KAUST Rising Stars in AI Symposium 2024

# Learning Text-to-Audio Synthesis from Videos

**Hao-Wen (Herman) Dong**

UC San Diego

February 20, 2024

## AI for Music & Audio

*New technology creates new art form*



AI

**Empowering music and audio creation with machine learning**

Music & Audio



## Music & Audio for AI

*New art form inspires new technology*

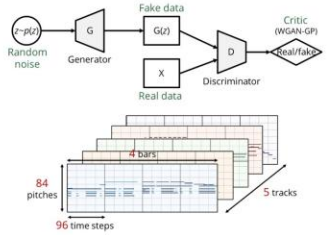
# Generative AI for Music & Audio

*Empowering music and audio creation with machine learning*

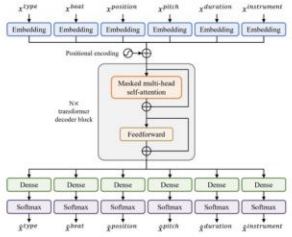
## Multitrack Music Generation

Advancing deep generative models for multitrack music

### MuseGAN (AAAI 2018)



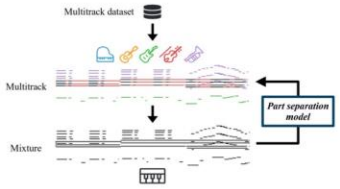

### MMT (ICASSP 2023)



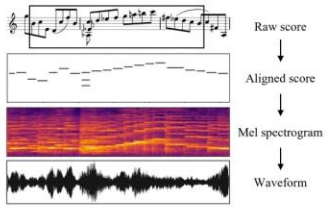

## Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

### Arranger (ISMIR 2021)



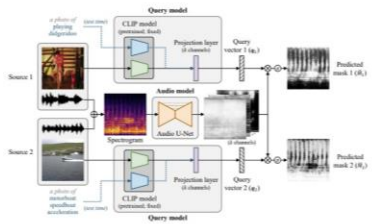

### Deep Performer (ICASSP 2022)



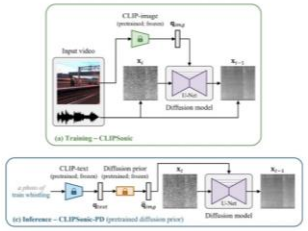

## Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

### CLIPSep (ICLR 2023)




### CLIPSonic (WASPAA 2023)



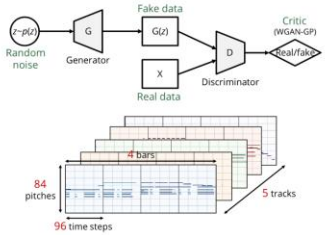
# Generative AI for Music & Audio

*Empowering music and audio creation with machine learning*

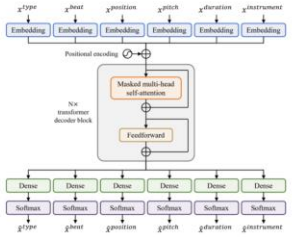
## Multitrack Music Generation

Advancing deep generative models for multitrack music

### MuseGAN (AAAI 2018)




### MMT (ICASSP 2023)




## Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools



### Arranger (ISMIR 2021)



### Deep Performer (ICASSP 2022)




## Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos



### CLIPSep (ICLR 2023)



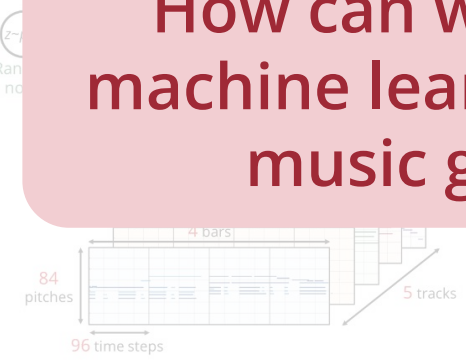
### CLIPSonic (WASPAA 2023)

# Generative AI for Music & Audio

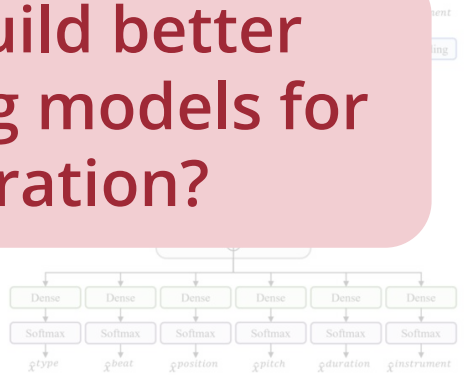
## Multitrack Music Generation

Advancing deep generative models for multitrack music

MuseGAN
(AAAI 2018)




MMT
(ICASSP 2023)



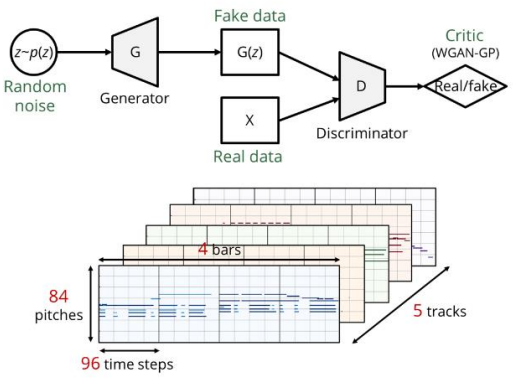

**How can we build better machine learning models for music generation?**

# Generative AI for Music & Audio

## Multitrack Music Generation

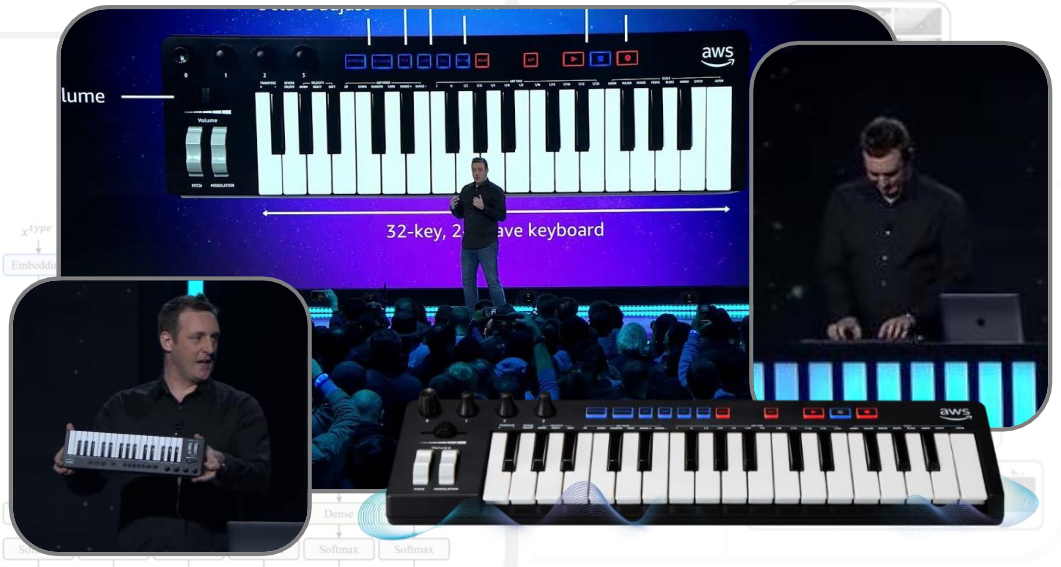
Advancing deep generative models for multitrack music

**MuseGAN**
(AAAI 2018)




First neural net that can **generate multi-instrument music from scratch**

Pop Music Generation



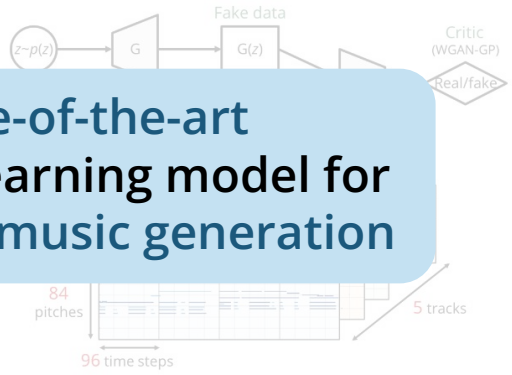

Featured in
Amazon AWS DeepComposer

# Generative AI for Music & Audio

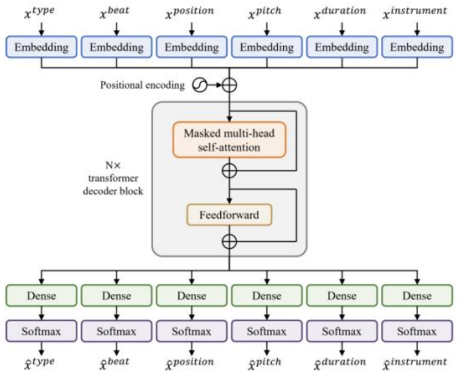
## Multitrack Music Generation

Advancing deep generative models for multitrack music

### MuseGAN (AAAI 2018)




### MMT (ICASSP 2023)




**State-of-the-art machine learning model for orchestral music generation**

**Orchestral Music Generation**

# Generative AI for Music & Audio

*Empowering music and audio creation with machine learning*

## Multitrack Music Generation

Advancing deep generative models for multitrack music



**MuseGAN**
(AAAI 2018)



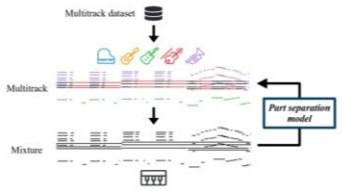
**MMT**
(ICASSP 2023)



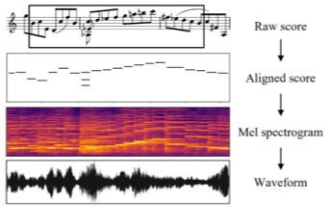
## Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

**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)

# Generative AI for Music & Audio

## Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

### Arranger (ISMIR 2021)


### Deep Performer (ICASSP 2022)

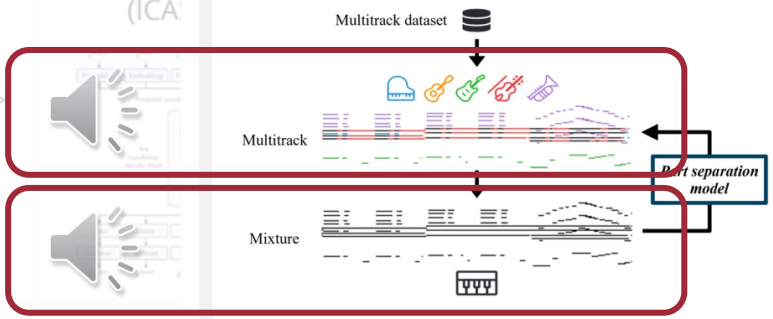

**How can AI help professionals and amateurs create music?**

# Generative AI for Music & Audio

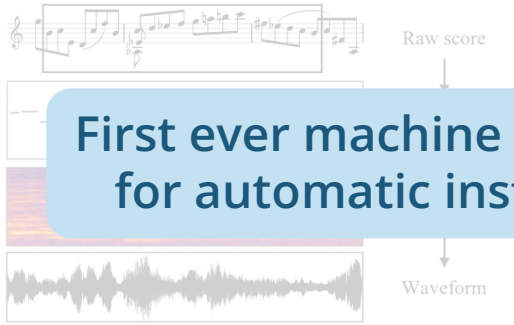
## Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

### Arranger
(ISMIR 2021)




### Deep Performer
(ICASSP 2022)



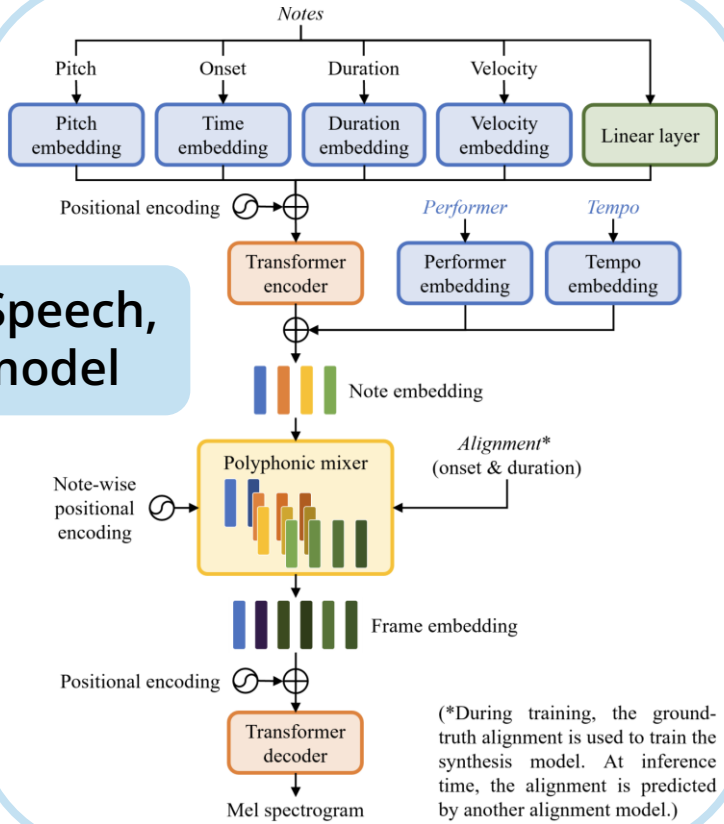

First ever machine learning model for automatic instrumentation

**Automatic Instrumentation**

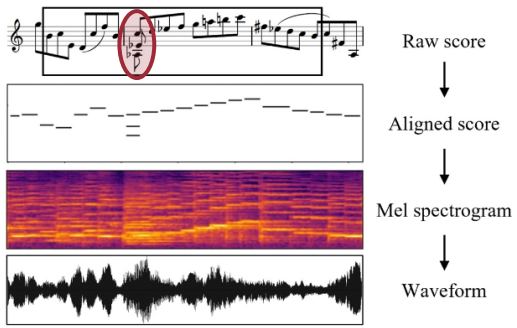
# Generative AI for Music & Audio

## Assistive Music Creation Tools

**Adapted from FastSpeech, a text-to-speech model**




### Deep Performer (ICASSP 2022)




**Score-to-audio synthesis**

# Generative AI for Music & Audio

*Empowering music and audio creation with machine learning*

## Multitrack Music Generation

Advancing deep generative models for multitrack music



### MuseGAN (AAAI 2018)



### MMT (ICASSP 2023)



## Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools



### Arranger (ISMIR 2021)



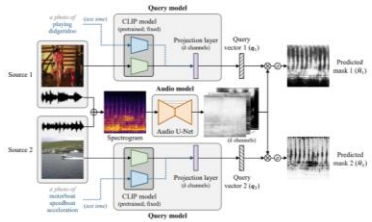
### Deep Performer (ICASSP 2022)



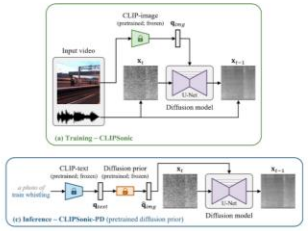
## Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

### CLIPSep (ICLR 2023)



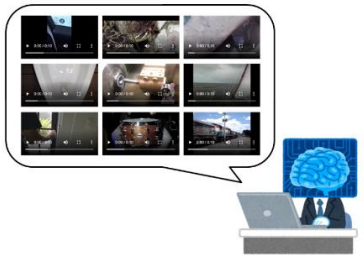
### CLIPSonic (WASPAA 2023)



# Generative AI for Music & Audio

## Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

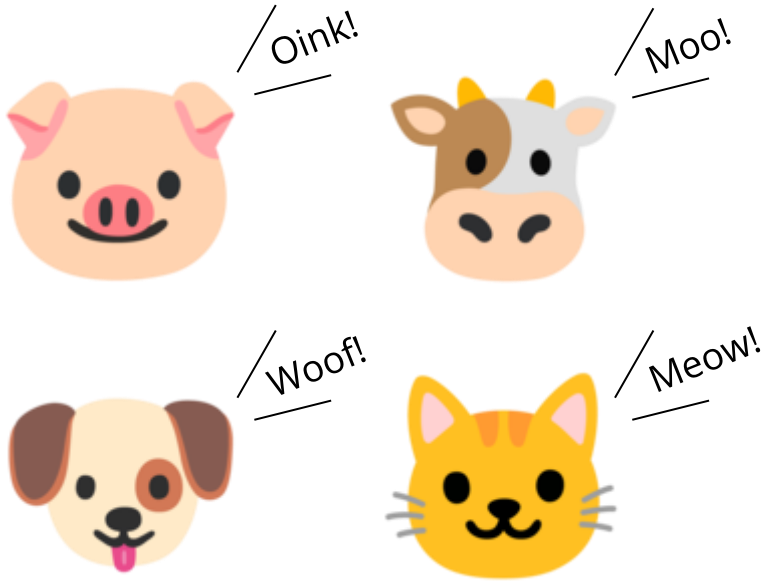


CLIPSep (ICLR 2023)

CLIPSonic (WASPAA 2023)

**How can we build AI systems that learn audio concepts like how humans do?**

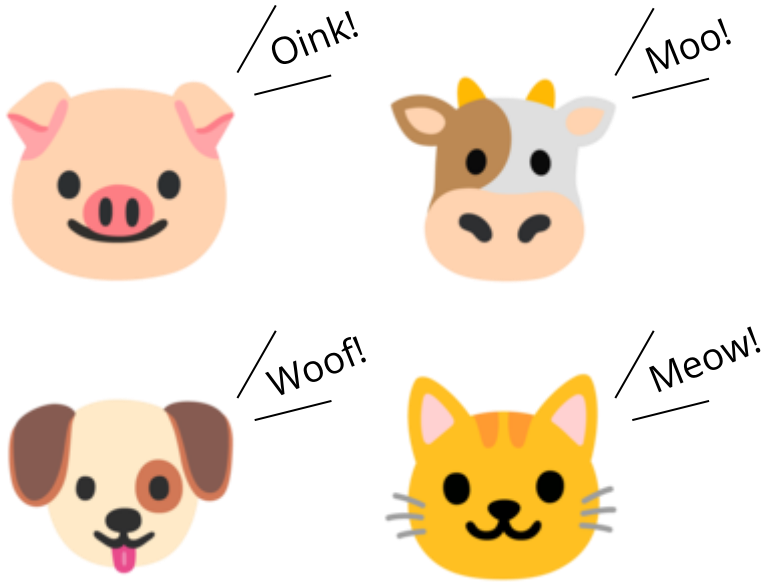
# Learning Sounds from Observations





*What does the fox say?*

# Learning Sounds from Observations



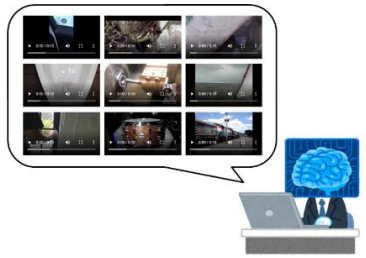

**Can machines learn to synthesize sounds from watching *noisy* videos?**

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

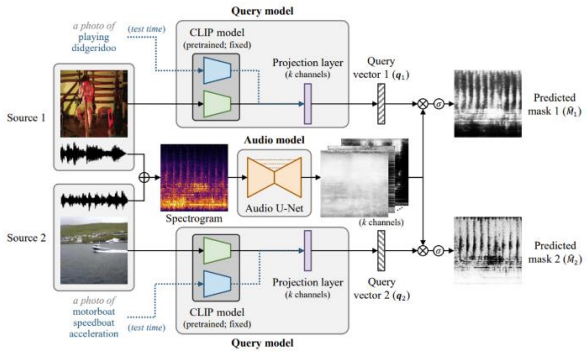
# Generative AI for Music & Audio

## Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos



### CLIPSep (ICLR 2023)

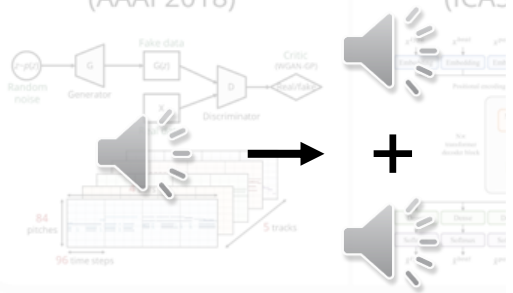



### CLIPSonic (WASPAA 2023)

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

**Text-queried sound separation**

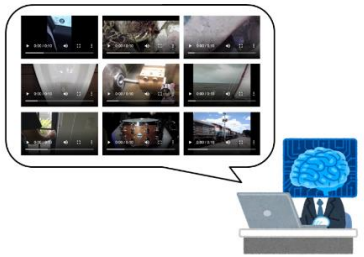
Query:
"*playing harpsichord*"



# Generative AI for Music & Audio

## Multimodal Learning for Audio & Music

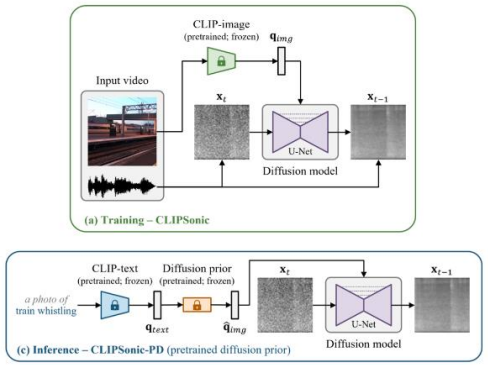
Learning sound separation and synthesis from videos



### CLIPSep (ICLR 2023)

First text-to-sound synthesis model that can be trained **using only unlabeled videos**

### CLIPSonic (WASPAA 2023)




Text-to-audio synthesis

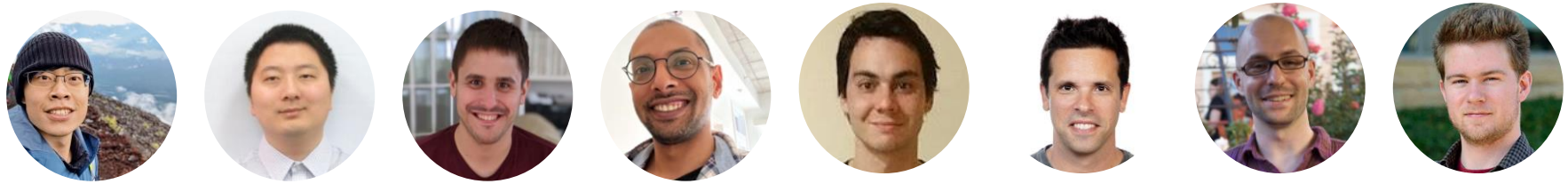
IEEE
WASPAA
2023

# 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

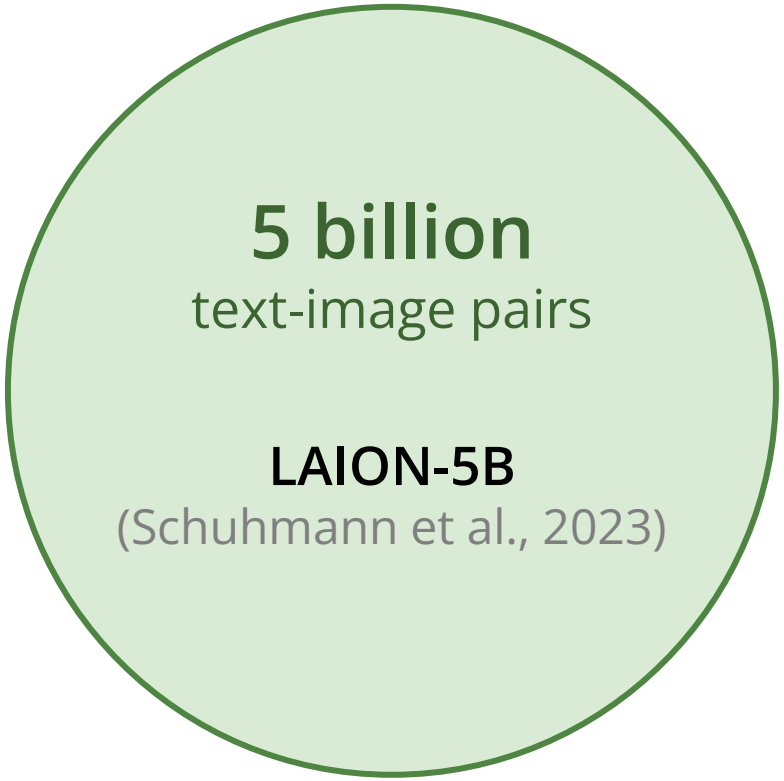
# What is Text-to-Audio Synthesis?

- Goal: Given a text query, generate the corresponding sounds

# Why NOT Text-audio Pairs?

**YouTube videos!**

500 hours of videos uploaded per minute



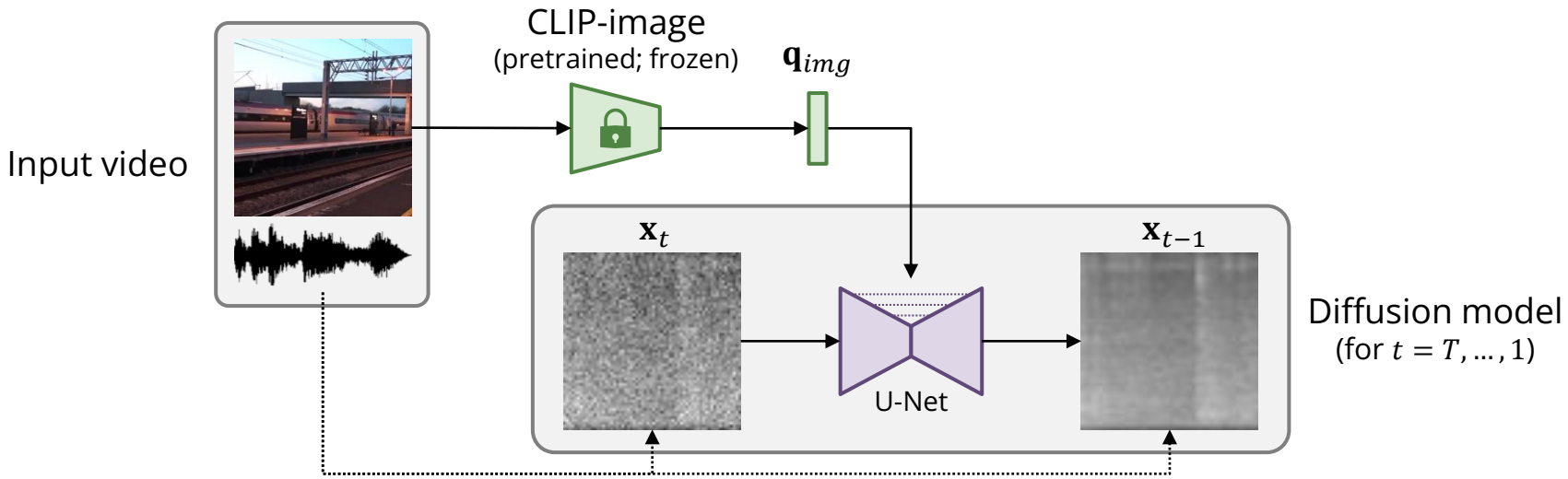

**0.6 million**
text-audio pairs

**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.

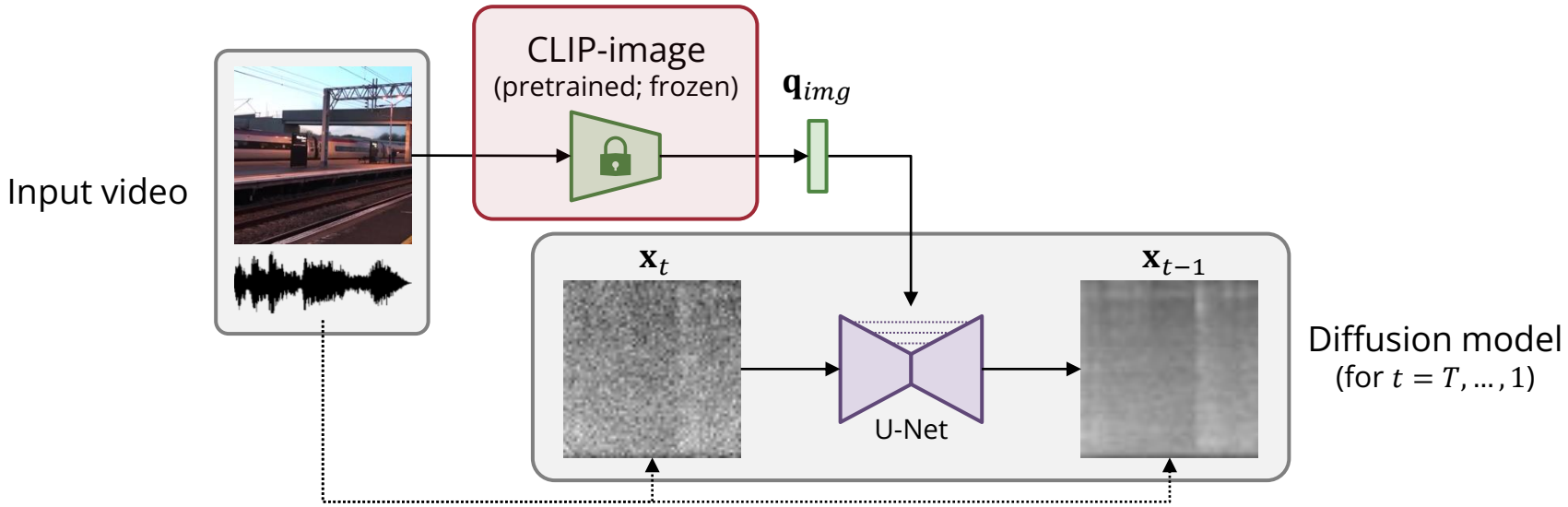
# Training an Image-to-Audio Synthesis Model

- We start by training an image-to-audio synthesis model



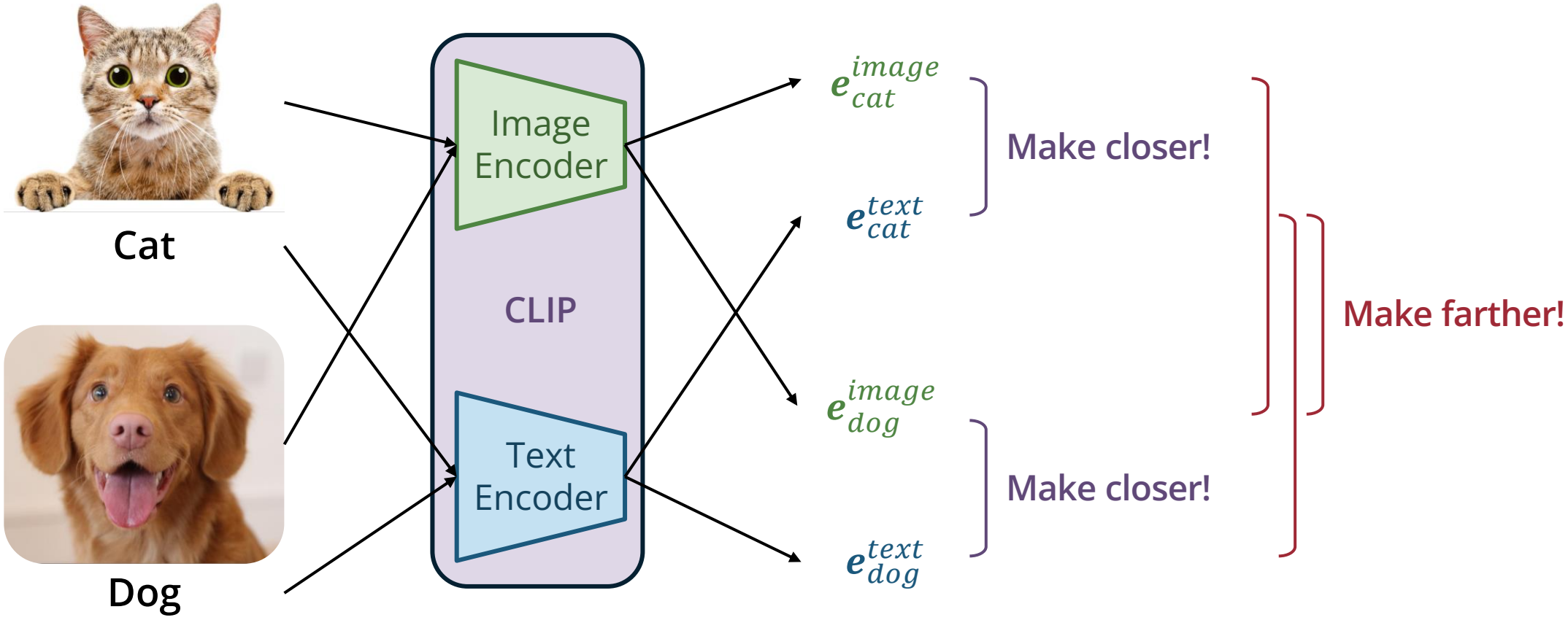
# Training an Image-to-Audio Synthesis Model

- We start by training an image-to-audio synthesis model



# CLIP (Contrastive Language-Image Pretraining)

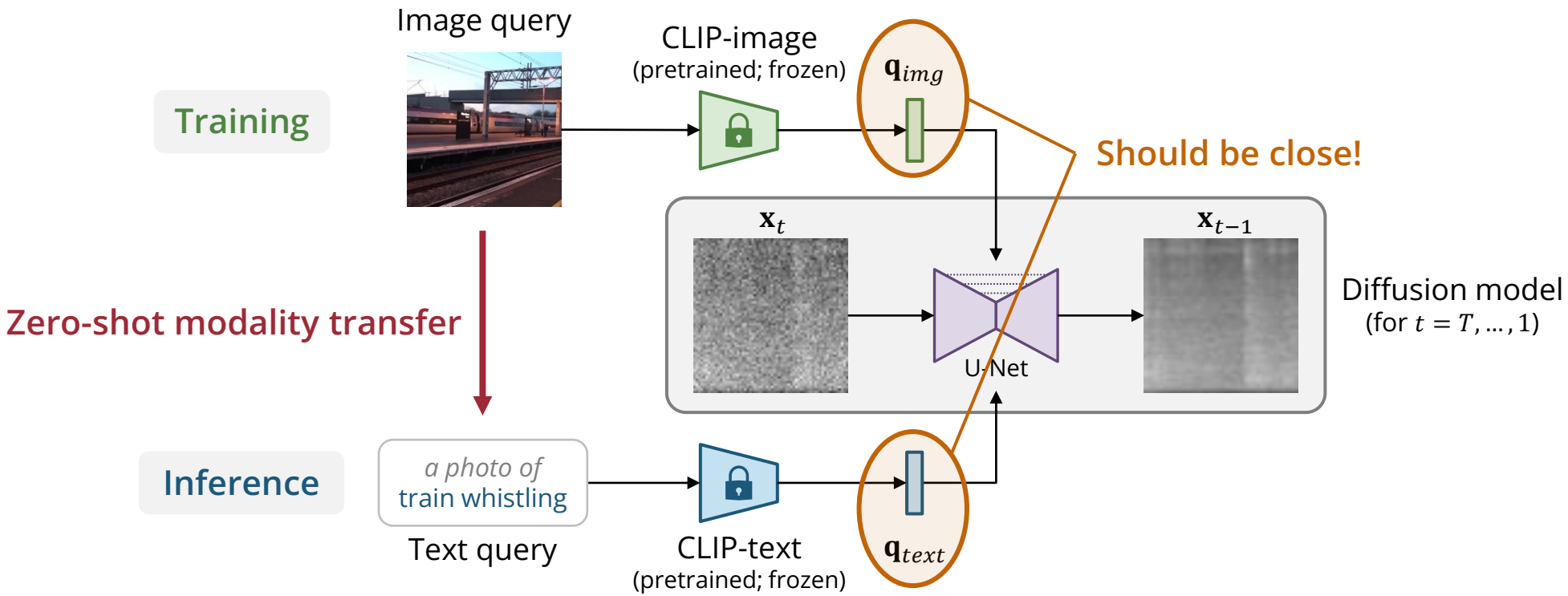
- Learn a shared embedding space for images and texts via *contrastive learning*




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

# Inference – Zero-shot Modality Transfer

- We switch to a pretrained CLIP-text encoder for text-to-sound synthesis



# Inference – Zero-shot Modality Transfer

- We switch to a pretrained CLIP-text encoder for text-to-sound synthesis

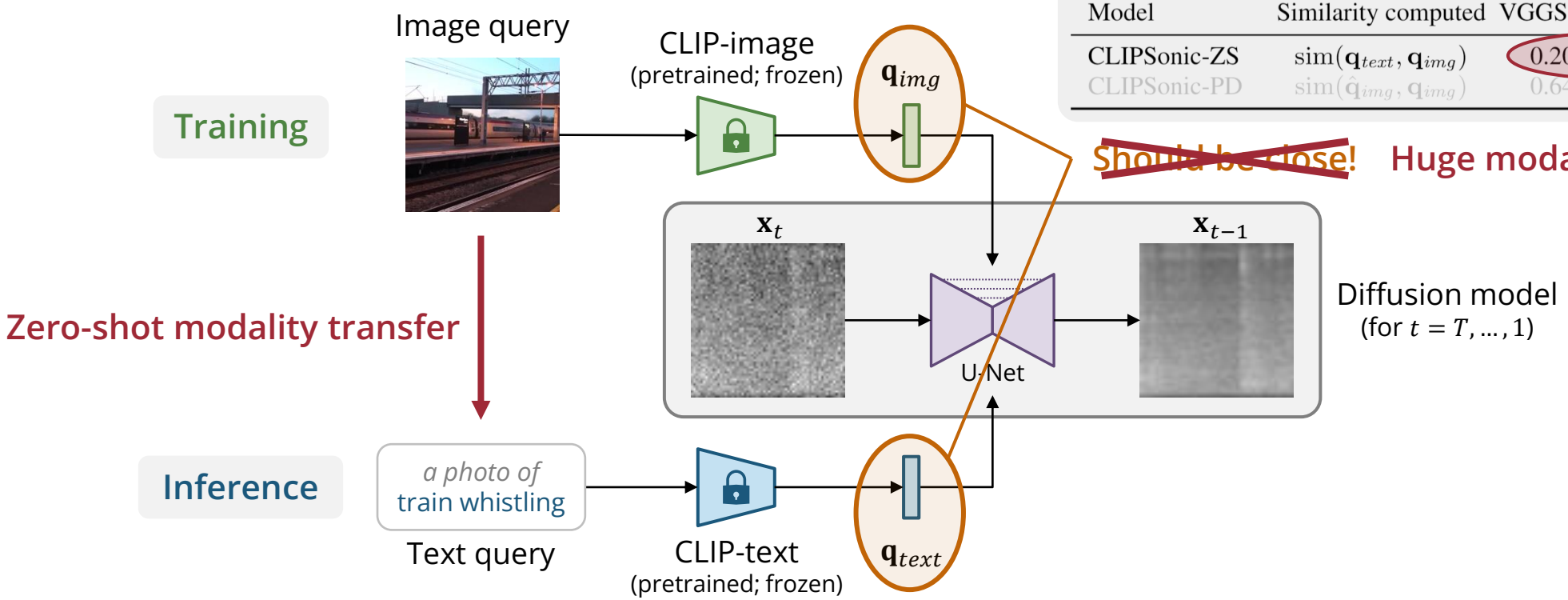


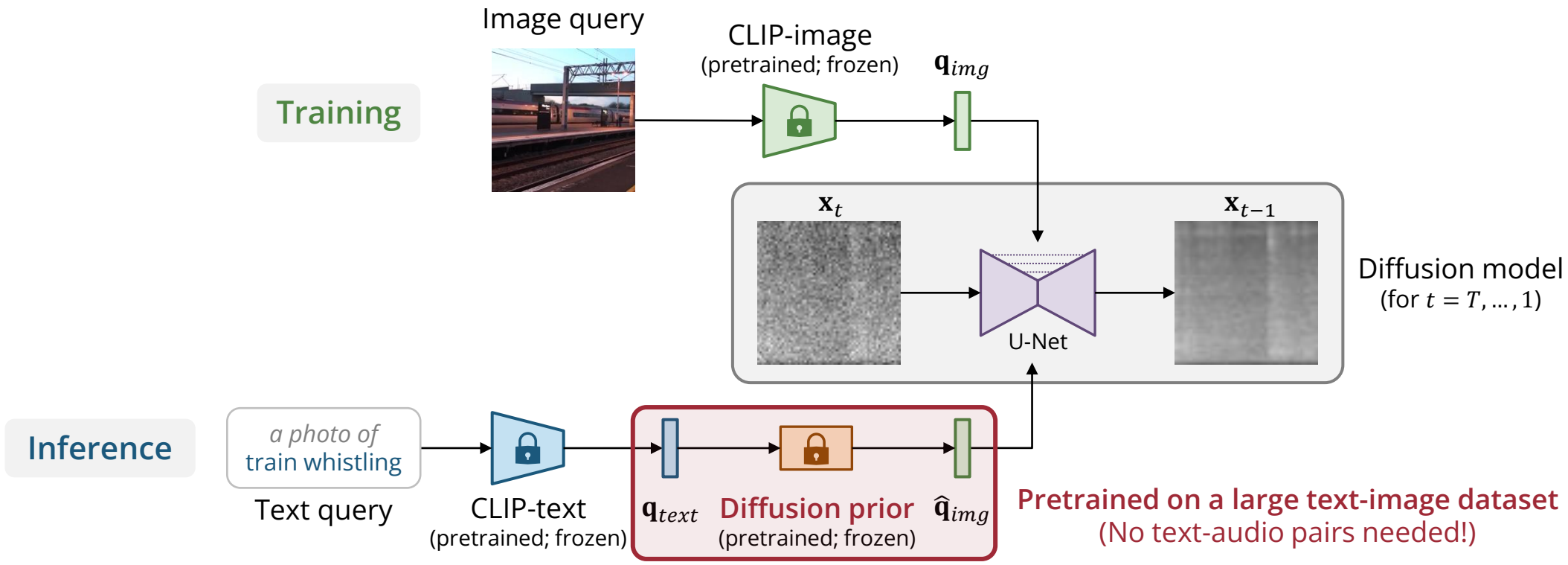

Table 2: Cosine similarities between various query embeddings.

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

~~Should be close!~~ Huge modality gap!

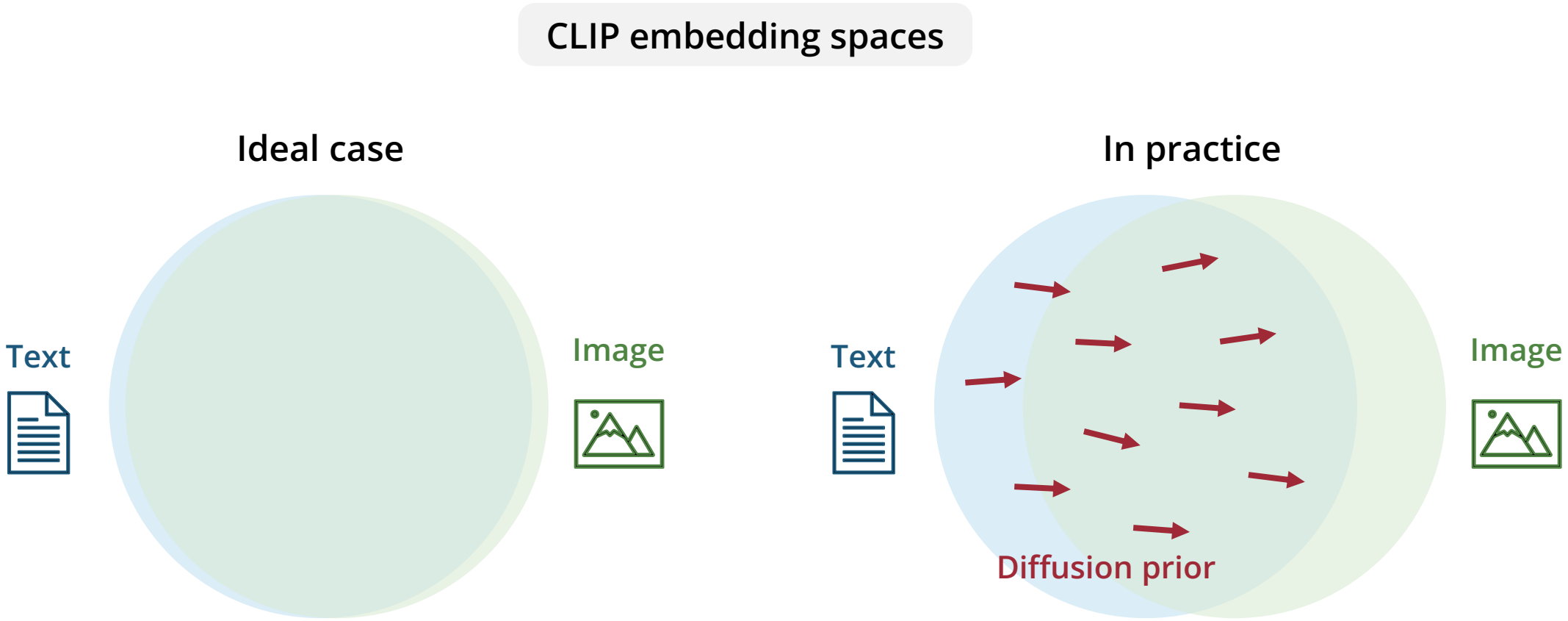
# Leveraging Diffusion Prior to Close the Modality Gap

- We adopt a pretrained diffusion prior model to reduce the modality gap



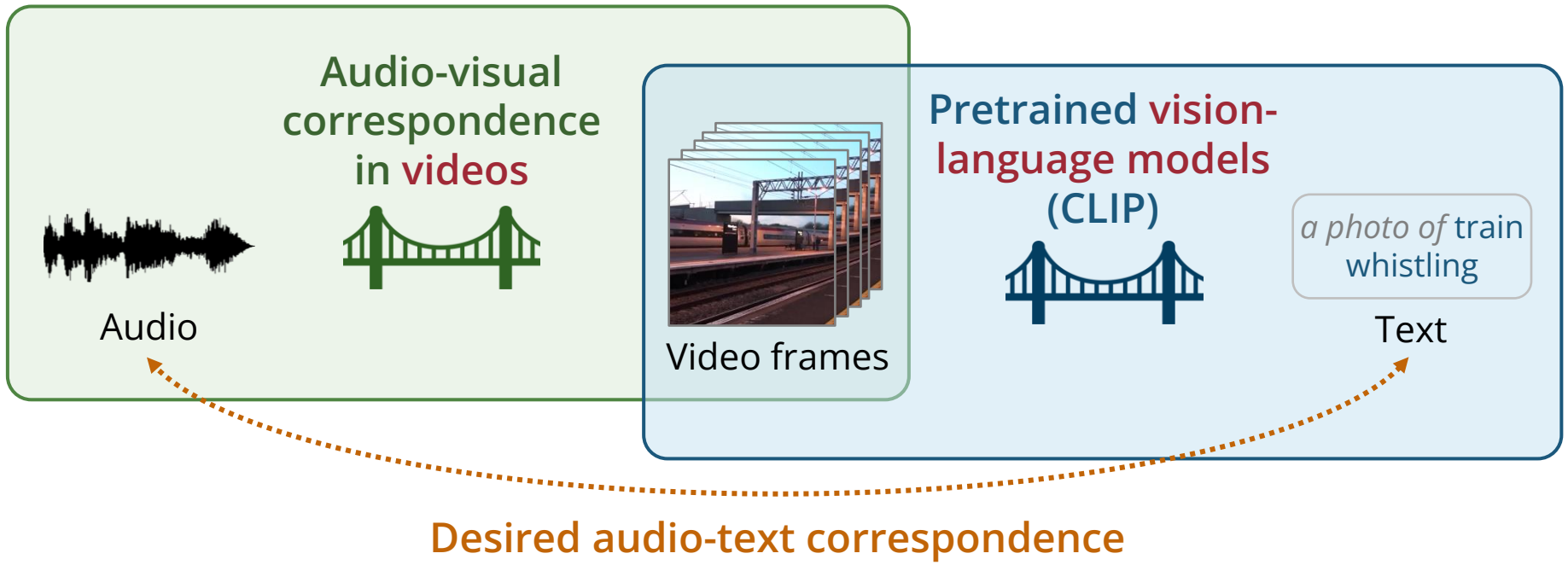

Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint arXiv:2204.06125*, 2022.

# Diffusion Prior (Ramesh et al., 2022)




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

# Leveraging the Visual Domain as a Bridge




No text-audio pairs required!

Scalable to large video datasets!

# 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.

# Example Text-to-Audio Synthesis Results

Rapping

Sea waves

Thunder

Smoke detector beeping

Playing table tennis

Playing violin fiddle

# Example Image-to-Audio Synthesis Results (Out-of-distribution)

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

# Subjective & Objective Evaluation Results

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

| Model | VGGSound | | MUSIC | |
|---|---|---|---|---|
| | Fidelity | Relevance | Fidelity | Relevance |
| CLIPSonic-ZS | $2.55 \pm 0.22$ | $2.01 \pm 0.27$ | $2.98 \pm 0.23$ | $3.87 \pm 0.24$ |
| CLIPSonic-PD | $\mathbf{3.04} \pm \mathbf{0.20}$ | $2.86 \pm 0.25$ | $\mathbf{3.67} \pm \mathbf{0.18}$ | $3.91 \pm 0.24$ |
| Ground truth | $3.78 \pm 0.19$ | $3.54 \pm 0.29$ | $3.90 \pm 0.17$ | $4.34 \pm 0.18$ |

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

| Model | Fidelity | Relevance |
|---|---|---|
| CLIPSonic-IQ (image-queried) | $\mathbf{3.29} \pm \mathbf{0.16}$ | $3.80 \pm 0.19$ |
| SpecVQGAN [20] | $2.15 \pm 0.17$ | $2.54 \pm 0.23$ |
| im2wav [21] | $2.19 \pm 0.15$ | $\mathbf{3.90} \pm \mathbf{0.22}$ |

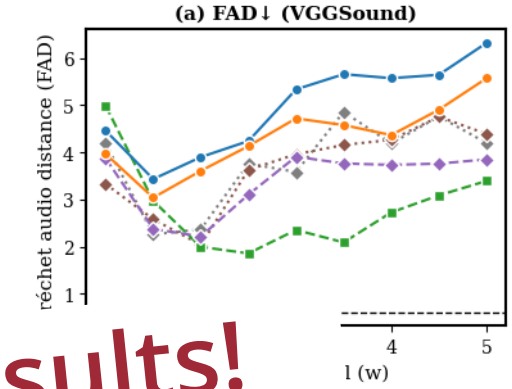
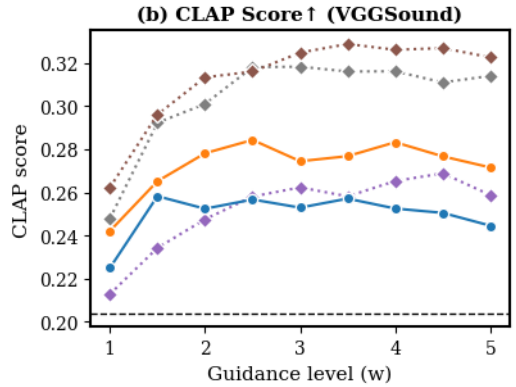



Table 1: Evaluation results on VGGSound and MUSIC

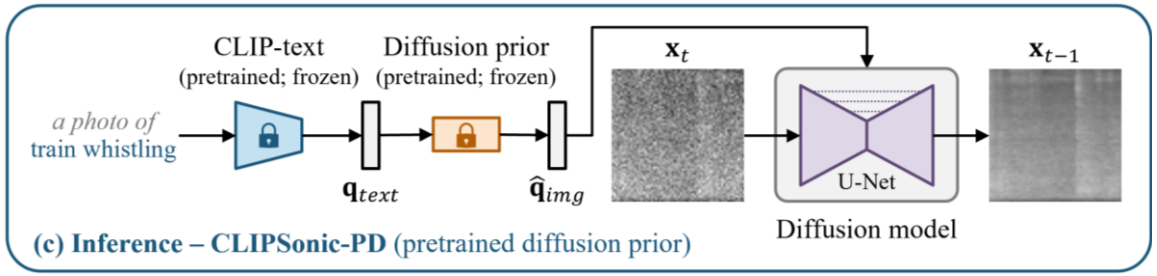
| Model | | | | VGGSound | | MUSIC | |
|---|---|---|---|---|---|---|---|
| | | | | FAD ↓ | CLAP score ↑ | FAD ↓ | CLAP score ↑ |
| CLIPSonic-IQ | - | Image | Image | 2.97 | - | 4.71 | - |
| CLIPSonic-ZS (zero-shot modality transfer) | ✓ | Image | Text | 3.43 | 0.258 | 19.30 | 0.284 |
| CLIPSonic-PD (pretrained 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 | 0.299 |
| CLIP-TTA | ✗ | Text | Text | 2.26 | 0.292 | 9.39 | 0.298 |
| CLAP-TTA | ✗ | Text | Text | 2.58 | 0.296 | 10.92 | 0.303 |
| BigVGAN mel spectrogram reconstruction | - | - | - | 0.60 | 0.204 | 6.21 | 0.272 |

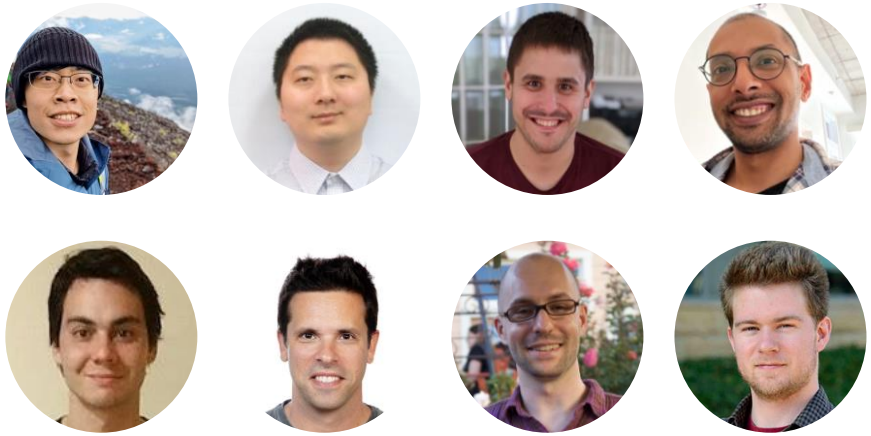



**Check out our paper for more results!**

# Summary

- First text-to-audio synthesis model that requires *no* text-audio pairs
- Strong text-to-audio synthesis performance without text-audio data
- State-of-the-art image-to-audio synthesis performance



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

# What’s Next?

**Video → Music & sound effects**
**Text → Video with music & sound effects**

OpenAI, Sora: Creating video from text, https://openai.com/sora, Feb 15, 2024.

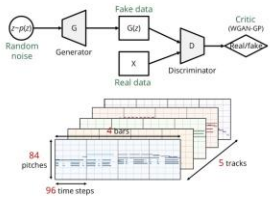
# Generative AI for Music & Audio

*Empowering music and audio creation with machine learning*

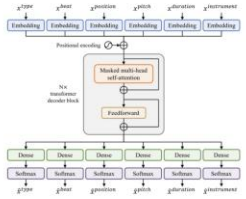
## Multitrack Music Generation

Advancing deep generative models for multitrack music

### MuseGAN (AAAI 2018)



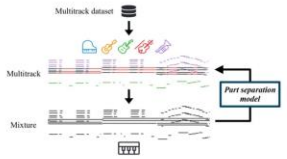

### MMT (ICASSP 2023)



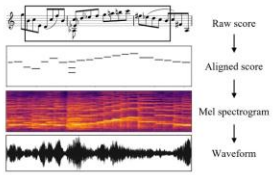
## Assistive Music Creation Tools

Developing AI-augmented assistive music creation tools

### Arranger (ISMIR 2021)



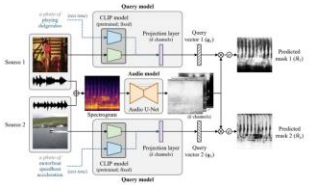

### Deep Performer (ICASSP 2022)



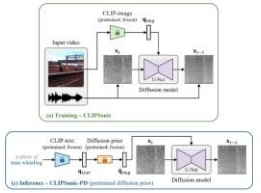

## Multimodal Learning for Audio & Music

Learning sound separation and synthesis from videos

### CLIPSep (ICLR 2023)



### CLIPSonic (WASPAA 2023)



UC San Diego · 中央研究院 ACADEMIA SINICA · Dolby · SONY · amazon · Ministry of Education 教育部 · erc