger

(ISMIR 2021)



Deep Performer

(ICASSP 2022)

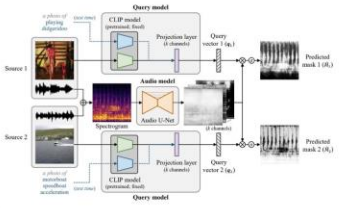


Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

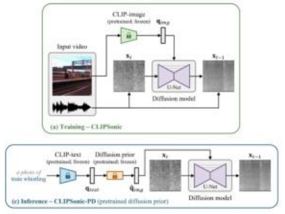
CLIPSep

(ICLR 2023)



CLIPSonic

(WASPAA 2023)


CLIP-image (pretrained, frozen)
Input video
U-Net
Diffusion model
(a) Training – CLIPSonic
CLIP-text (pretrained, frozen)
Diffusion prior (pretrained, frozen)
(c) Inference – CLIPSonic-PD (pretrained diffusion prior)

My Research

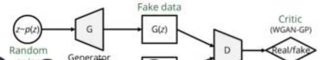
AI × Music

Multitrack Music Generation

Generating new music contents automatically

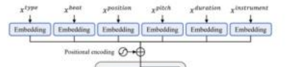
MuseGAN

(AAAI 2018)



Multitrack Music Transformer

(ICASSP 2023)



Featured in
Amazon AWS DeepComposer

Assistive Music Creation Tools

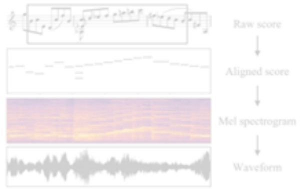
Assisting humans to create and perform music

Arranger

(ISMIR 2021)

Deep Performer

(ICASSP 2022)

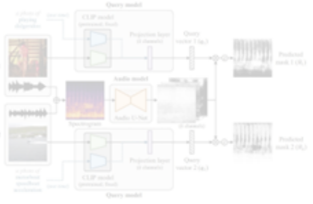


Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep

(ICLR 2023)



CLIPSonic

(WASPAA 2023)

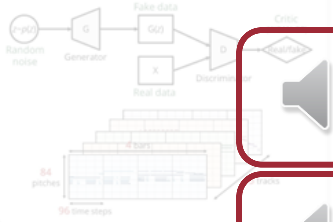
My Research

AI × Music

Multitrack Music Generation

Generating new music contents automatically

MuseGAN
(AAAI 2018)

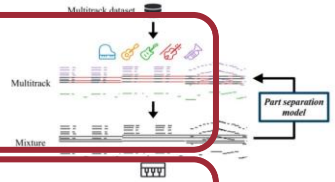


Multitrack Music Transformer
(ICASSP 2023)

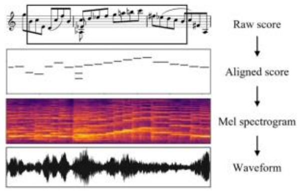
Assistive Music Creation Tools

Assisting humans to create and perform music

Arranger
(ISMIR 2021)



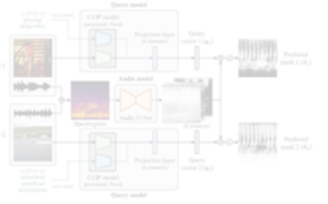
Deep Performer
(ICASSP 2022)



Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

CLIPSep
(ICLR 2023)



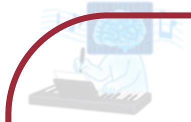
CLIPSonic
(WASPAA 2023)

My Research

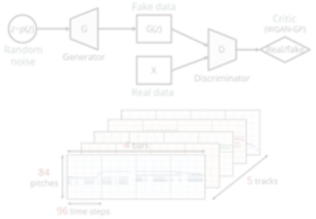
AI × Music

Multitrack Music Generation

Generating new music contents automatically



MuseGAN
(AAAI 2018)



Multitrack Music Transformer
(ICASSP 2023)

Assistive Music Creation Tools

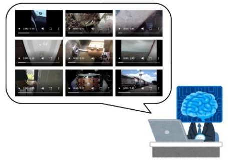
Assisting humans to create and perform music

Arranger
(ISMIR 2021)

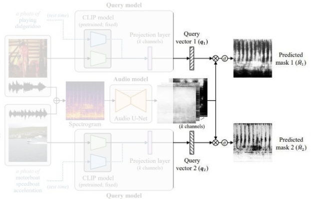
Deep Performer
(ICASSP 2022)

Multimodal Learning for Audio & Music

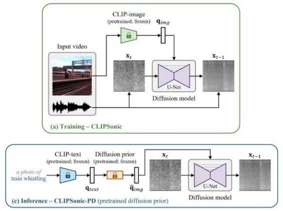
Learning sound separation and synthesis from videos



CLIPSep
(ICLR 2023)


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

CLIPSonic
(WASPAA 2023)


CLIP-image (pretrained; frozen)
$\mathbf{q}_{img}$
Input video
$\mathbf{x}_t$
$\mathbf{x}_{t-1}$
U-Net
Diffusion model
(a) Training – CLIPSonic
CLIP-text (pretrained; frozen)
Diffusion prior (pretrained; frozen)
a photo of train whistling
$\mathbf{q}_{text}$
$\hat{\mathbf{q}}_{img}$
(c) Inference – CLIPSonic-PD (pretrained diffusion prior)

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

Hao-Wen Dong[1,2] * Naoya Takahashi[1] † Yuki Mitsufuji[1]

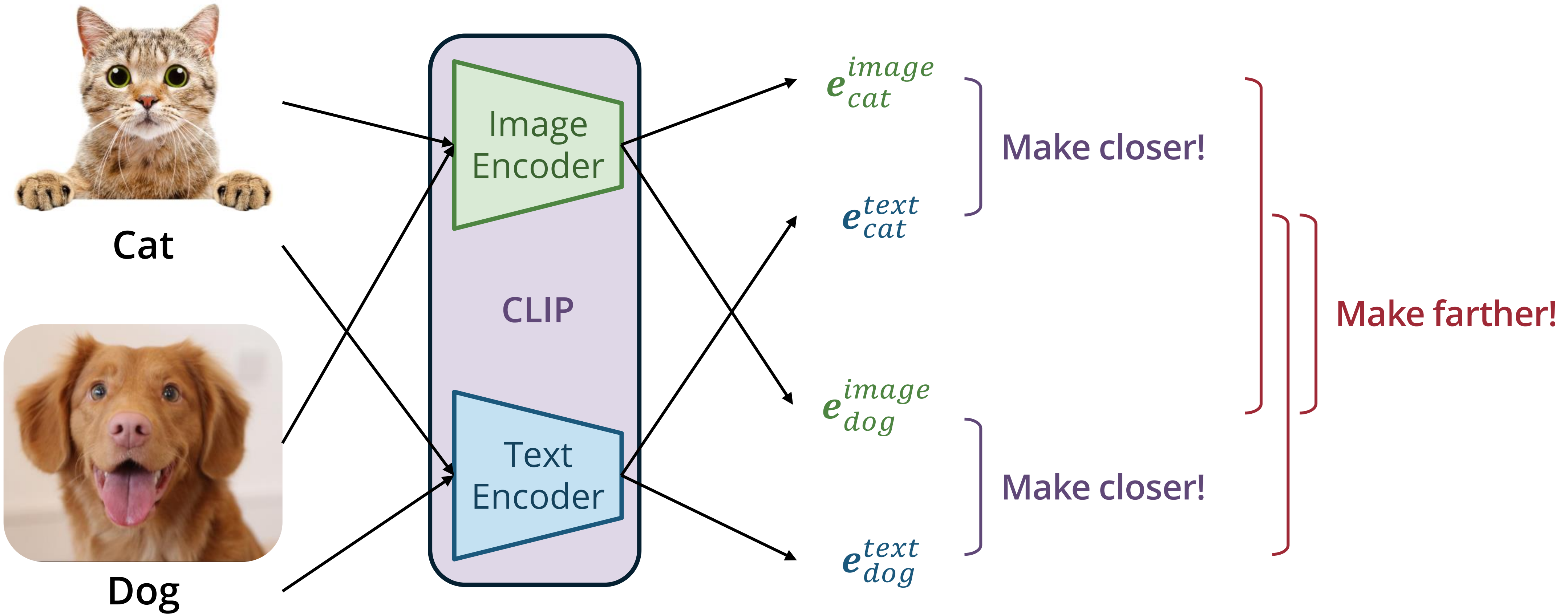
Julian McAuley[2] Taylor Berg-Kirkpatrick[2]

[1]Sony Corporation [2]University of California San Diego

* Work done during an internship at Sony † Corresponding author

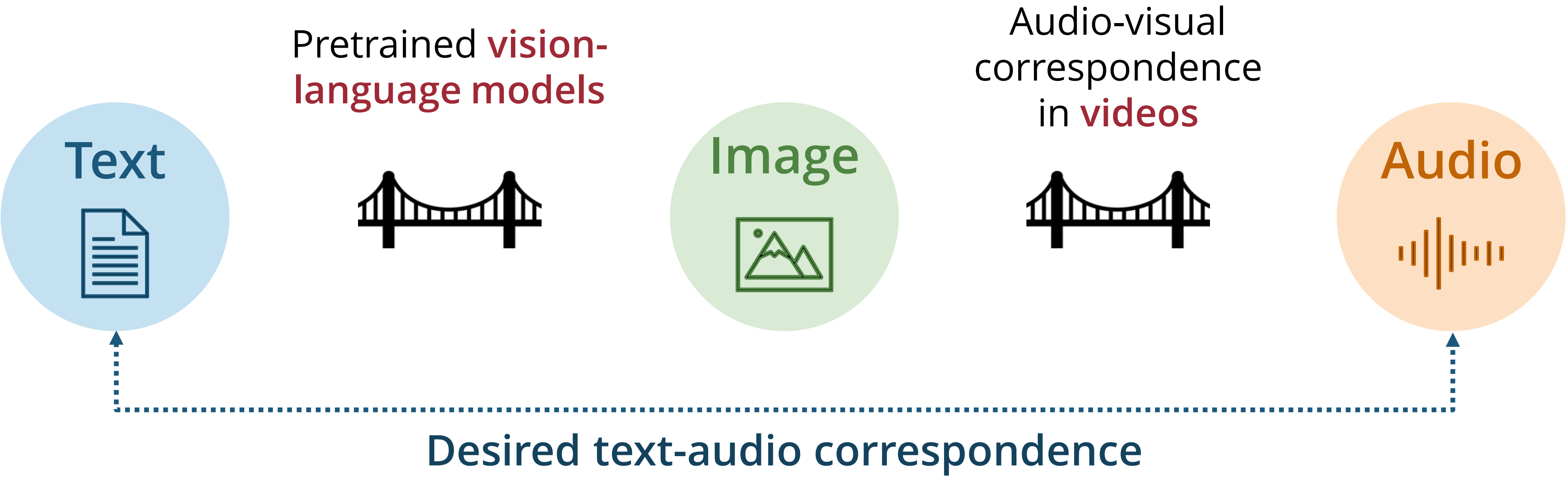
CLIP (Contrastive Language-Image Pretraining)

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



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

Leveraging the Visual Domain as a Bridge



Text

Pretrained **vision-language models**

Image

Audio-visual correspondence in **videos**

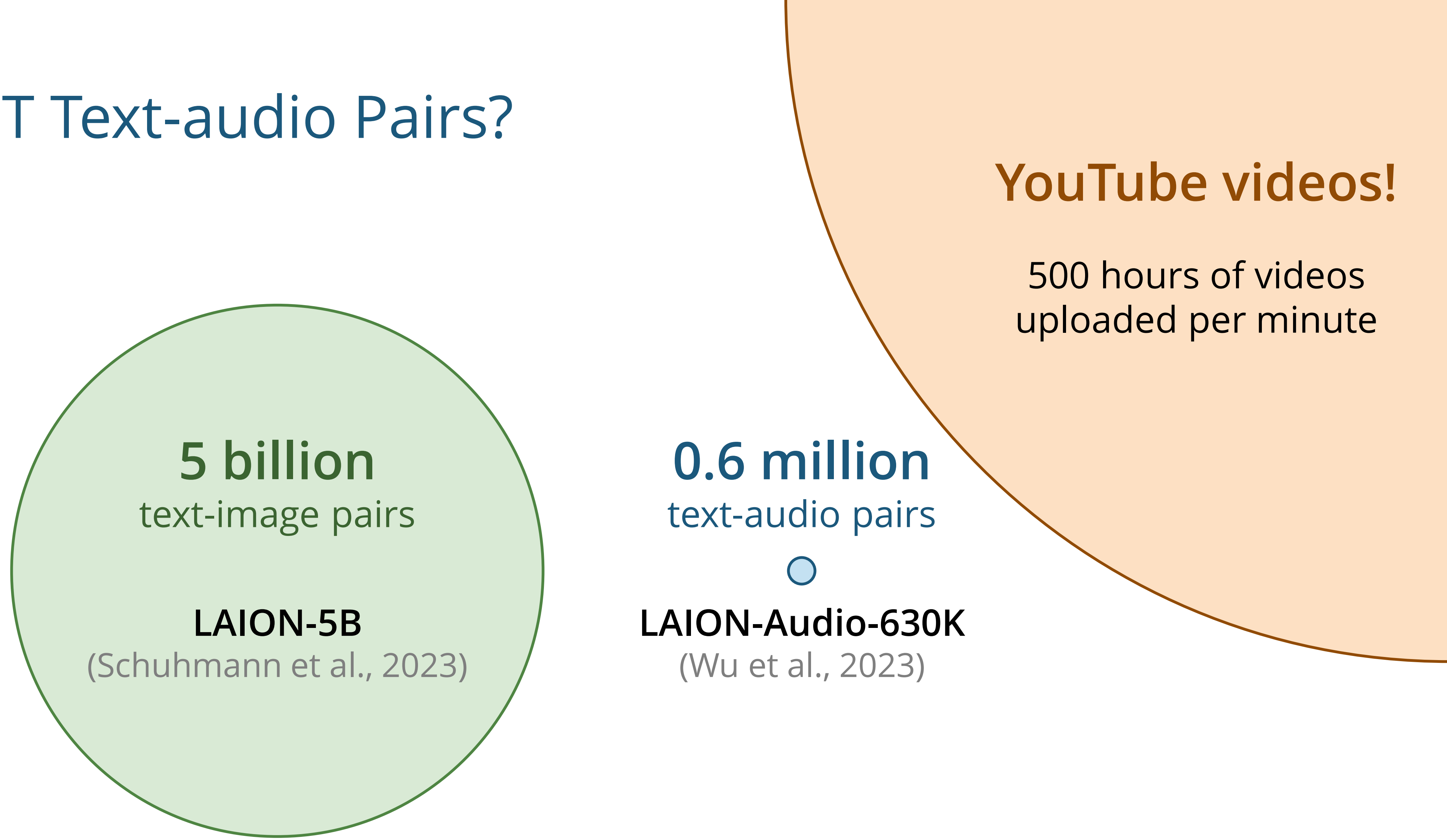
Audio

Desired text-audio 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

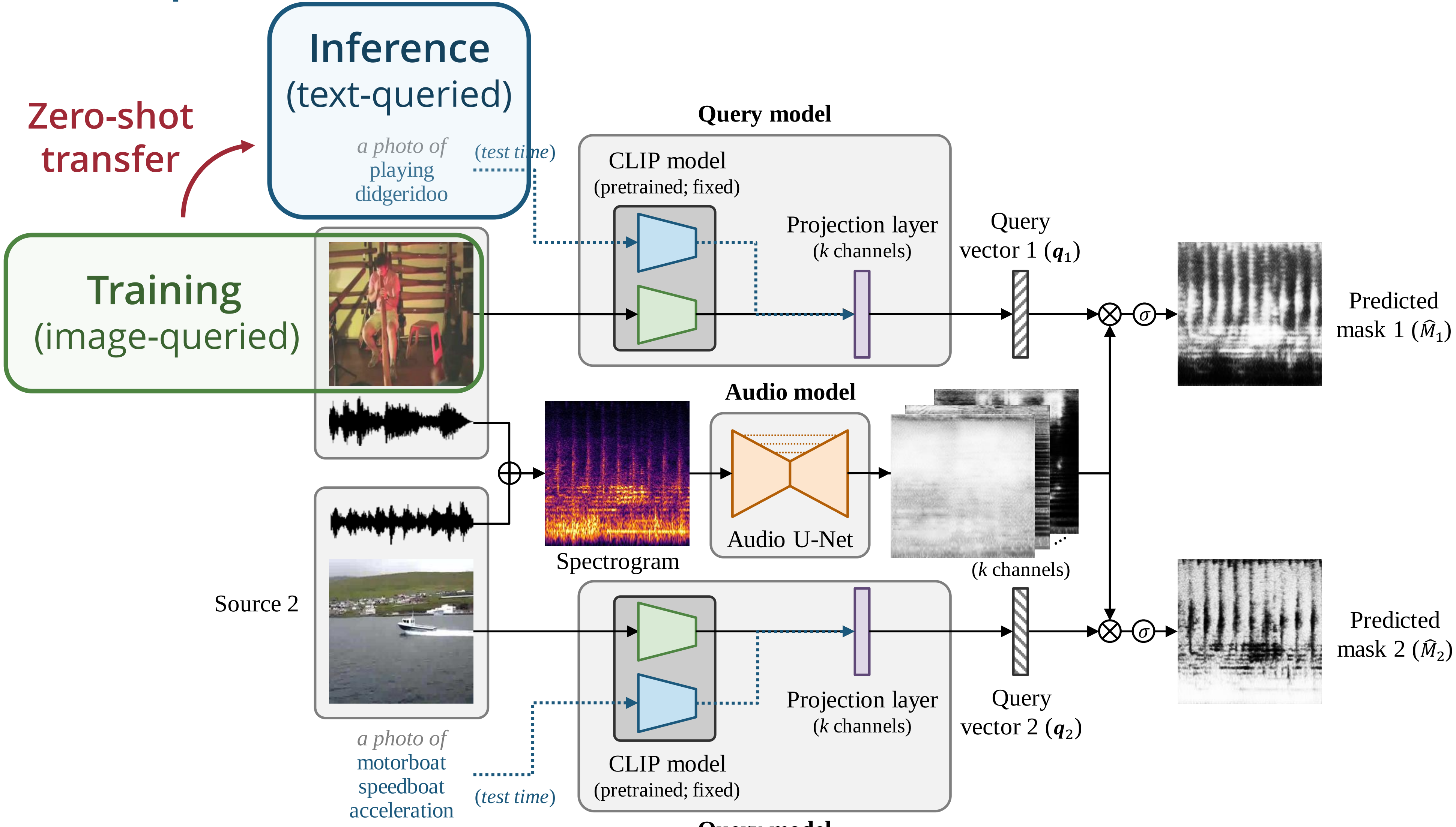
LAION-5B
(Schuhmann et al., 2023)

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.

CLIPSep


Zero-shot transfer
Inference
(text-queried)
a photo of
playing
didgeridoo
(test time)
Training
(image-queried)
Query model
CLIP model
(pretrained; fixed)
Projection layer
(k channels)
Query
vector 1 ($\boldsymbol{q}_1$)
Predicted
mask 1 ($\hat{M}_1$)
Audio model
Spectrogram
Audio U-Net
(k channels)
Source 2
Projection layer
(k channels)
Query
vector 2 ($\boldsymbol{q}_2$)
Predicted
mask 2 ($\hat{M}_2$)
CLIP model
(pretrained; fixed)
Query model
a photo of
motorboat
speedboat
acceleration
(test time)
σ

Data

MUSIC

(Zhao et al., 2018)

Violin

Acoustic guitar

Accordion

Music instrument playing videos

VGGSound

(Chen et al., 2020)

Hedge trimmer running

Dog bow-wow

Bird chirping, tweeting

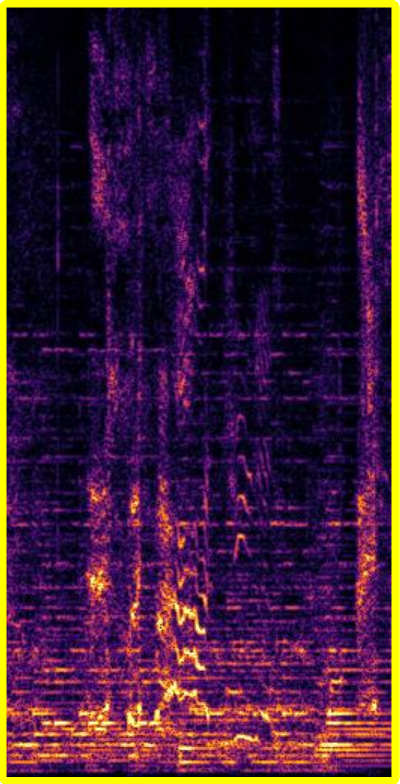
Noisy videos with diverse sounds

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

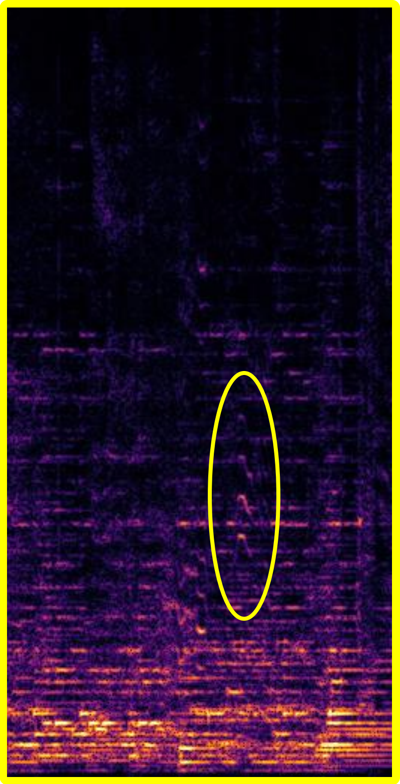
Demo – CLIPSep

Query: “*playing harpsichord*”

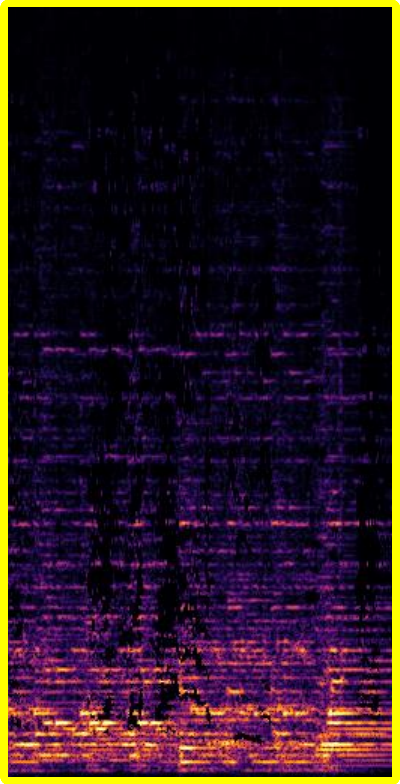
Mixture



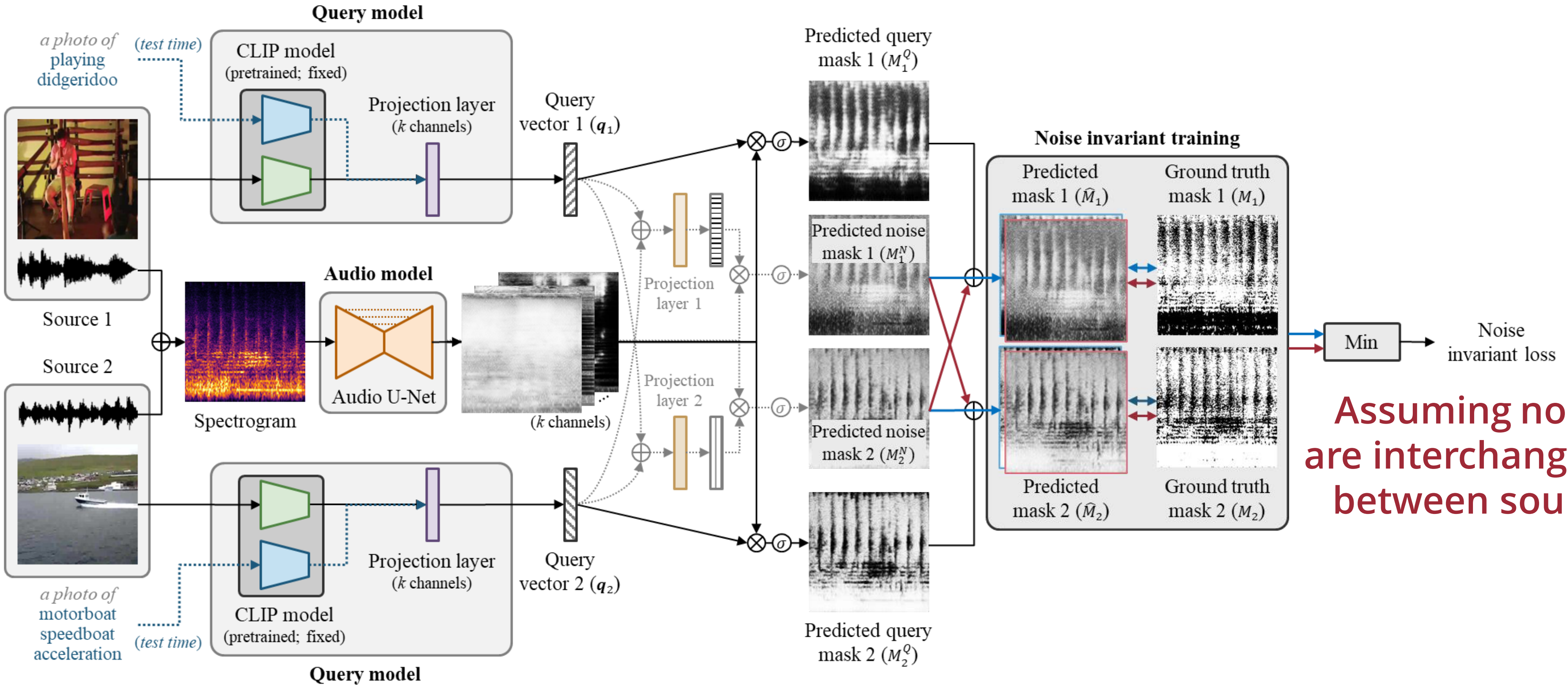
CLIPSep



Ground truth



Noise Invariant Training (NIT)

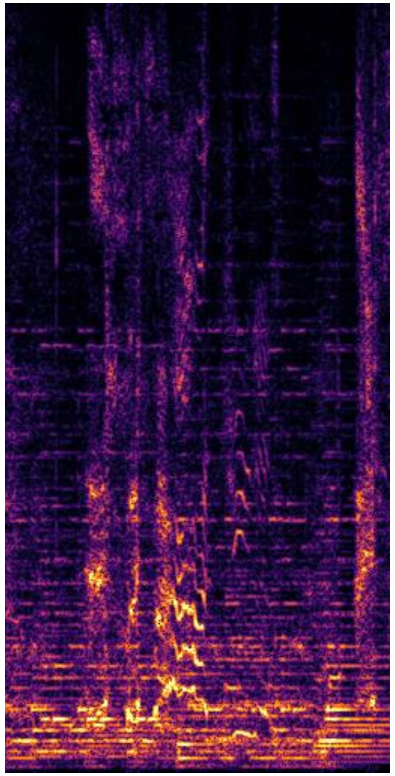


Assuming noises are interchangeable between sources

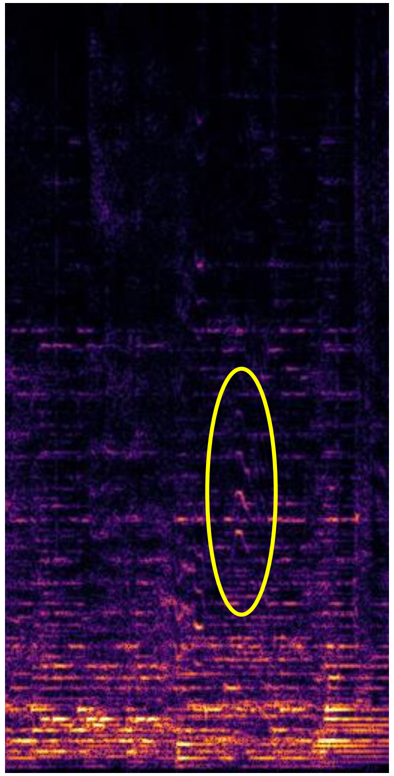
Demo – CLIPSep-NIT

Query: “*playing harpsichord*”

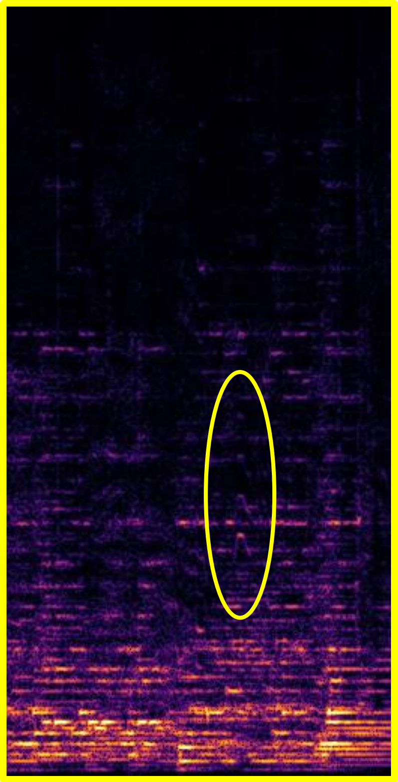
Mixture



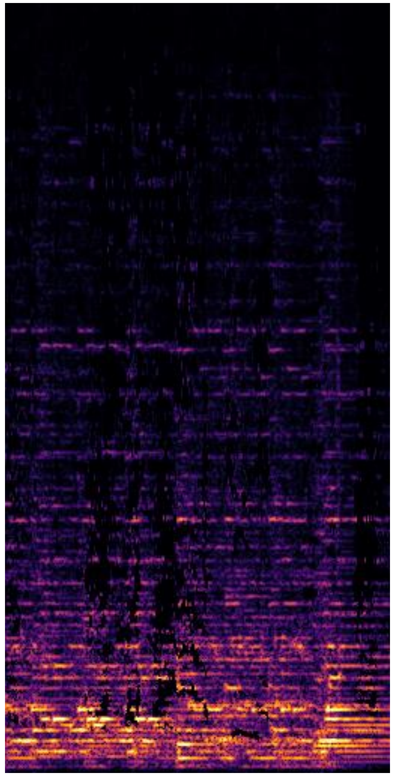
CLIPSep



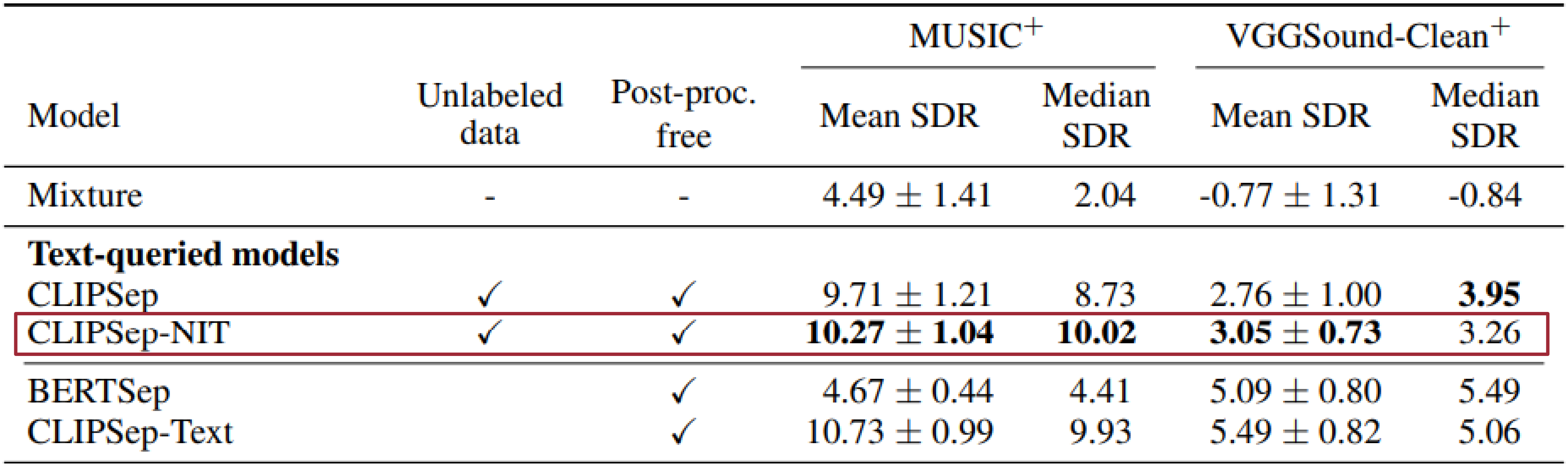
CLIPSep-NIT



Ground truth



Quantitative Results

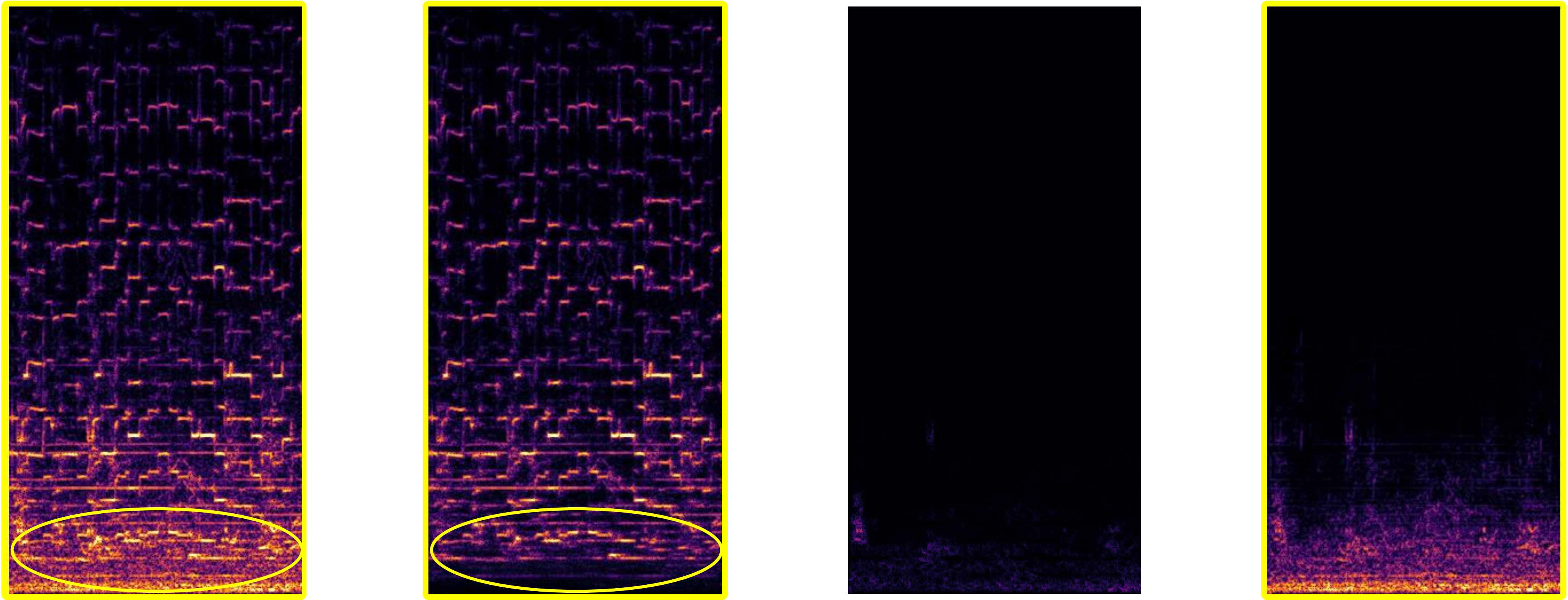


Model	Unlabeled data	Post-proc. free	MUSIC$^+$		VGGSound-Clean$^+$	
			Mean SDR	Median SDR	Mean SDR	Median SDR
Mixture	-	-	4.49 ± 1.41	2.04	-0.77 ± 1.31	-0.84
Text-queried models						
CLIPSep	✓	✓	9.71 ± 1.21	8.73	2.76 ± 1.00	**3.95**
CLIPSep-NIT	✓	✓	**10.27 ± 1.04**	**10.02**	**3.05 ± 0.73**	3.26
BERTSep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

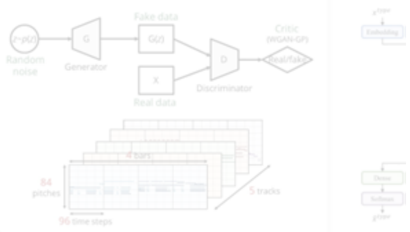
My Research

Multitrack Music Ge

Generating new music contents automatically

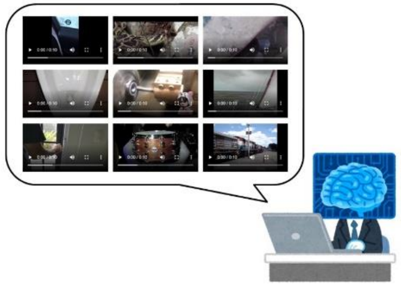
MuseGAN
(AAAI 2018)

Multit

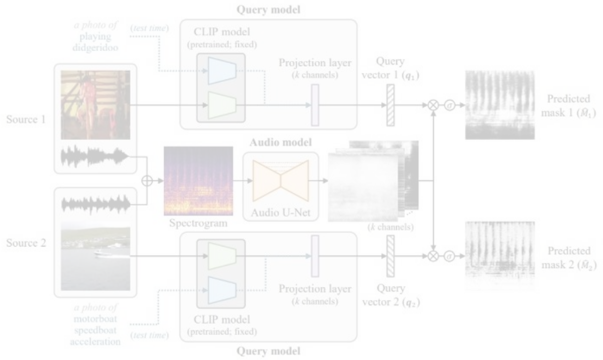


Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

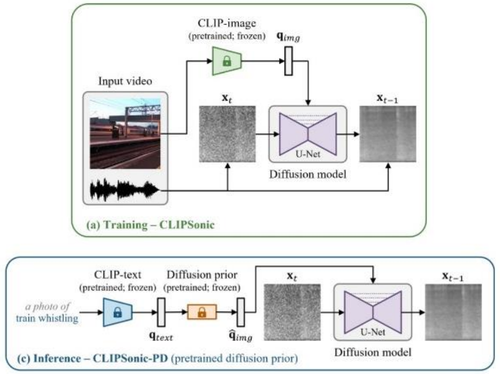


CLIPSep
(ICLR 2023)



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
a photo of motorboat speedboat acceleration
Query vector 2 (q_2)
Predicted mask 2 ($\hat{M}_2$)

CLIPSonic
(WASPAA 2023)



(a) Training – CLIPSonic

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

earning for Audio & Music

separation
rom videos

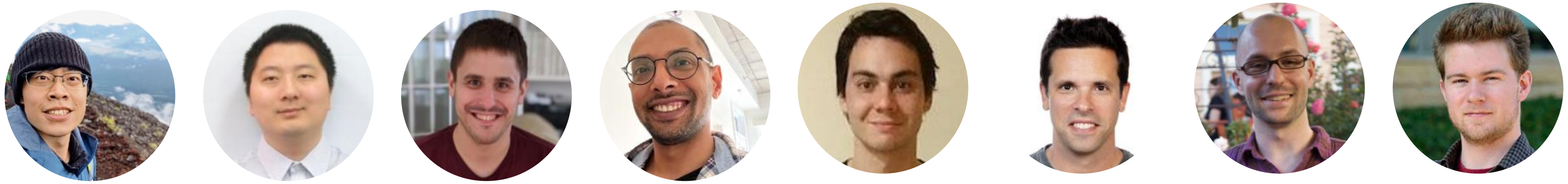
CLIPSonic
(WASPAA 2023)

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

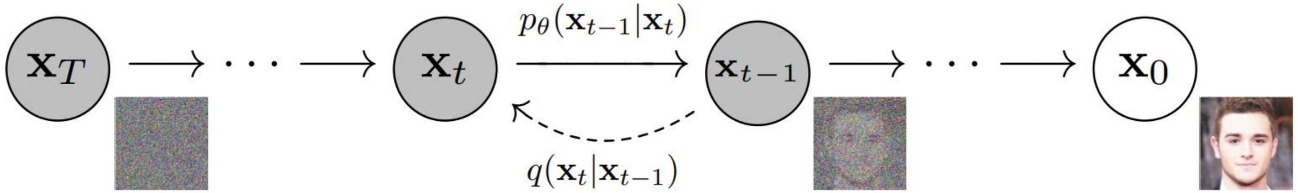


Dolby UC San Diego

Diffusion model

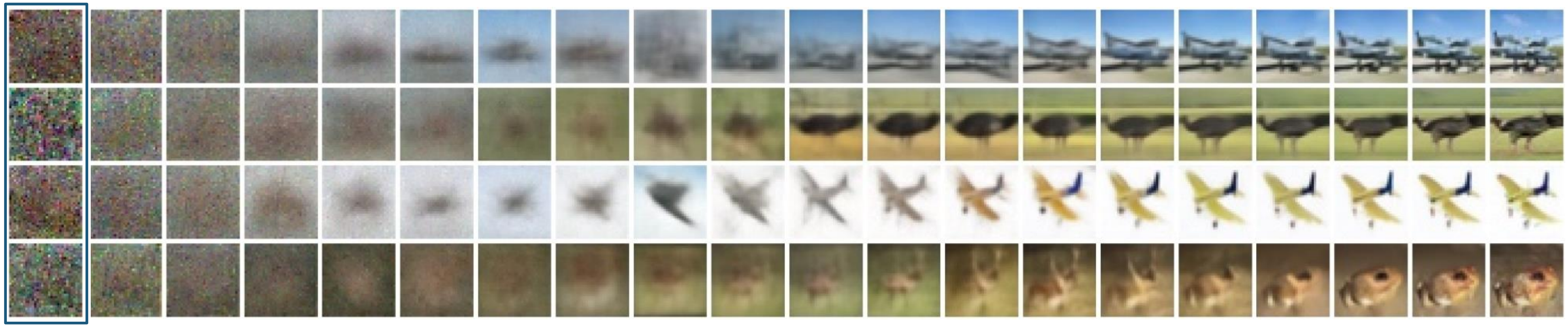
Add noise gradually

(Forward diffusion process)


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

Remove noise gradually

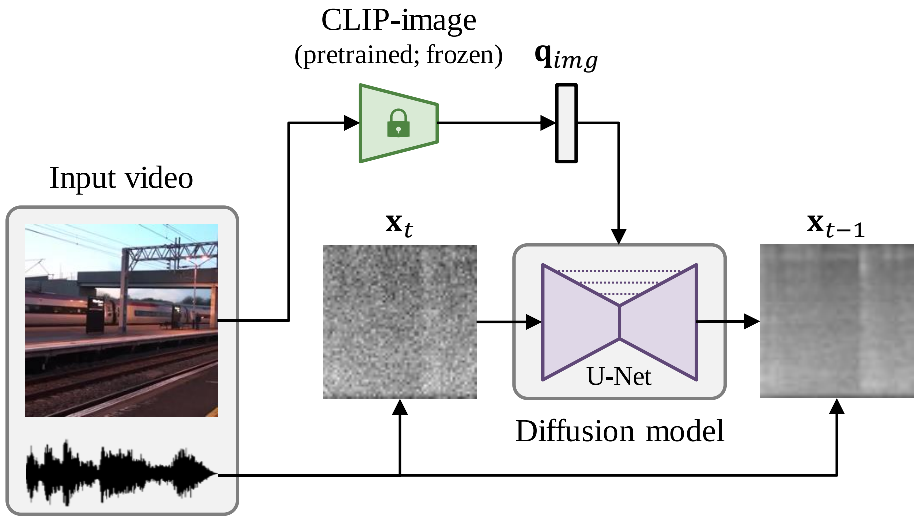
(Backward diffusion process)



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

CLIPSonic – Training (Image-queried)

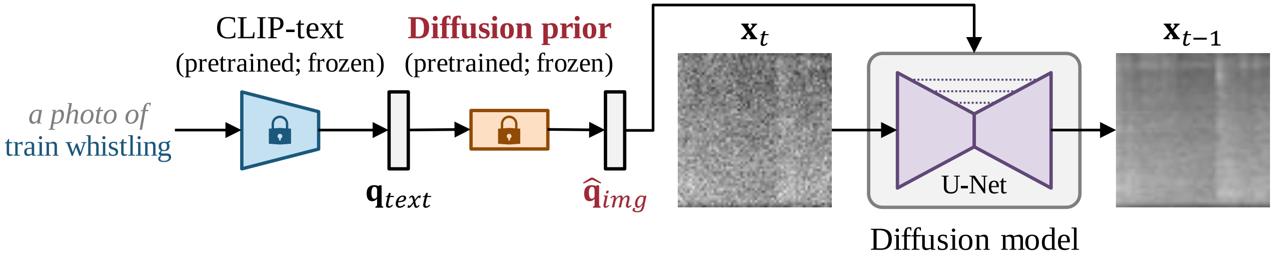
- We train the model to perform image-to-audio synthesis
  - Encode a video frame using a pretrained CLIP-image encoder (Radford et al., 2021)



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

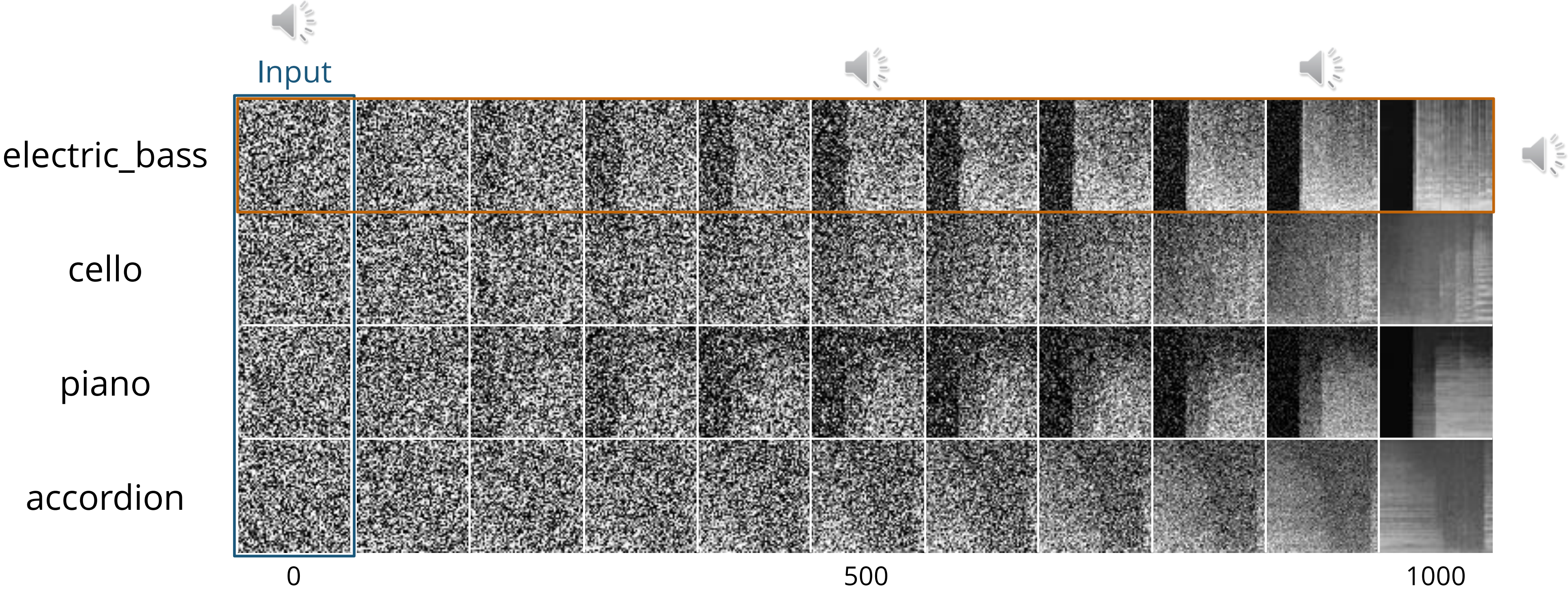
CLIPSonic – Inference (Text-queried)

- We use a pretrained diffusion prior model (Ramesh et al., 2022)
  - To generate a CLIP-image embedding given a CLIP-text embedding



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

CLIPSonic – Inference Examples



Data

MUSIC

(Zhao et al., 2018)

Violin

Acoustic guitar

Accordion

Music instrument playing videos

VGGSound

(Chen et al., 2020)

Hedge trimmer running

Dog bow-wow

Bird chirping, tweeting

Noisy videos with diverse sounds

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

Text-to-Audio Synthesis – Demo

Rapping

Sea waves

Thunder

Smoke detector beeping

Playing table tennis

Playing violin fiddle

Text-to-Audio Synthesis – Listening Test

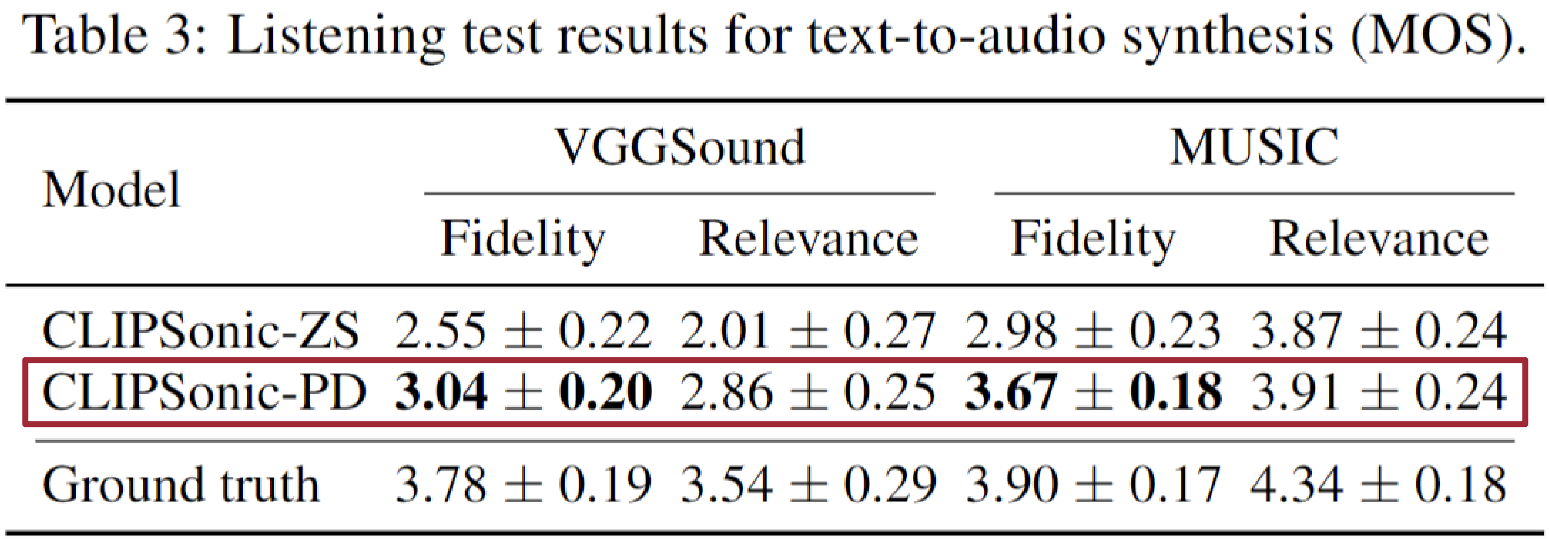


Table 3: Listening test results for text-to-audio synthesis (MOS).

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 – Listening Test

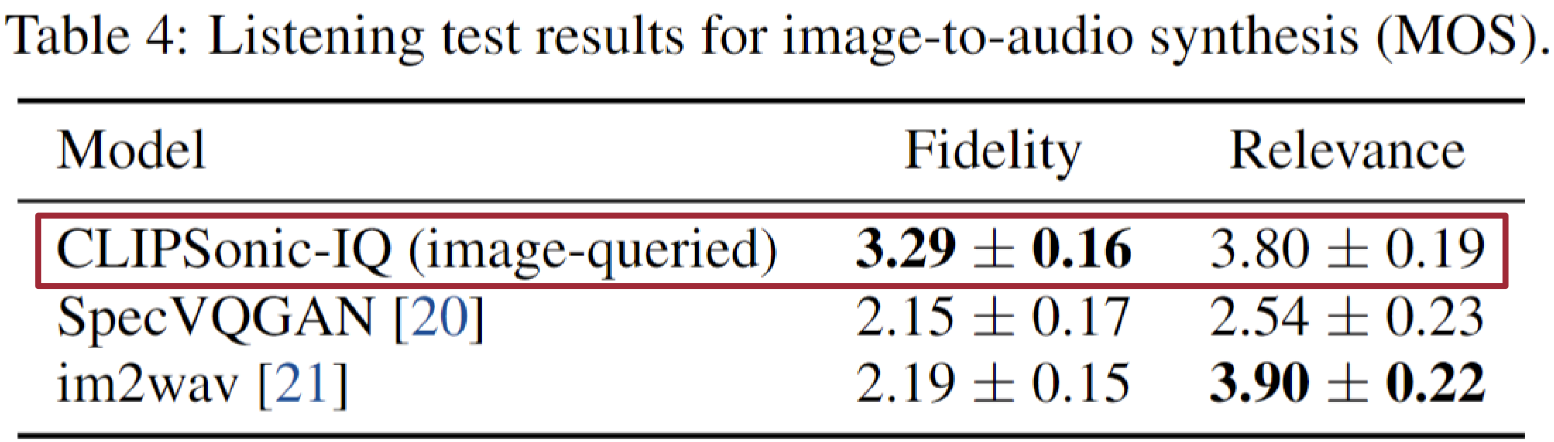


Table 4: Listening test results for image-to-audio synthesis (MOS).

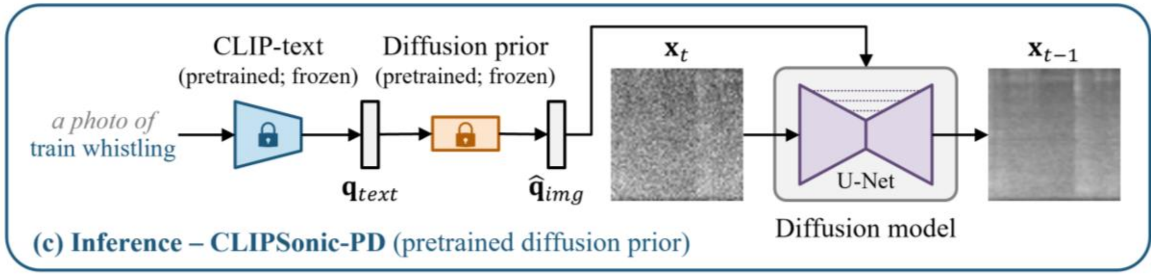
Model	Fidelity	Relevance
CLIPSonic-IQ (image-queried)	**3.29 ± 0.16**	3.80 ± 0.19
SpecVQGAN [20]	2.15 ± 0.17	2.54 ± 0.23
im2wav [21]	2.19 ± 0.15	**3.90 ± 0.22**

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

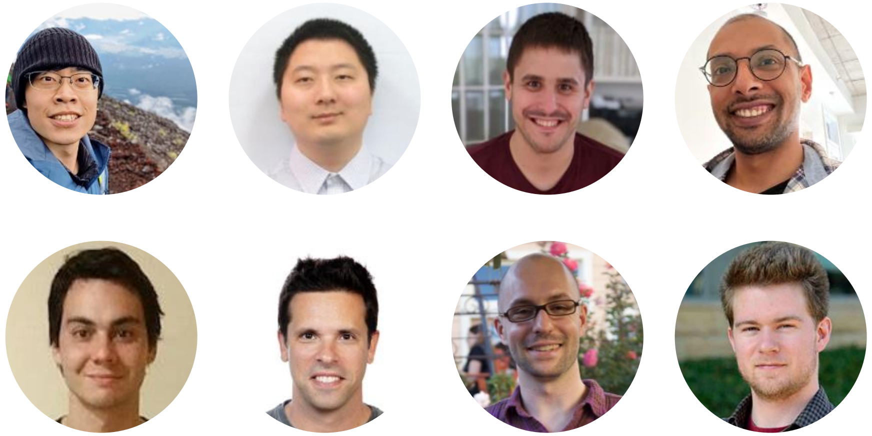
Sheffer and Adi, "I Hear Your True Colors: Image Guided Audio Generation," *ICASSP*, 2023.
Iashin and Rahtu, "Taming Visually Guided Sound Generation," *BMVC*, 2021.

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

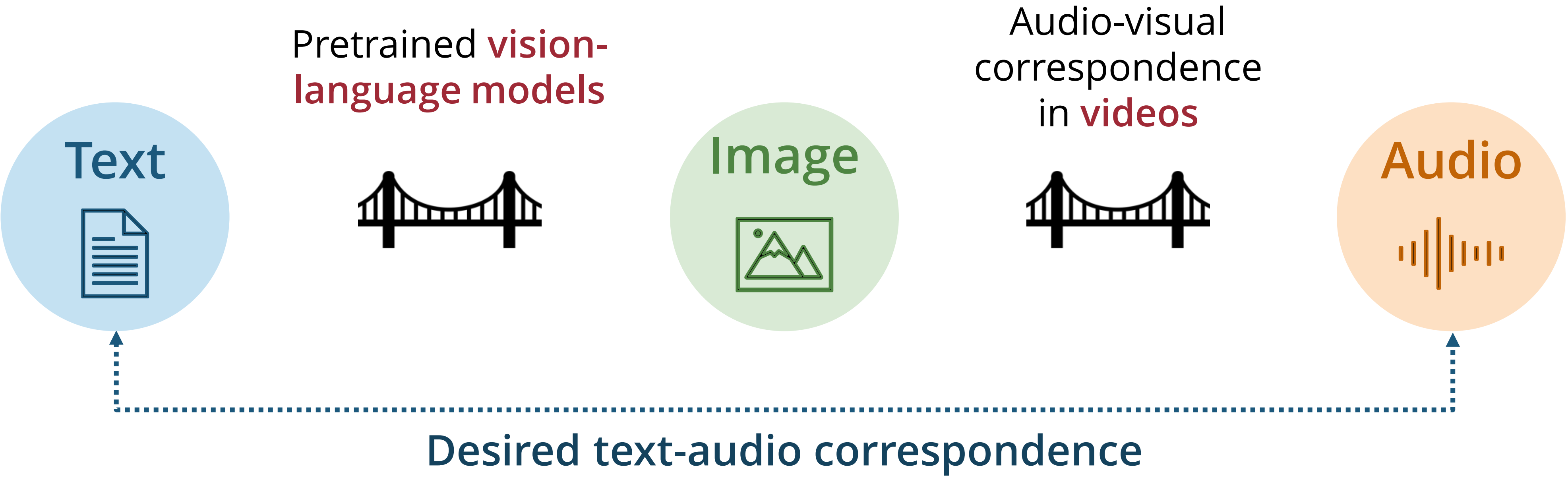


Paper: arxiv.org/abs/2306.09635
Demo: salu133445.github.io/clipsonic



Conclusion

Leveraging the Visual Domain as a Bridge

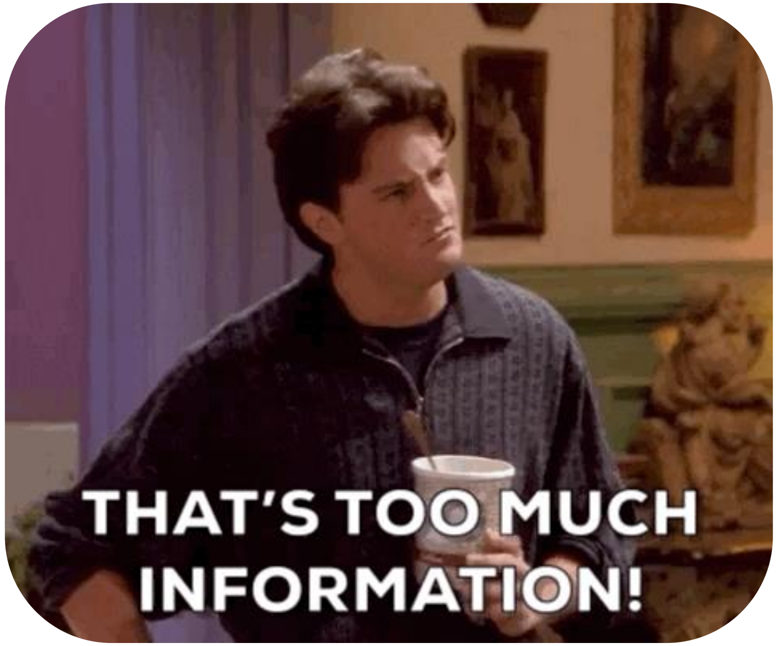

Pretrained vision-language models
Audio-visual correspondence in videos
Text
Image
Audio
Desired text-audio correspondence

No text-audio pairs required!

Scalable to large video datasets!

A Lot More to Learn from Videos

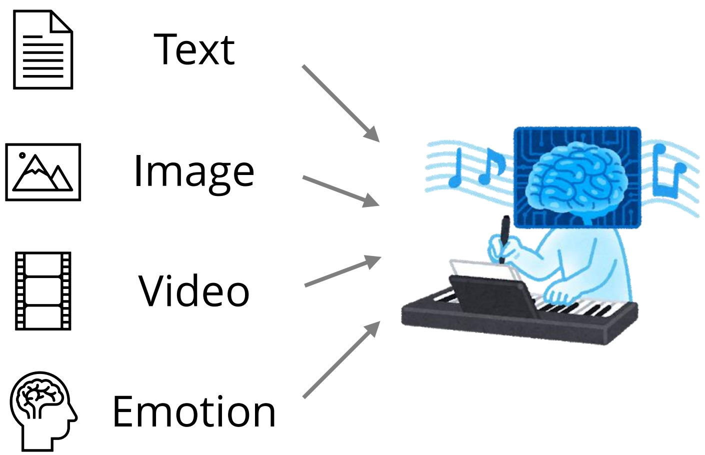
- Free audio-visual correspondence
- Rich context information
- Rich temporal dynamics



Future Directions

Challenges

Multimodality



Text

Image

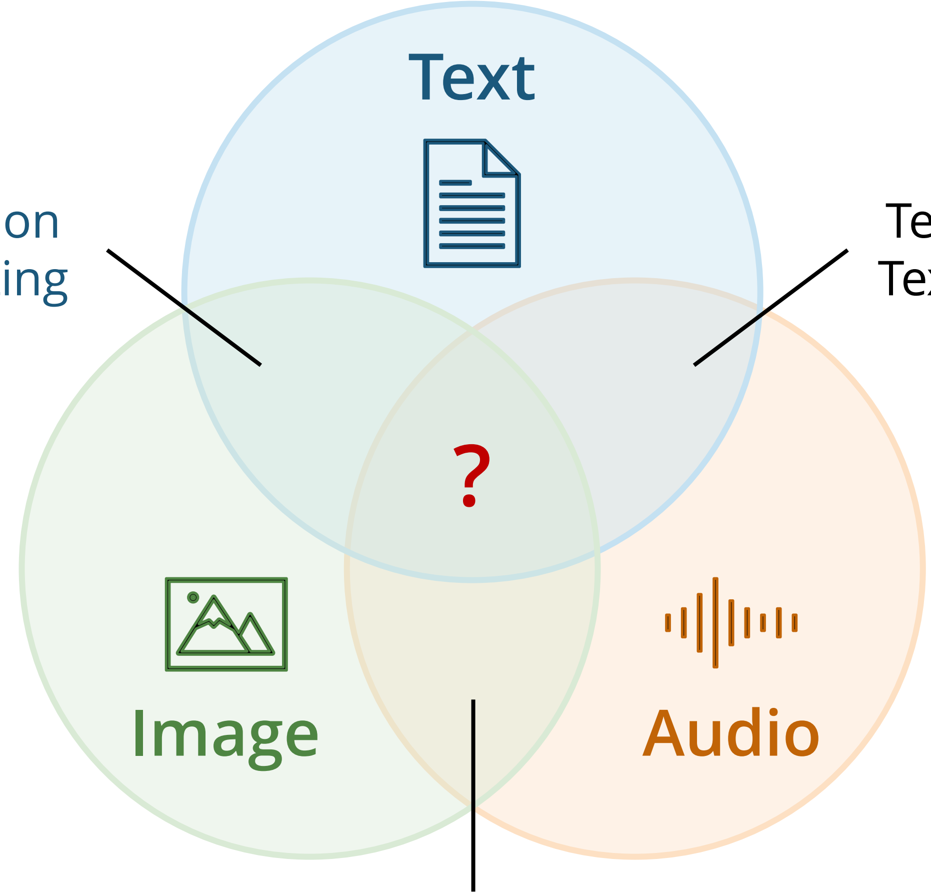
Video

Emotion

Usability

Licensing

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

Multimodal generative AI for Ads

Video	**Runway Gen-2**
Music	**MusicGen**

進捗共有チャンネル, “Runway Gen-2 Sample / Tiger Whiskey,” *YouTube*, July 9, 2023.
Kaoru Naito, *Twitter*, https://twitter.com/ka0ru_1620/status/1678313226453520385, July 10, 2023.

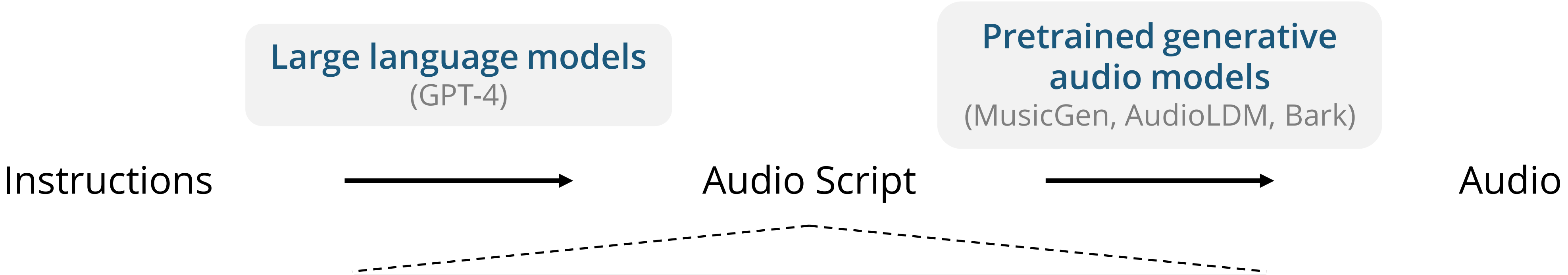
Generative AI for News

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.

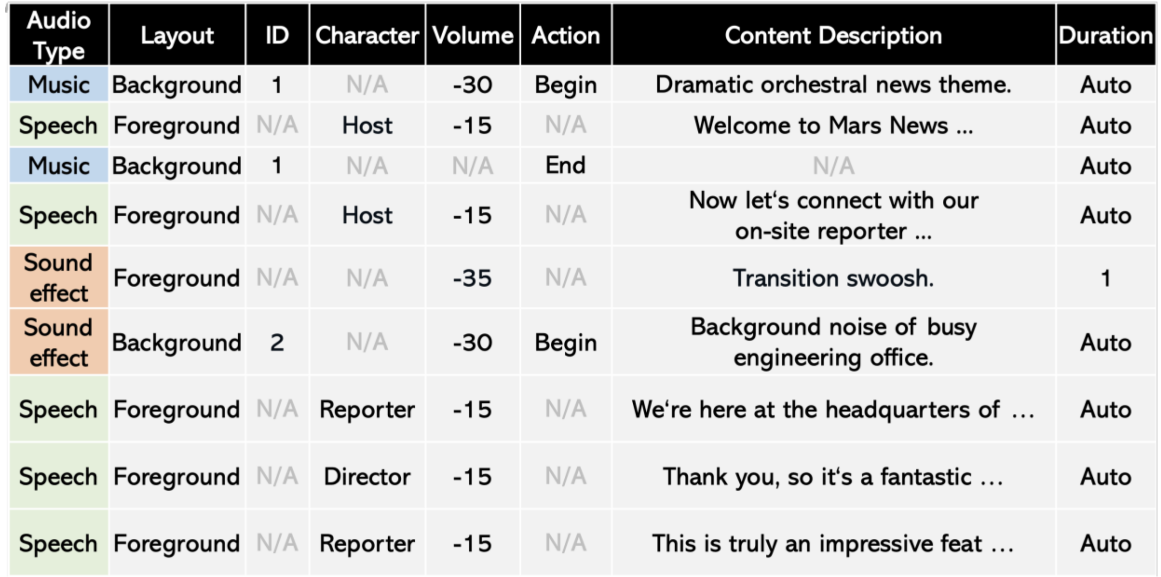
Controllable Generative AI



Large language models
(GPT-4)

Pretrained generative audio models
(MusicGen, AudioLDM, Bark)

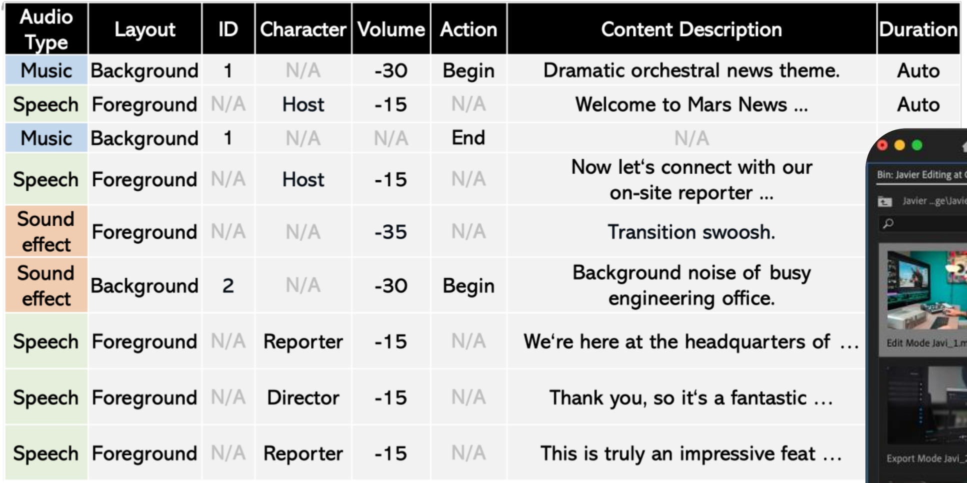
Instructions → Audio Script → Audio



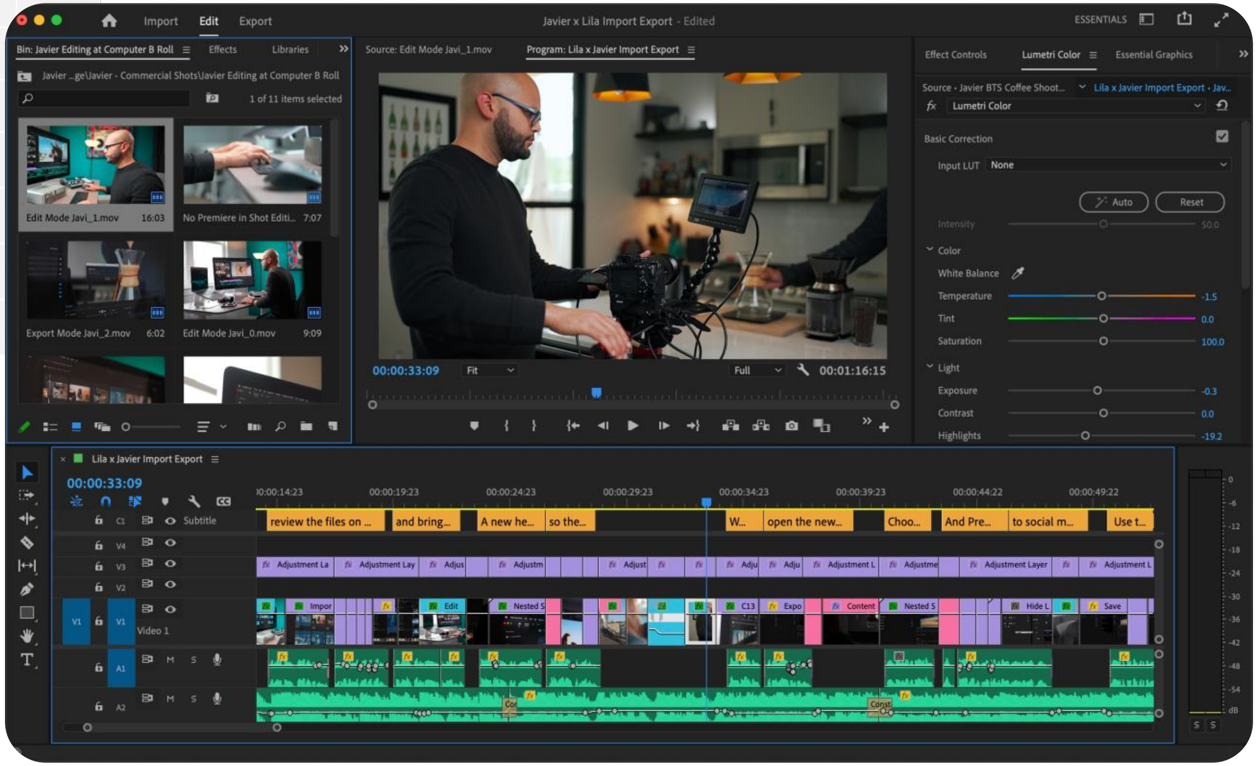
Audio Type	Layout	ID	Character	Volume	Action	Content Description	Duration
Music	Background	1	N/A	-30	Begin	Dramatic orchestral news theme.	Auto
Speech	Foreground	N/A	Host	-15	N/A	Welcome to Mars News ...	Auto
Music	Background	1	N/A	N/A	End	N/A	Auto
Speech	Foreground	N/A	Host	-15	N/A	Now let's connect 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

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

Controllable Generative AI

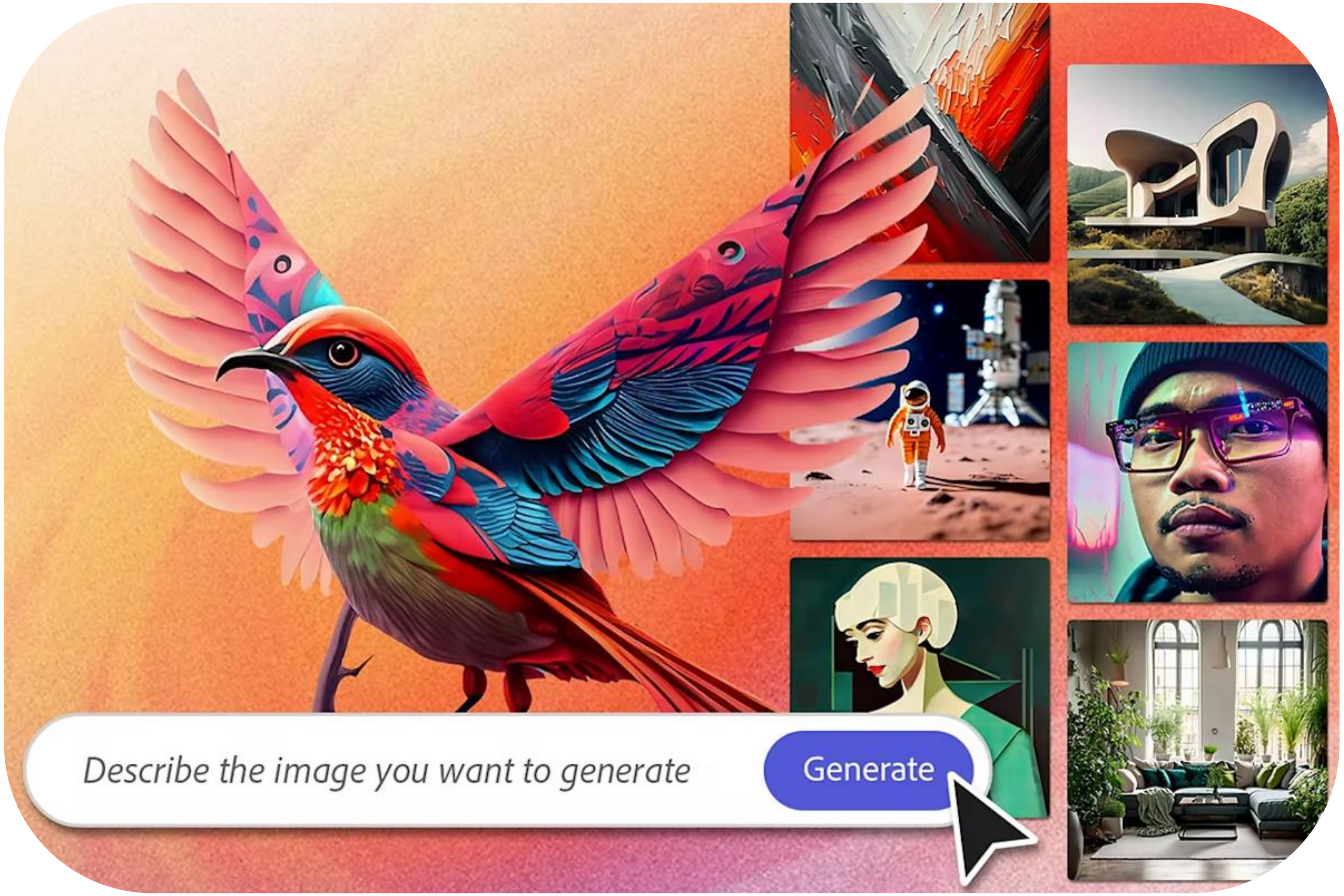


Audio Type	Layout	ID	Character	Volume	Action	Content Description	Duration
Music	Background	1	N/A	-30	Begin	Dramatic orchestral news theme.	Auto
Speech	Foreground	N/A	Host	-15	N/A	Welcome to Mars News ...	Auto
Music	Background	1	N/A	N/A	End	N/A	
Speech	Foreground	N/A	Host	-15	N/A	Now let's connect with 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 …	



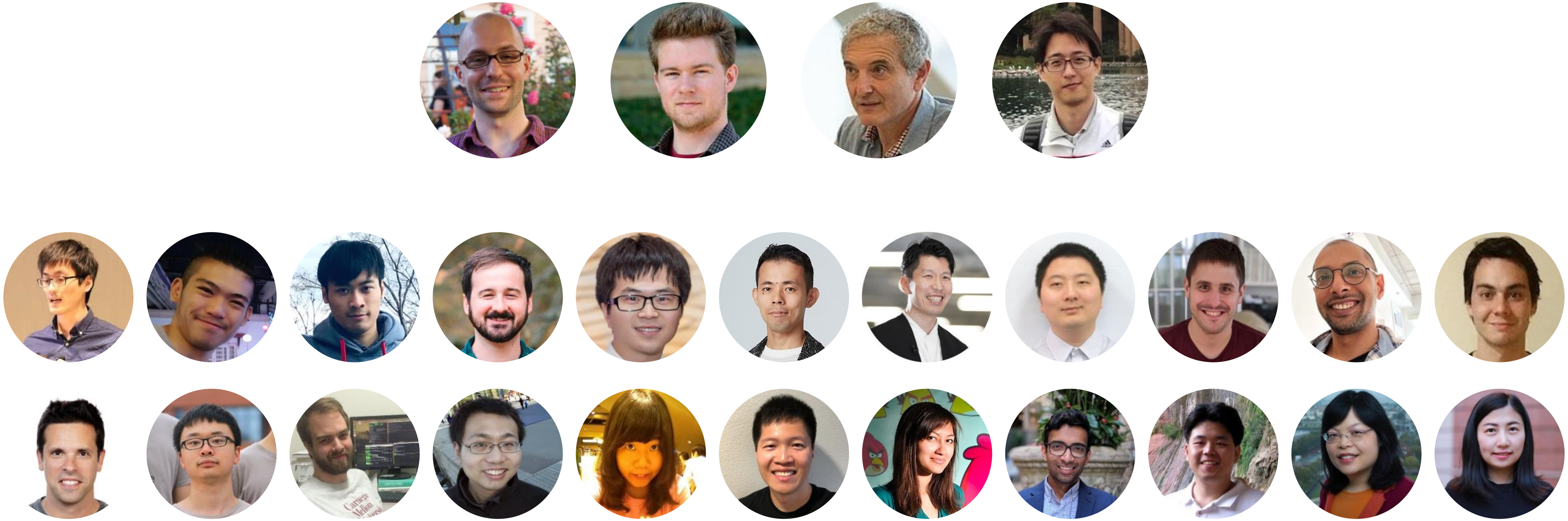
Integration into professional creative workflow

Licensing Example – Adobe Firefly



Trained with royalty-free Adobe Stock images

Acknowledgements



UC San Diego

中央研究院 ACADEMIA SINICA

Dolby

SONY

amazon

MINISTRY OF EDUCATION 教育部

erc

Thank you!