AudioLDM: Text-to-Audio Generation with Latent Diffusion Models

Haohe Liu*, Zehua Chen*, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, Mark D. Plumley

UNIVERSITY OF SURREY

Imperial College London

Authors

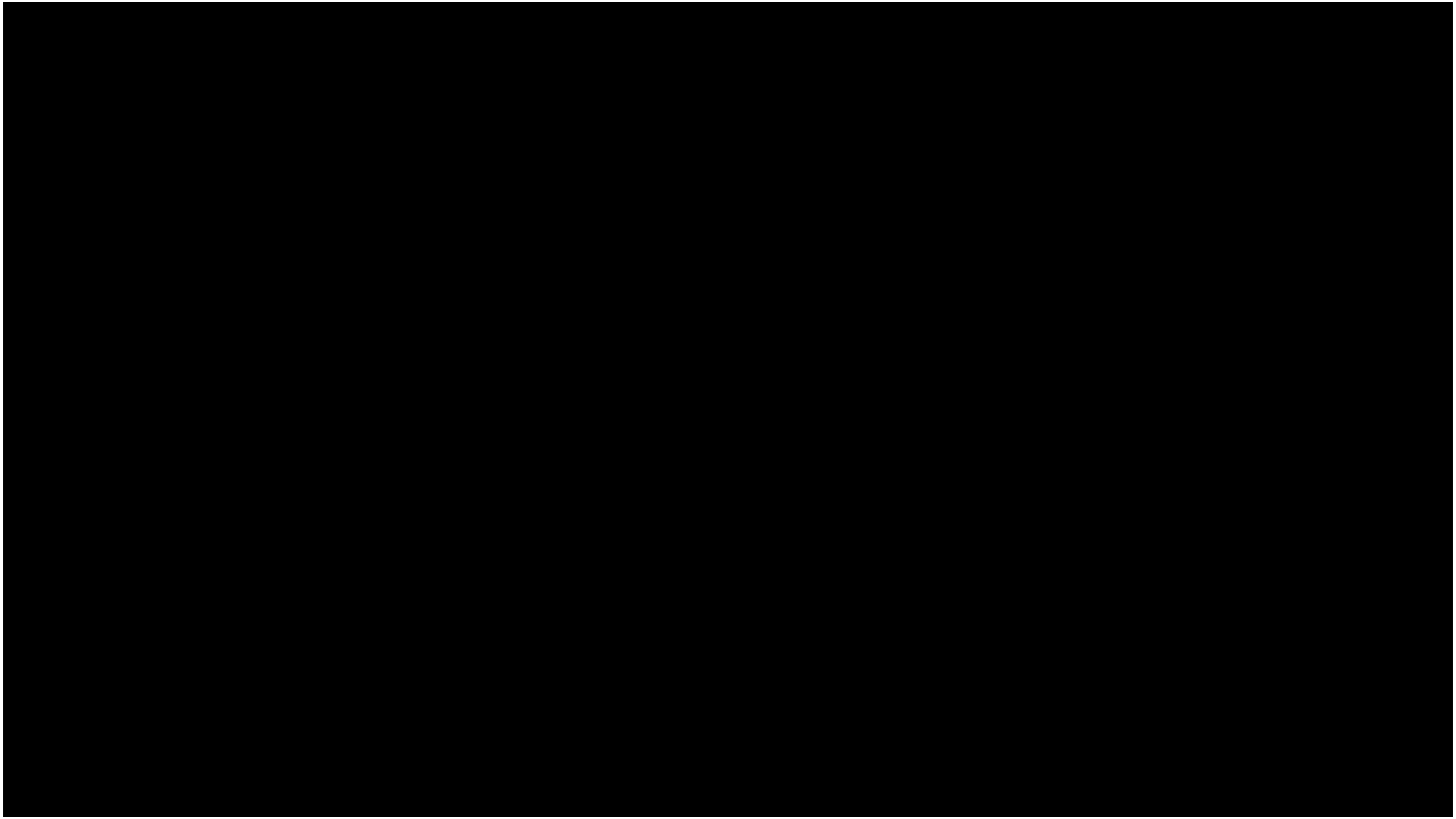
Haohe (Leo) Liu

University of Surrey, Guildford, UK
Centre for Vision, Speech and Signal Processing (CVSSP)
Supervisor: Prof. Mark D. Plumbley

Zehua Chen

Imperial College London, London, UK
Department of Electrical and Electronic Engineering
Supervisor: Prof. Danilo Mandic

- Many thanks to other co-authors who made this work possible:
  - Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, Mark D. Plumbley



What is Audio Generation

Definition, history, and related works

Audio Generation

- **The creation of sound through various ways**
- **The targets include:**
  - *Sound Effect* (Natural, Human-made objects, Animal, etc.)
  - *Speech* (Emotion, Pace, Gender, etc.)
  - *Music* (Genre, Rhythm, Instruments, etc.)
  - *Other* (Imaginary sound, compositional sound)

History of Sound Effect Creation

Jack Foley (1891-1967)
American sound effects artist

Foley Artist

Recreation of the realistic ambient sounds

Jack Foley

Modern foley art

Physical Modeling

Synthesis by modeling physical process

Generate sound based on shape, material, strength, and excitations.

Sound Effect Library

Digital collection of sound effect

Sound Ideas

BBC SFX

Freesound

…

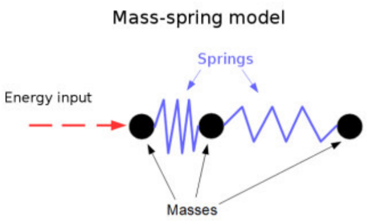
Sound Ideas released the Series 1000 (1979), which was the world's first fully digital sound effect library.

BBC Sound Effect Library is a large collection of sound effect

Freesound is a collaborative repository of CC licensed audio samples, and non-profit organization

Add live sound effects 1920s. .

Modern Foley Artist



The mass-string model

NESS Next Generation Sound Synthesis

Project from the University of Edinburgh

History of Speech Creation

Christian Gottlieb Kratzenstein (1723-1795)
Kratzenstein's resonators that can produce:
[a:], [e:], [i:], [o:] and [u:]

Machenical Synthesis

Simulating vocal tract, tongue, and lips

Kratzenstein Resonators

Kempelen's Speaking Machine

…

Electronic Signal Processing

Synthesis by modeling physical process

The VODER

Concatenation synthesis

Formant synthesis

Articulatory synthesis

…

Deep learning-based

Digital collection of sound effect

Tacotron

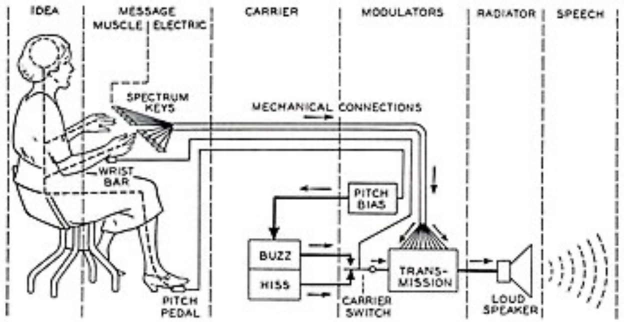
FastSpeech

NaturalSpeech

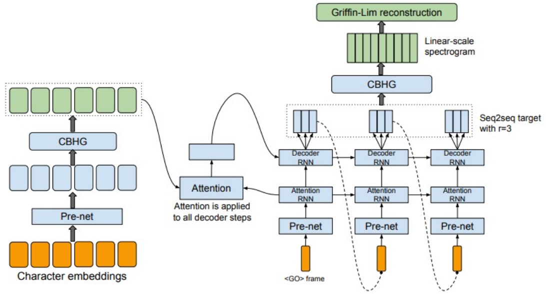
VALL-E

…

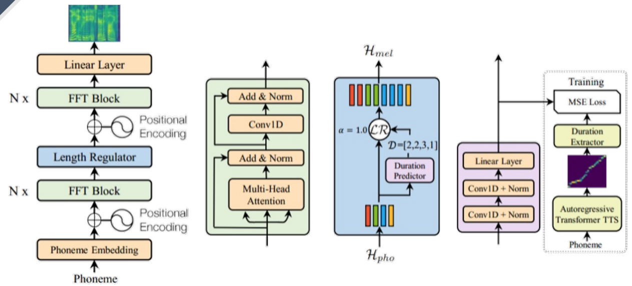
Kempelen's speaking machine (replica, 1837)



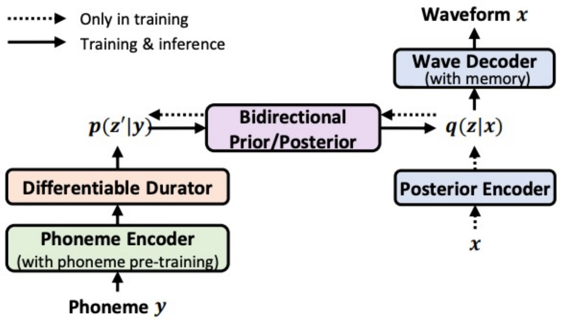
The Vocal Demonstrator (VODER, 1939)



Tacotron by Google (Wang et al., 2017)



FastSpeech by Microsoft (Ren et al., 2019)



NaturalSpeech by Microsoft (Tan et al., 2022)

University of Surrey
Imperial College London

History of Music Creation

Aurignacian flute
(43000 and 35000 years ago)

Musical Instrument

a device created or adapted to make musical sound

Aurignacian flute

Kempelen's Speaking Machine

…

Synthesizers, MIDI, and DAWs

Create music with electrical devices

Moog Synthesizer

Modern DAW

MIDI

…

Deep learning-based

Create music with machine learning

Synthesizer (DDSP)

Symbolic / MIDI (MuseNet)

Waveform (JukeBox)

Controllability (MIDI-DDSP)

…

MuseNet by OpenAI (2019)

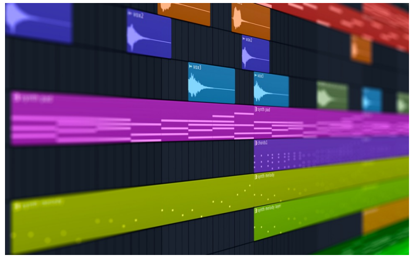
DDSP by Google (Engel et al., 2020)

MIDI-DDSP by Yusong et al. (2021)

JukeBox by OpenAI (Dhariwal, 2020)

Modern Musical Instruments

The Moog Synthesizer by Robert Moog (1970s)



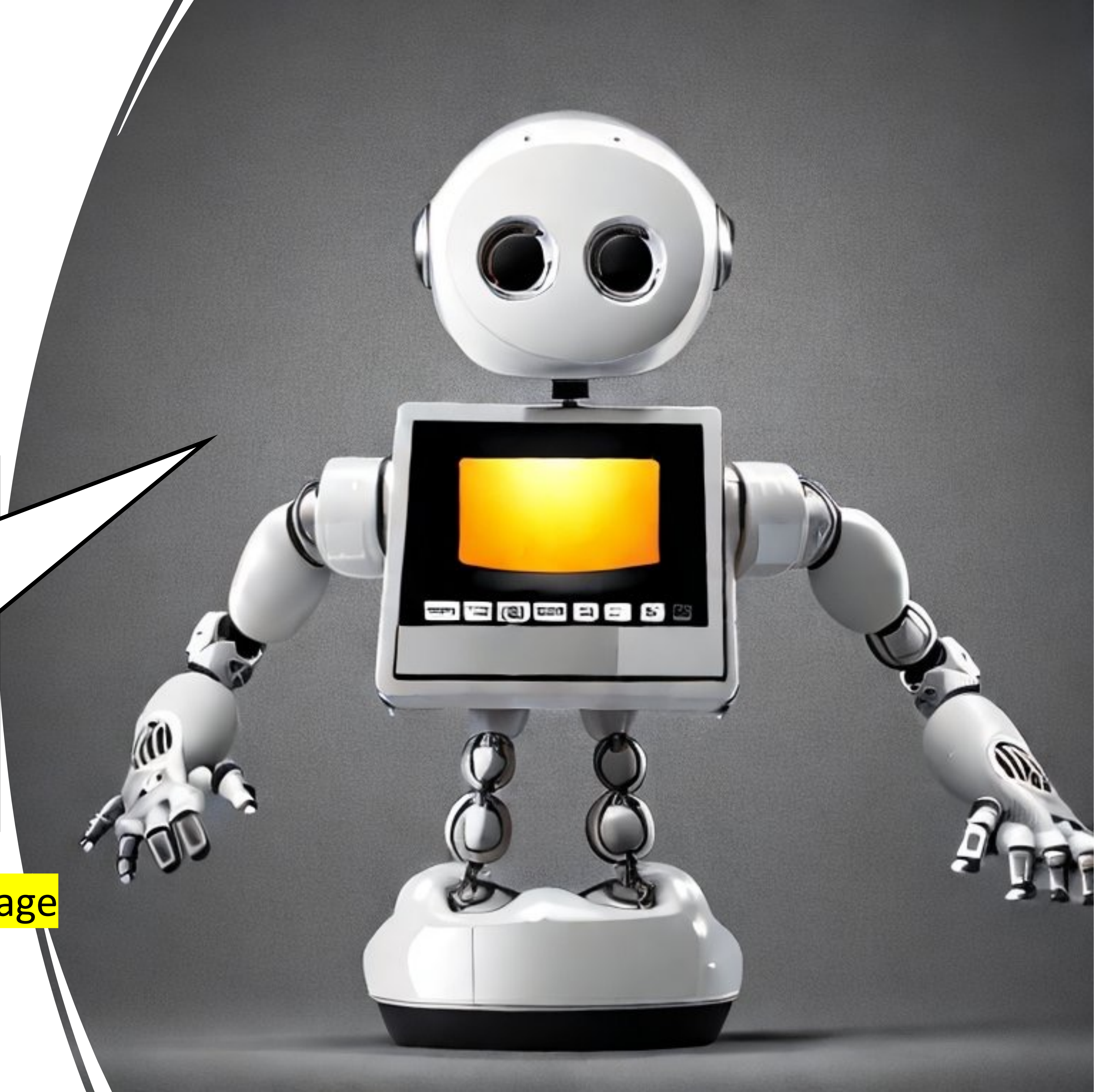
Digital Audio Workstation (DAW)

UNIVERSITY OF SURREY

Imperial College London

Can machine do general audio generation?

- **I'm a**
  - foley artist,
  - musical instruments performer,
  - oral broadcaster,
  - sound imaginer,
  - …



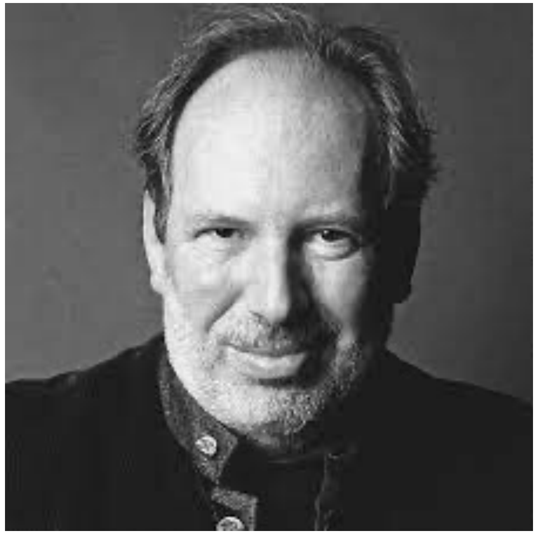
- Communicate with AI by natural language
  - Text-to-Audio Generation

Why: Text-to-Audio Generation

Applications, and motivations

Text-to-Audio Generation Usage Cases

- Computational “foley artist”: (e.g., https://www.thefoleybarn.com)
  - *Game developer: e.g., A ghost is haunting a house.*
  - *Audio producer: e.g., high heels hitting metal ground.*
  - *Movie producer: e.g., the laser sound from a laser gun.*
  - …
- Automatic content creation (> 60 startups[1])
  - Endless music
  - Audiobook with ambient noises
  - White noise for meditation
  - …
- Data Augmentations



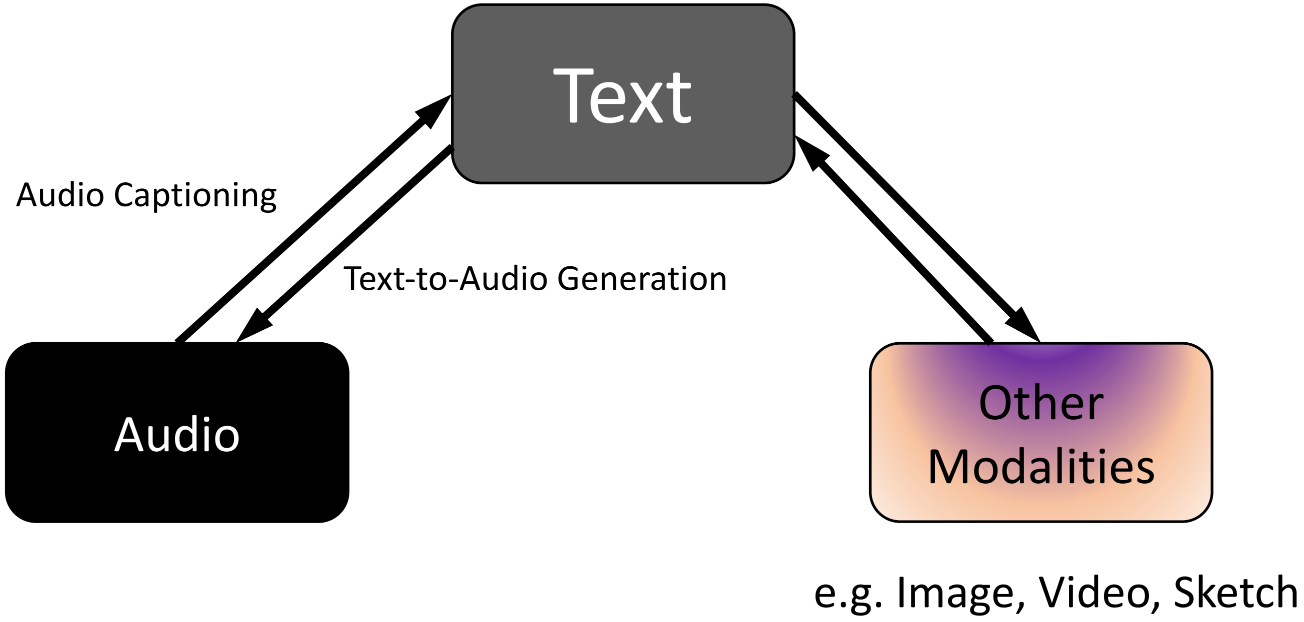
Sound is often the unsung hero of the movie world
- Hans zimmer

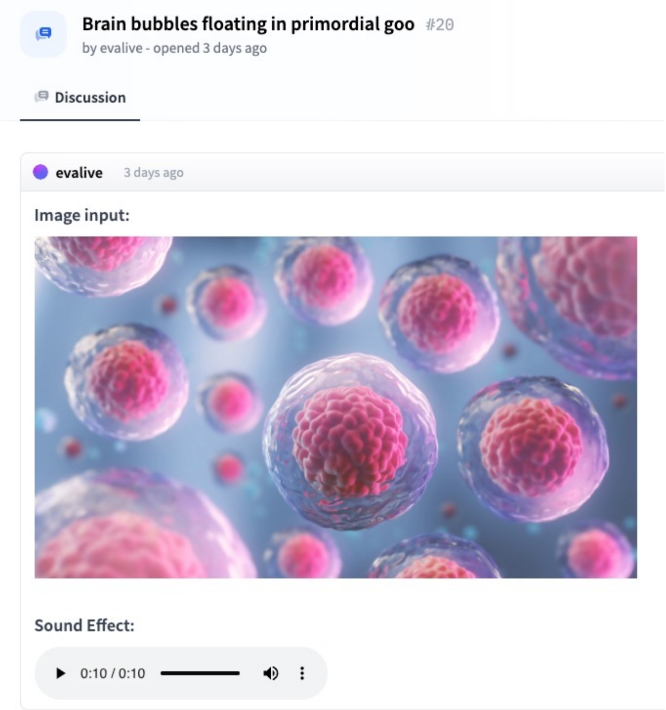
[1]https://github.com/csteinmetz1/ai-audio-startups

UNIVERSITY OF SURREY
Imperial College London

Text-to-Audio Generation Usage Cases

- Text is a bridge between audio and other modalities





[1]https://github.com/csteinmetz1/ai-audio-startups

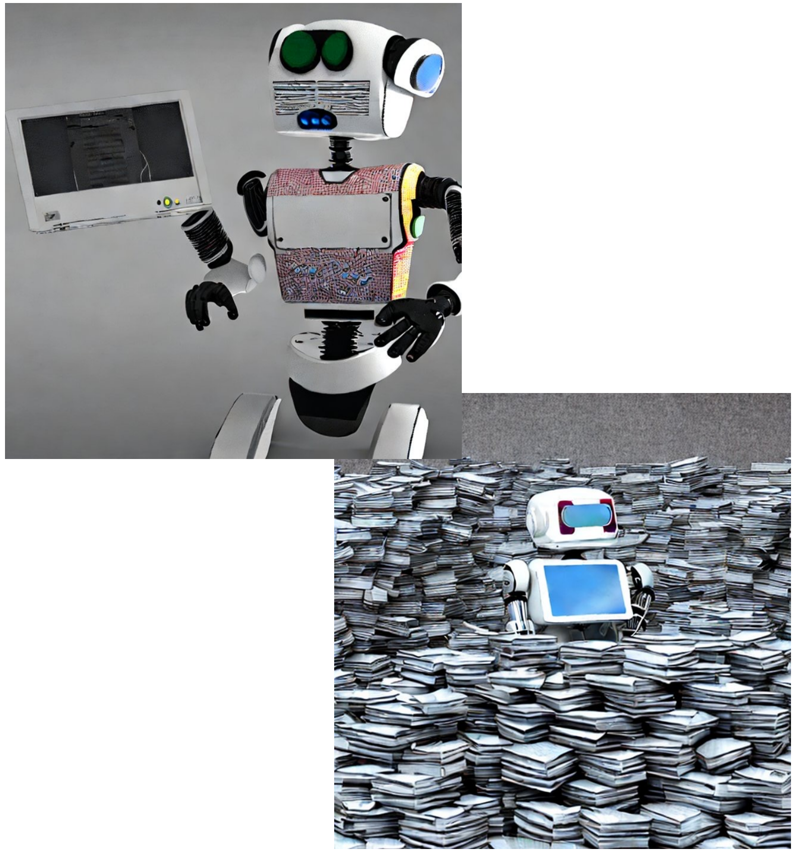
Generation VS Retrieval

Efficiency

- No need for retrieval
- Endless audio samples
- Fine-grained control on sound
  - Emotion, pitch, materials, etc.
- Future way of fuzzy data storage
  - 2GB VS 2048 GB

Creativity

- Generate non-exist sound
  - e.g., Half cat Half sheep sound
- Inspire the content creation



UNIVERSITY OF SURREY

Imperial College London

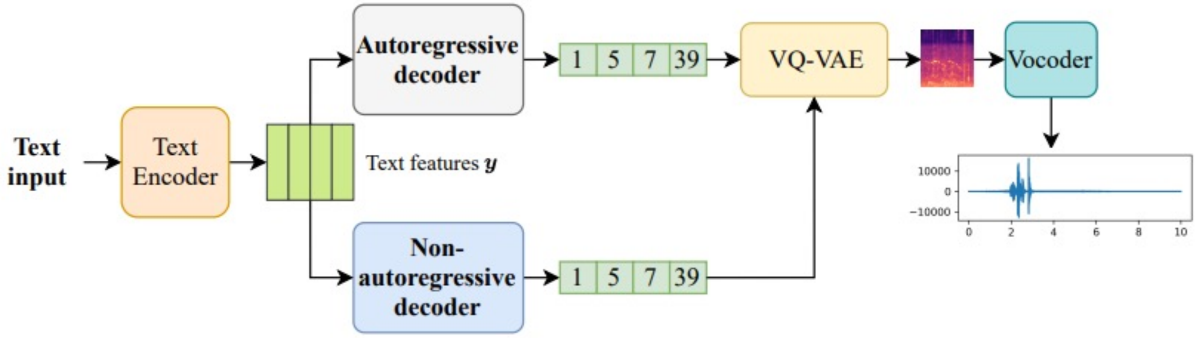
Related works

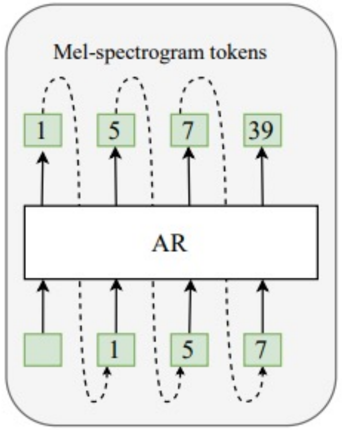
Introduction, and comparison

Related works

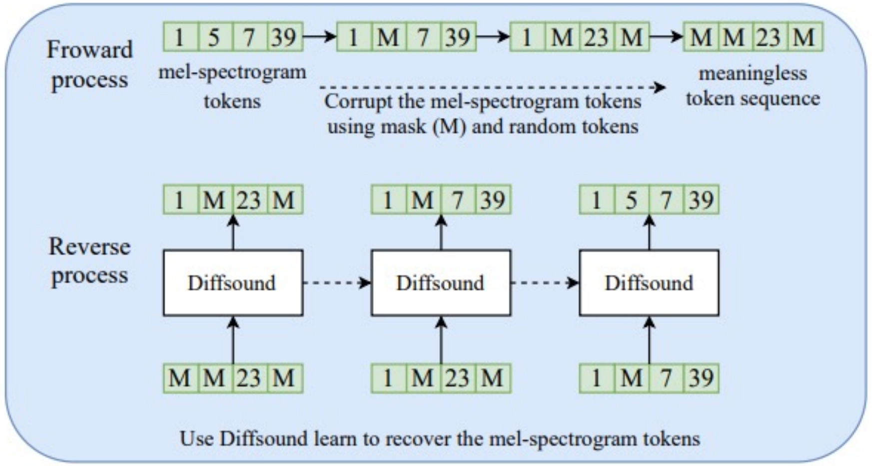
- **Label-to-Audio Generation**
  - Acoustic Scene (Kong et al., 2019), Sound event (Liu et al., 2019), FootStep (Comunit et al. 2019), …
- **Text-to-Audio Generation**
  - DiffSound (Yang et al., 2022), AudioGen (Kreuk et al., 2022), Make-an-Audio (Huang et al., 2023)
- **Text-to-Music Generation**
  - MusicLM (Andrea et al., 2023)
  - Moûsai (Flavio et al., 2023)
  - Noise2Music (Huang et al., 2023)
- **Others**
  - JukeBox (Dhariwal et al., 2020), AudioLM (Borsos et al., 2022), SingSong (Donahue et al., 2023),…

DiffSound (Yang et al., 2022)

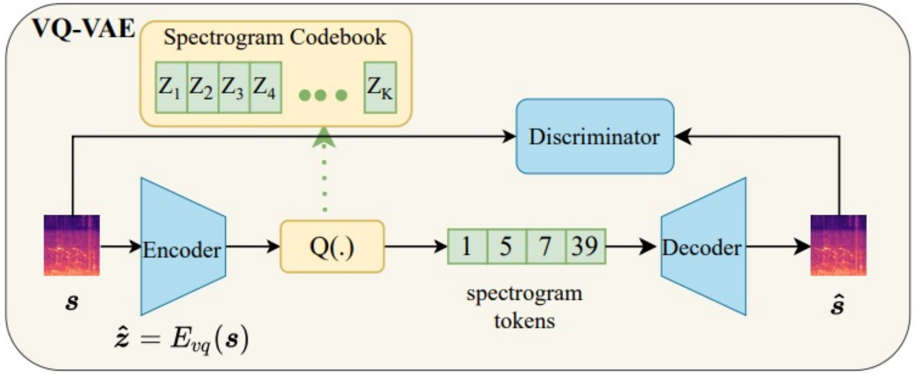




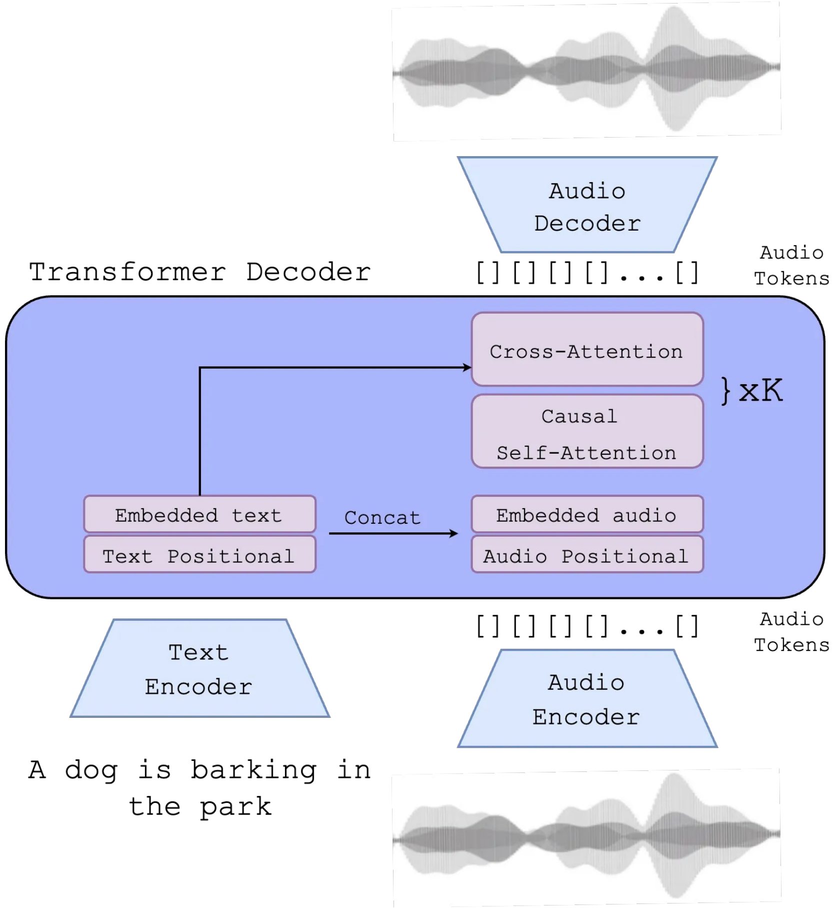
An example of autoregressive spectrogram tokens generation

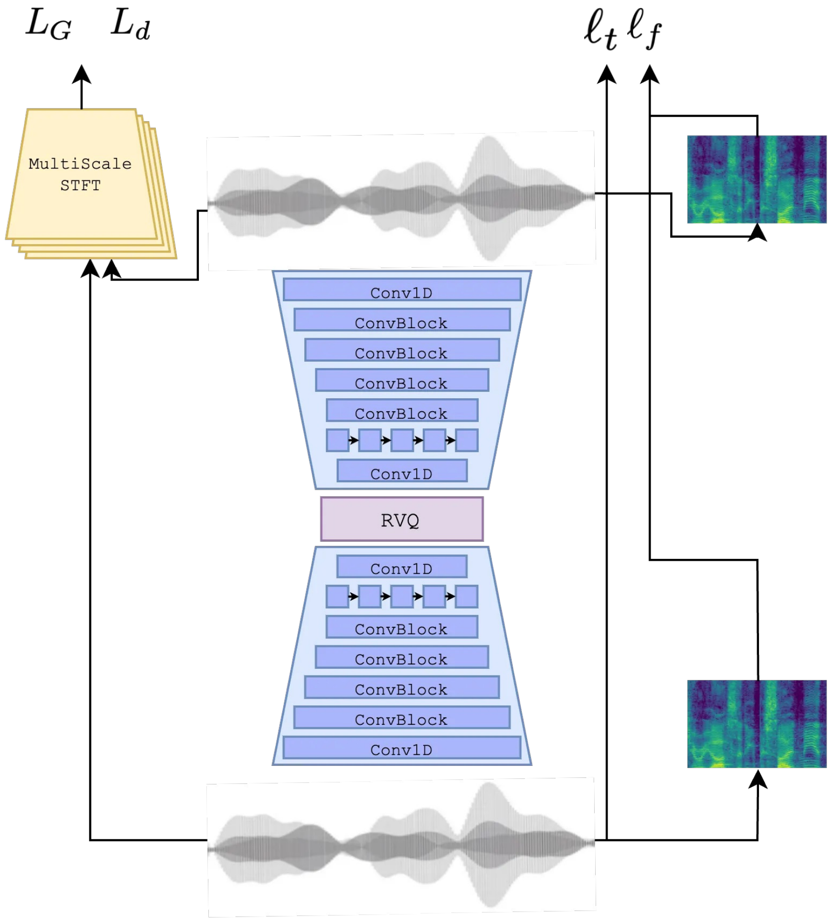


An example of non-autoregressive spectrogram tokens generation.

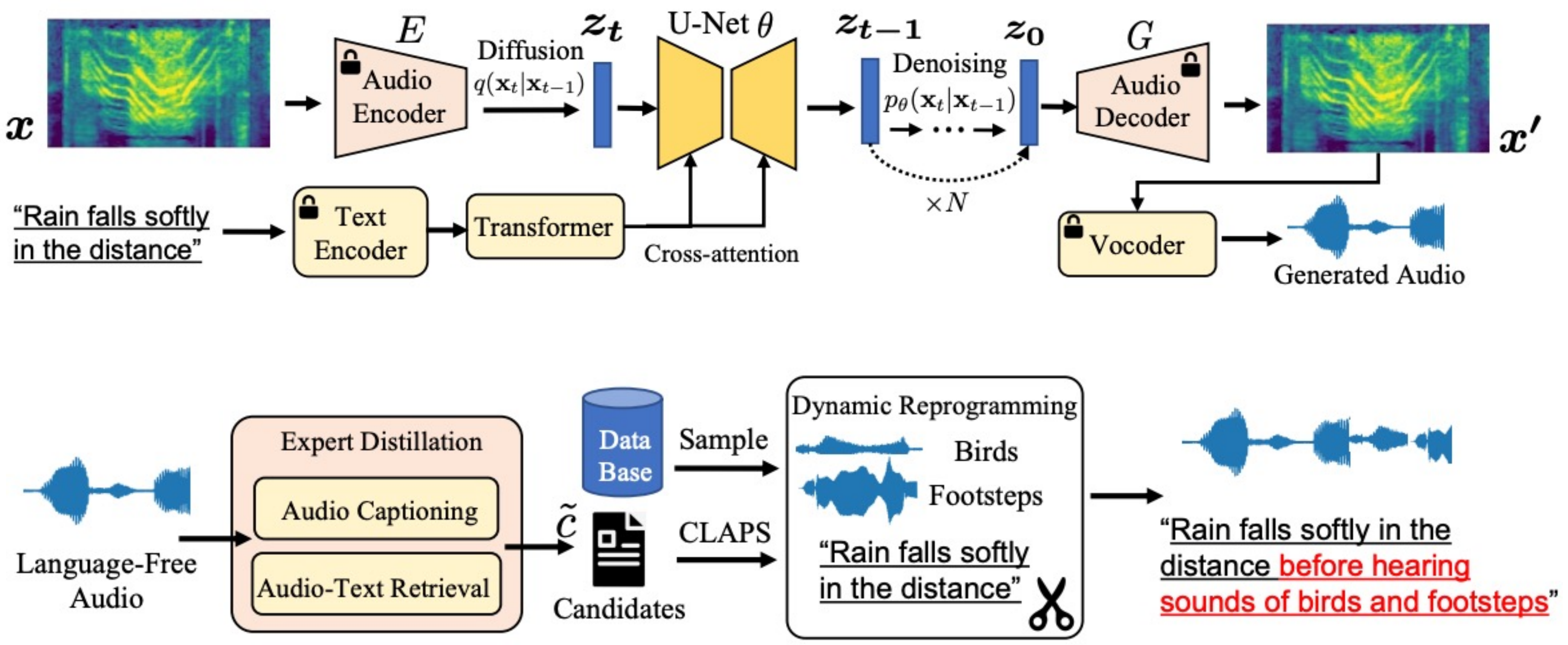


AudioGen (Kreuk et al., 2022)





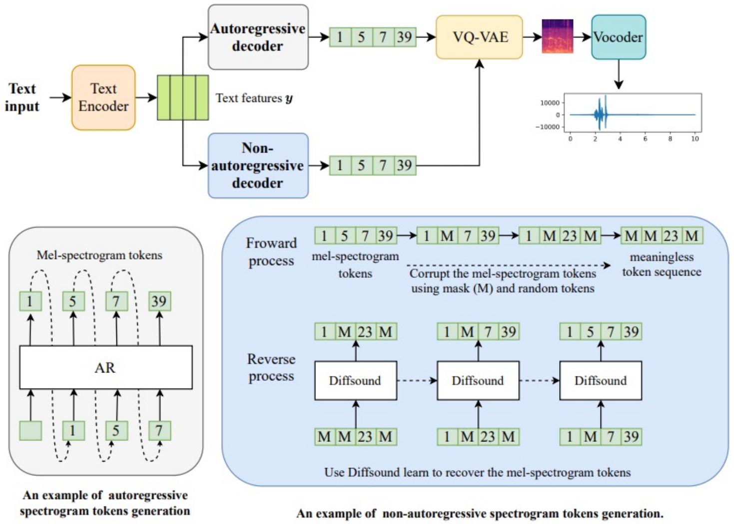
Make-an-Audio (Huang et al., 2023)



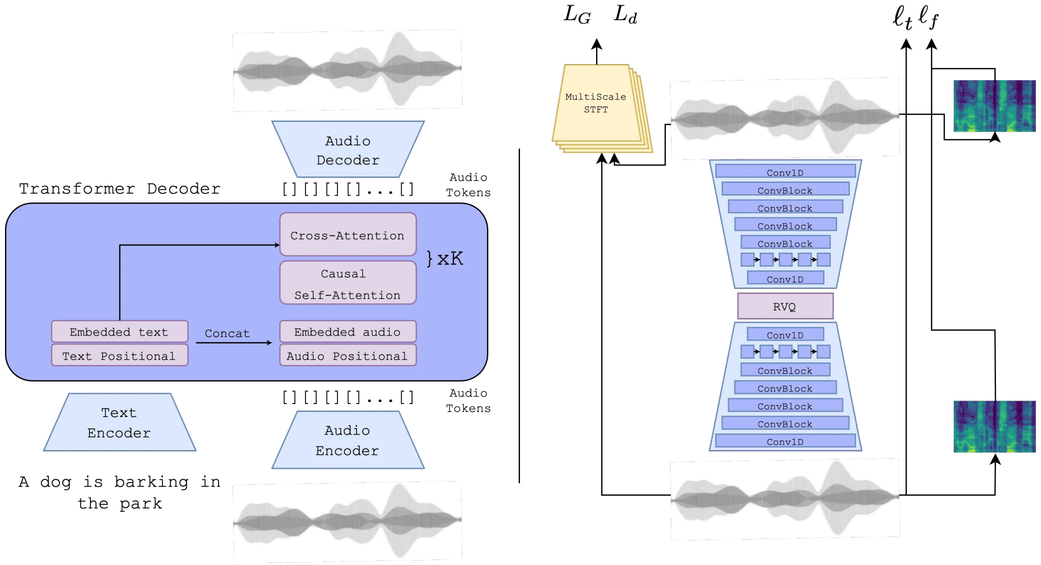
UNIVERSITY OF SURREY

Imperial College London

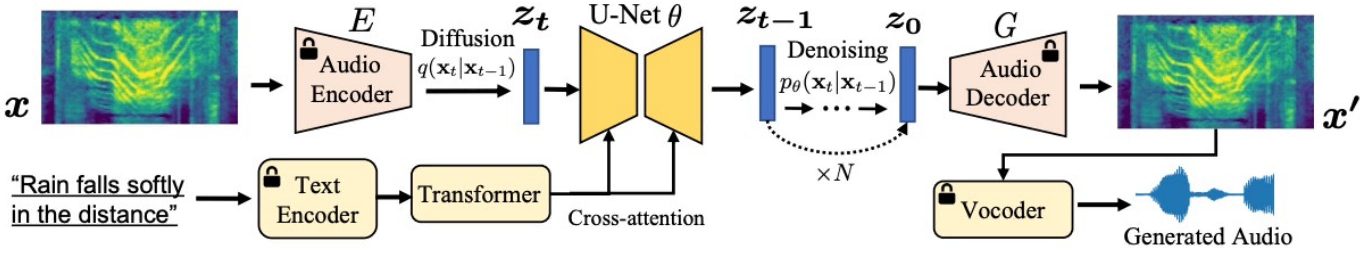
Related works



DiffSound (Yang et al., 2022)



AudioGen (Kreuk et al., 2022),



Make-an-Audio (Huang et al., 2023)

UNIVERSITY OF SURREY

Imperial College London

Comparison with previous studies

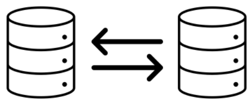
- Previous audio generation studies:
  - Requires large-scale audio-text pairs
    - Prev: Text → Audio → Loss → Backprop
    - Our: Audio → Audio → Loss → Backprop
  - High computational cost
    - Prev: 64 or 32 V100 GPUs (AudioGen, DiffSound)
    - Our: 1 GPUs
  - Limited generation quality and diversity.
  - Discrete latent space may limit model performance

Previous works:
10+ datasets, 800K audio-text pairs
(**still not enough**).

Self-supervised Learning for Audio Generation!

Self-supervised Audio Generation

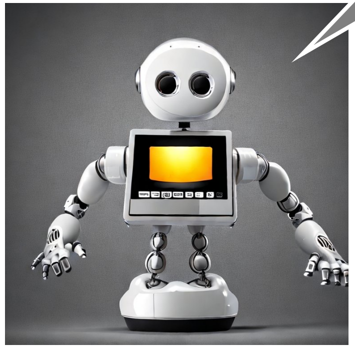
Step 1



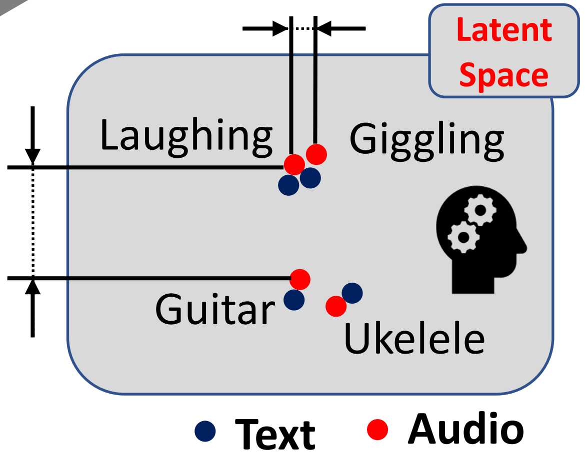
Audio Text

Human Developer :
Here are some audio-text pair,
try to figure out their relation!

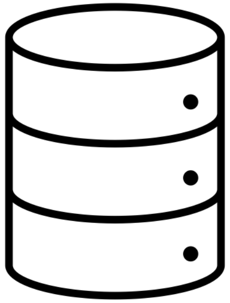
AI



OK Got it



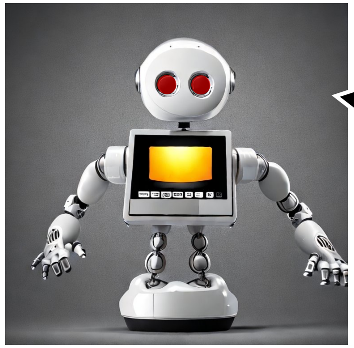
Step 2



Audio

Human Developer:
Here are more audio data,
Try to figure out how to generate them
using your knowledge!

AI



Nobody knows audio better than me!!!

Tell me what you want!

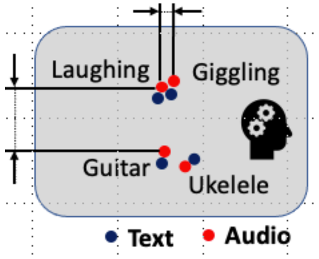
University of Surrey   Imperial College London

How: AudioLDM

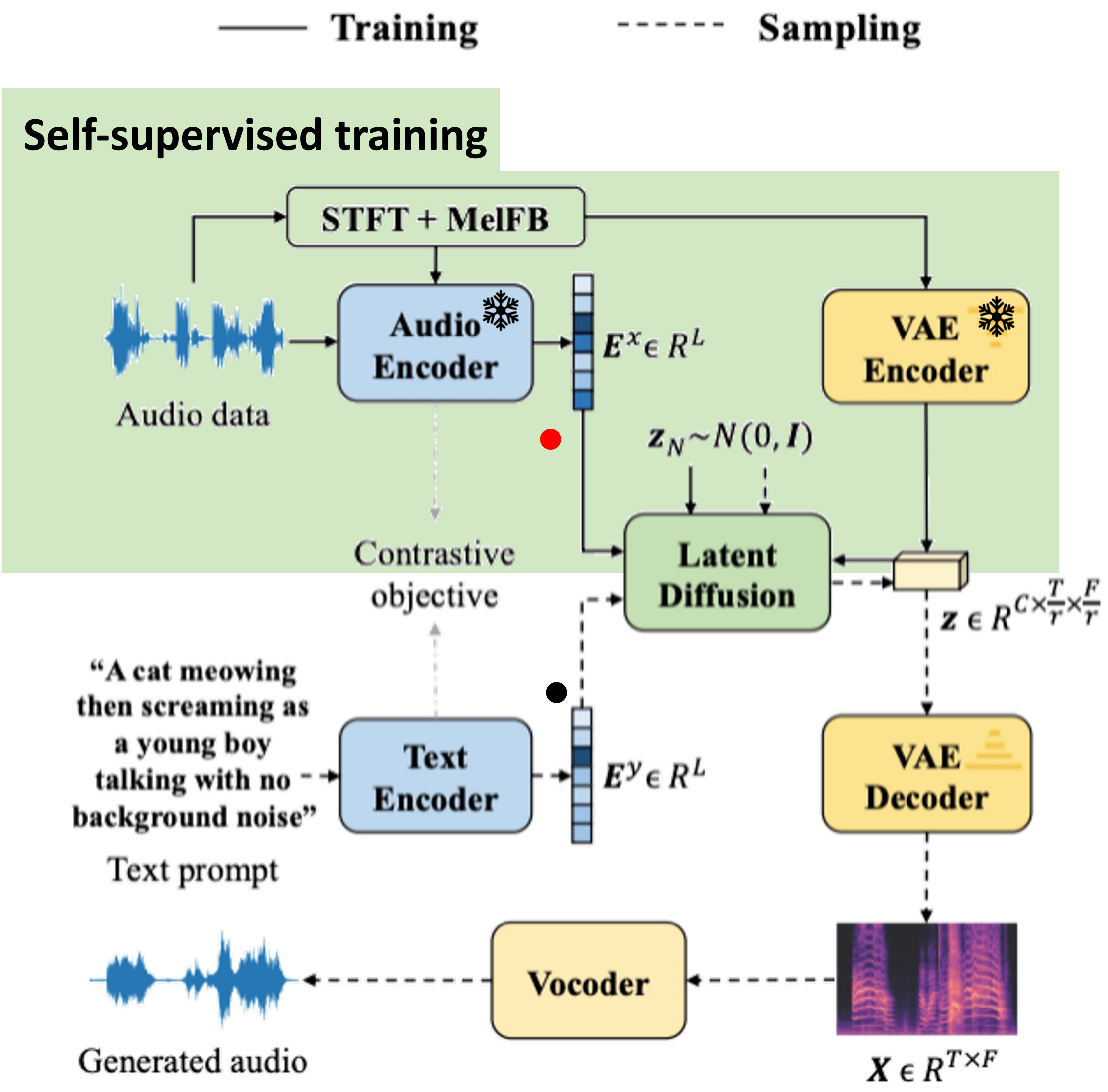
Methodology, Advantages, Experiment, and Result

UNIVERSITY OF SURREY
Imperial College London

AudioLDM

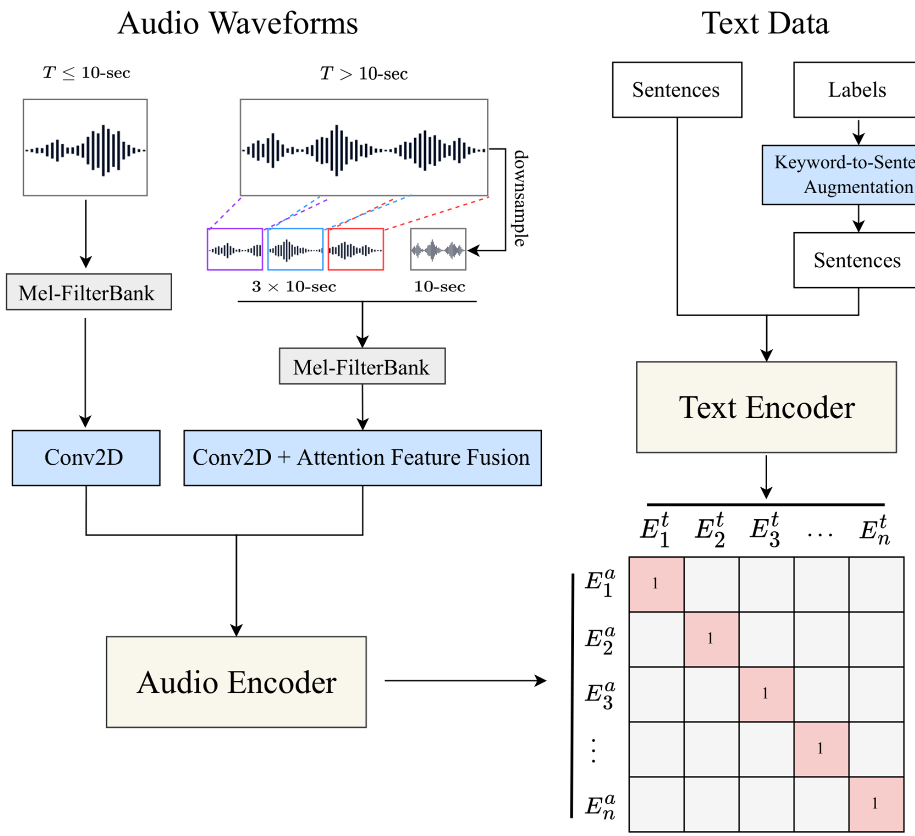


1. **Contrastive Language-Audio Learning (CLAP) Encoders**
   - Align audio and text in one space.
2. **Latent Diffusion Models**
   - Learn to generate VAE latent conditioned on CLAP embedding
3. **Mel-spectrogram Autoencoder**
   - Learn latent representations.
4. **Mel-to-Waveform Vocoder**
   - Reverse Mel back to waveform



UNIVERSITY OF SURREY
Imperial College London

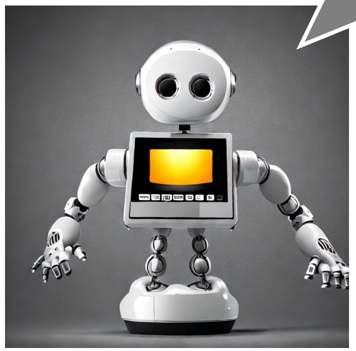
Step1: Contrastive Language-audio Pretraining

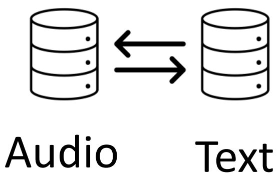


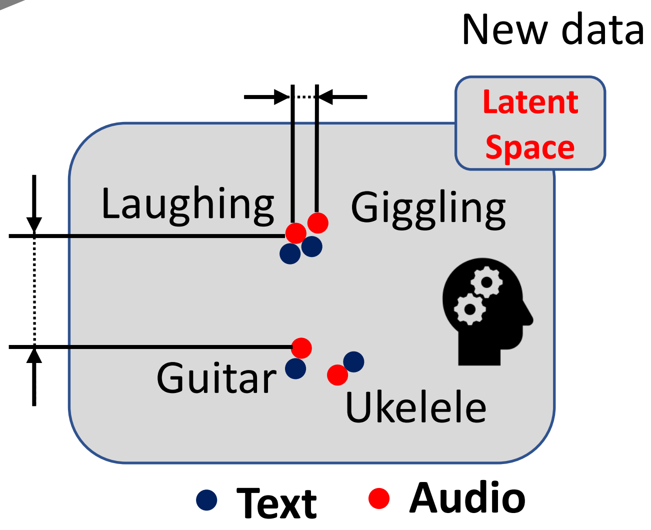
Contrastive Language-Audio Pretraining (Wu et al., 2022)

OK Got it
Laughing is similar to Giggle (Text).
Laughing also sounds like Giggle (Audio).
Laughing is **not** similar to Guitar (Text).
Laughing does **not** sound like Guitar (Audio)

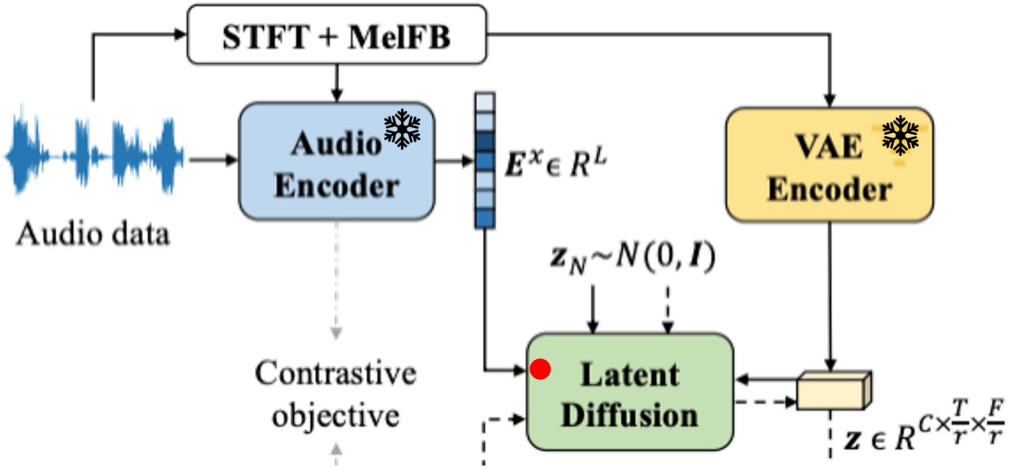
AI

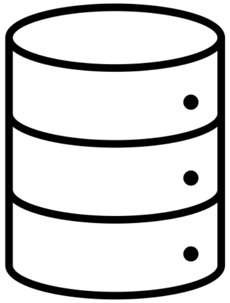






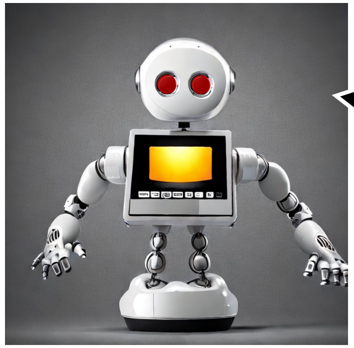
Step2: Self-supervised Audio Generation Training





Audio

Human Developer:
Here are more audio data,
Try to figure out how to generate them
using your knowledge!



Nobody knows audio better than me!!!

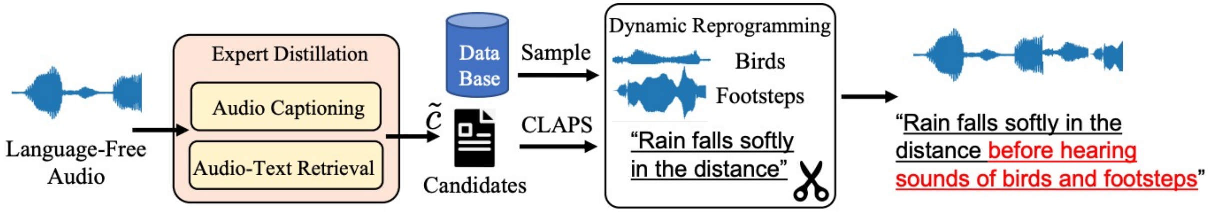
Tell me what you want!

UNIVERSITY OF SURREY

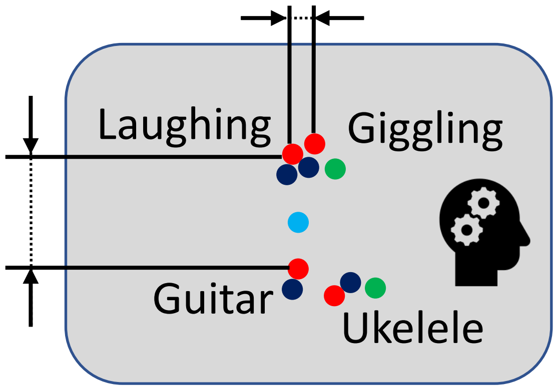
Imperial College London

Advantages of self-supervised training

- **Scale up training data easily!**
  - Collect Audio → Train model!
- **Perform data augmentation easily!**
  - .Previous works:
    - Mixup (Kreuk et al., 2022)
      - Text1 + Text2 → Audio1+Audio2
    - Pseudo prompt enhancement (Huang et al., 2023)



Make-an-Audio (Huang et al., 2023)

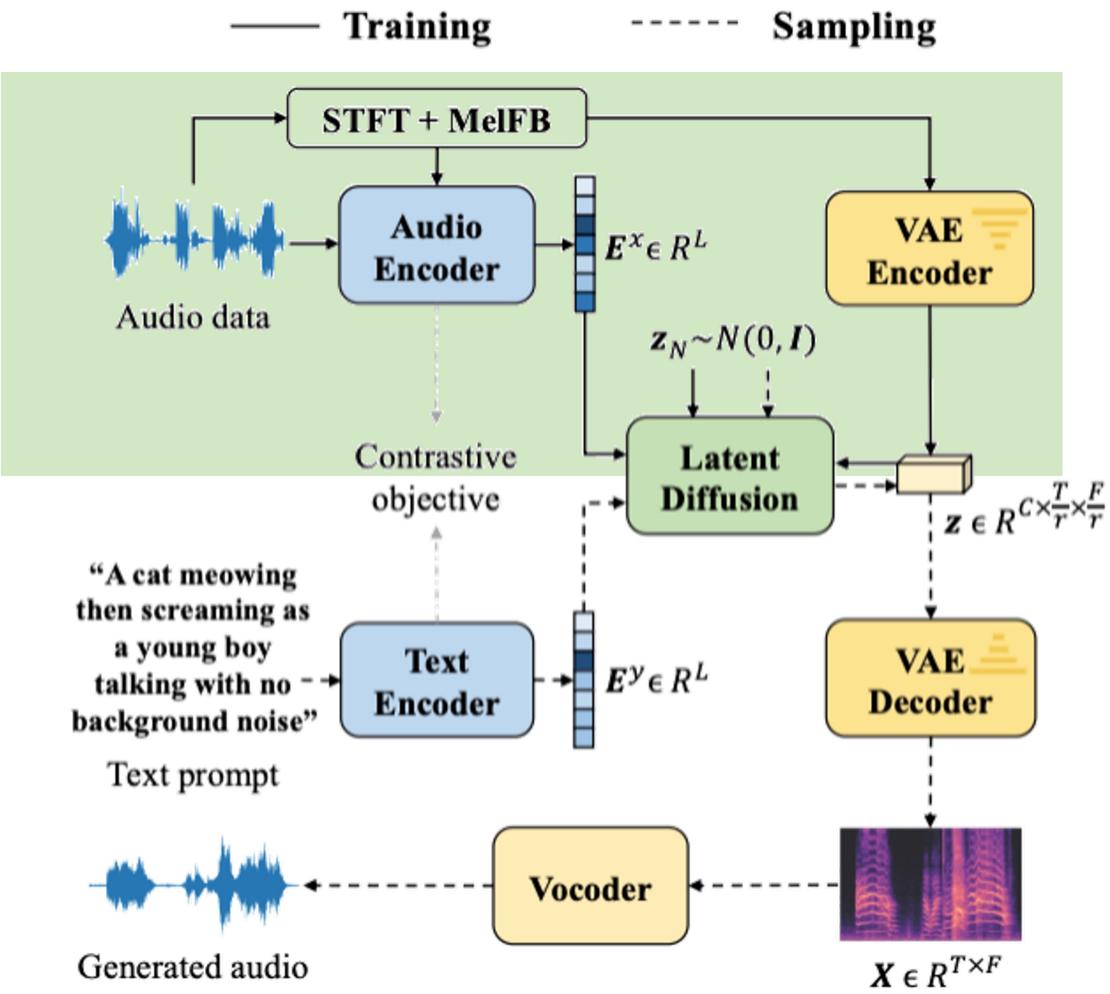


- **Text**
- **Audio**
- New audio data
- Augmented audio data

UNIVERSITY OF SURREY
Imperial College London

Overall Advantages

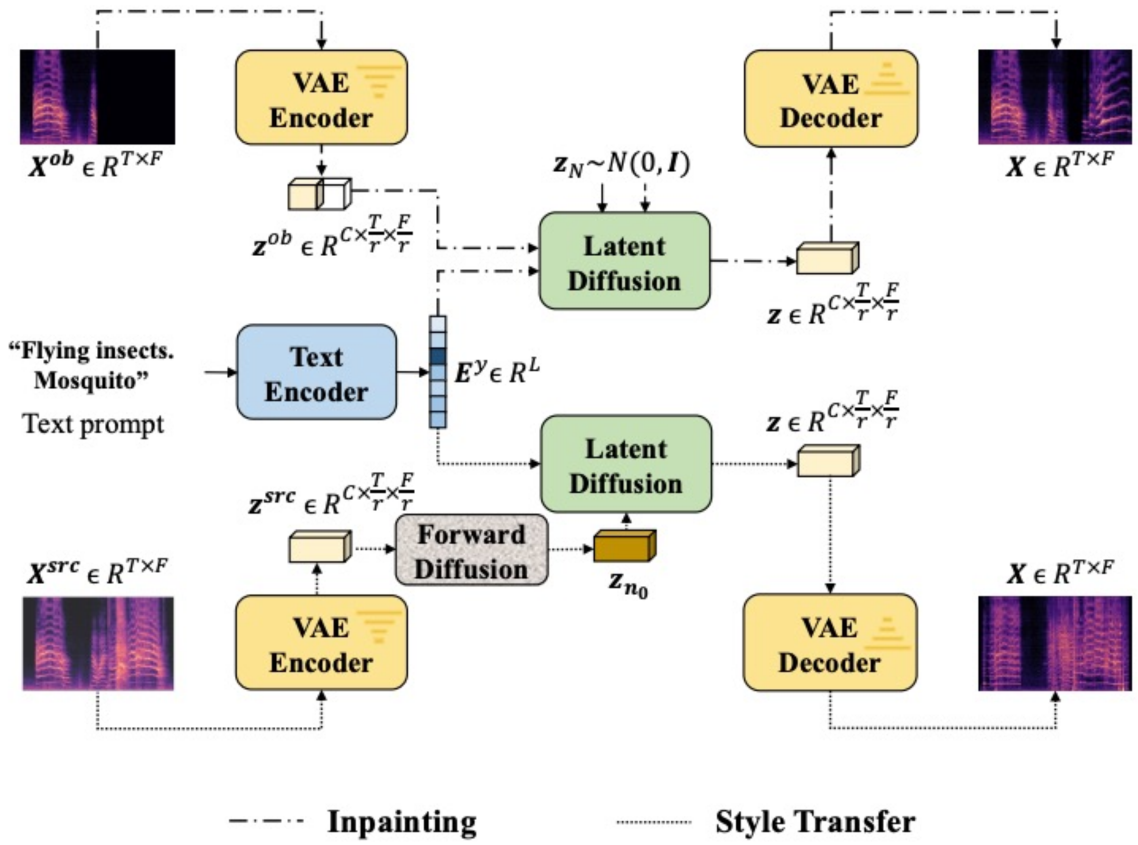
- **Less computation cost**
  - Latent Diffusion Models.
- **Less dependency on audio-text pairs.**
  - Train LDMs by self supervision
- **Continueous latent space**
  - Zero-shot audio style transfer.
  - Zero-shot audio super-resolution
  - Zero-shot audio inpainting.
  - …



UNIVERSITY OF SURREY
Imperial College London

Zero-shot down stream tasks

- Audio style transfers
  - Corrupt -> Reverse Diffusion
- Audio inpainting
  - Provide temporal hint during sampling.
- Audio super-resolutions
  - Provide frequency hint during sampling.



Training Data (16 kHz)

- AudioSet
- AudioCaps
- FreeSound
- BBC Sound Effect Library

Finally: 3,302,553 ten-seconds (9000+ hours) audio samples **without text labels.**

Largest scale so far

Evaluation Metrics

- Subjective evaluation
  - **OVL:** Overall quality
  - **REL:** relevance to text
- Objective evaluation
  - **FD**: Frechet Distance
  - **IS**: Inception Score
  - **KL**: Kullback-Leibler Divergence

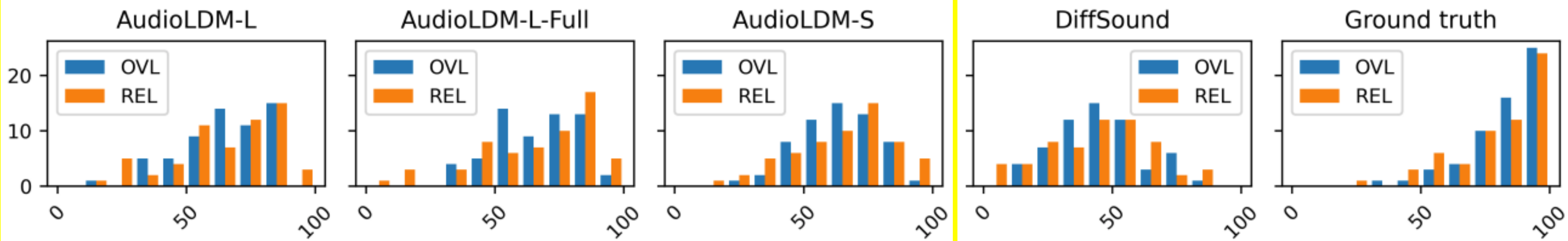
File name	Text description	Overall impression (1-100)	Relation to the text description (1-100)
random_name_108029.wav	A man talking followed by lights scrapping on a wooden surface	80	90
random_name_108436.wav	Bicycle Music Skateboard Vehicle	70	80
random_name_116883.wav	A power tool drilling as rock music plays	90	95
...	...	...	...

Example questionnaire for human evaluation. The participant will need to fill in the last two columns.

UNIVERSITY OF SURREY
Imperial College London

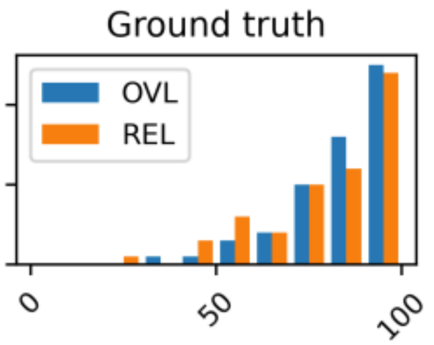
Result – SOTA comparison

Model	Datasets	Text	Params	FD ↓	IS ↑	KL ↓	FAD ↓	OVL ↑	REL ↑
Ground truth	-	-	-	-	-	-	-	83.61	80.11
DiffSound† (Yang et al., 2022)	AS+AC	✓	400M	47.68	4.01	2.52	7.75	45.00	43.83
AudioGen† (Kreuk et al., 2022)	AS+AC+8 others	✓	285M	-	-	2.09	3.13	-	-
AudioLDM-S	AC	✗	181M	29.48	6.90	1.97	2.43	63.41	64.83
AudioLDM-L	AC	✗	739M	27.12	7.51	1.86	2.08	64.30	64.72
AudioLDM-L-Full	AS+AC+2 others	✗	739M	**23.31**	**8.13**	**1.59**	**1.96**	**65.91**	**65.97**

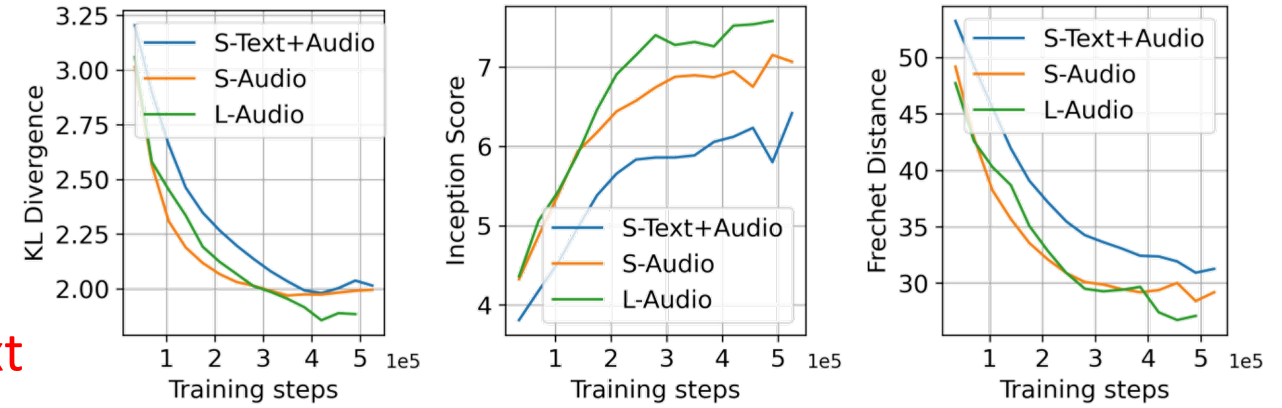


Trained on a single 3090 or A100 GPU!

Result – self-supervised LDMs training



- Training with audio can even outperform training with audio-text pairs.
- Reason:
  - Audio representaiton is better than Text
  1. Text labeling sometimes have weak relations to audio
     - e.g., Boats: Battleships-5.25 conveyor space
  2. Text labeling is error-prone
     - Missing labels in text.
     - Text is difficult to include every details.



Model	Text	Audio	FD ↓	IS ↑	KL ↓
AudioLDM-S	✓	✓	31.26	6.35	2.01
AudioLDM-S	✗	✓	29.48	6.90	1.97

UNIVERSITY OF SURREY

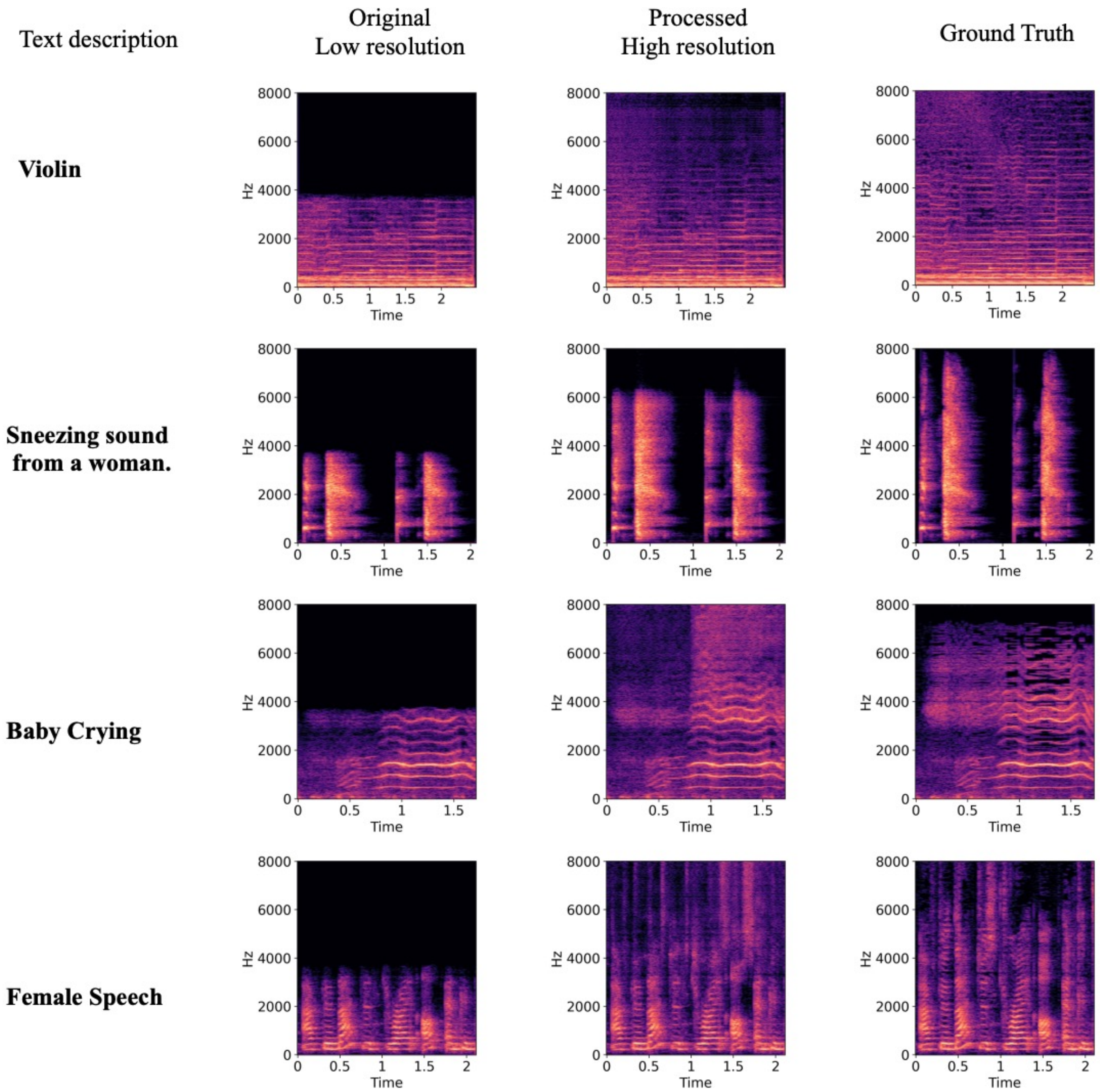
Imperial College London

Result – Super-resolution and Inpainting

- Super-resolution
  - VCTK (Speech)
  - AudioCaps (General Audio)
- Inpainting
  - AudioCaps

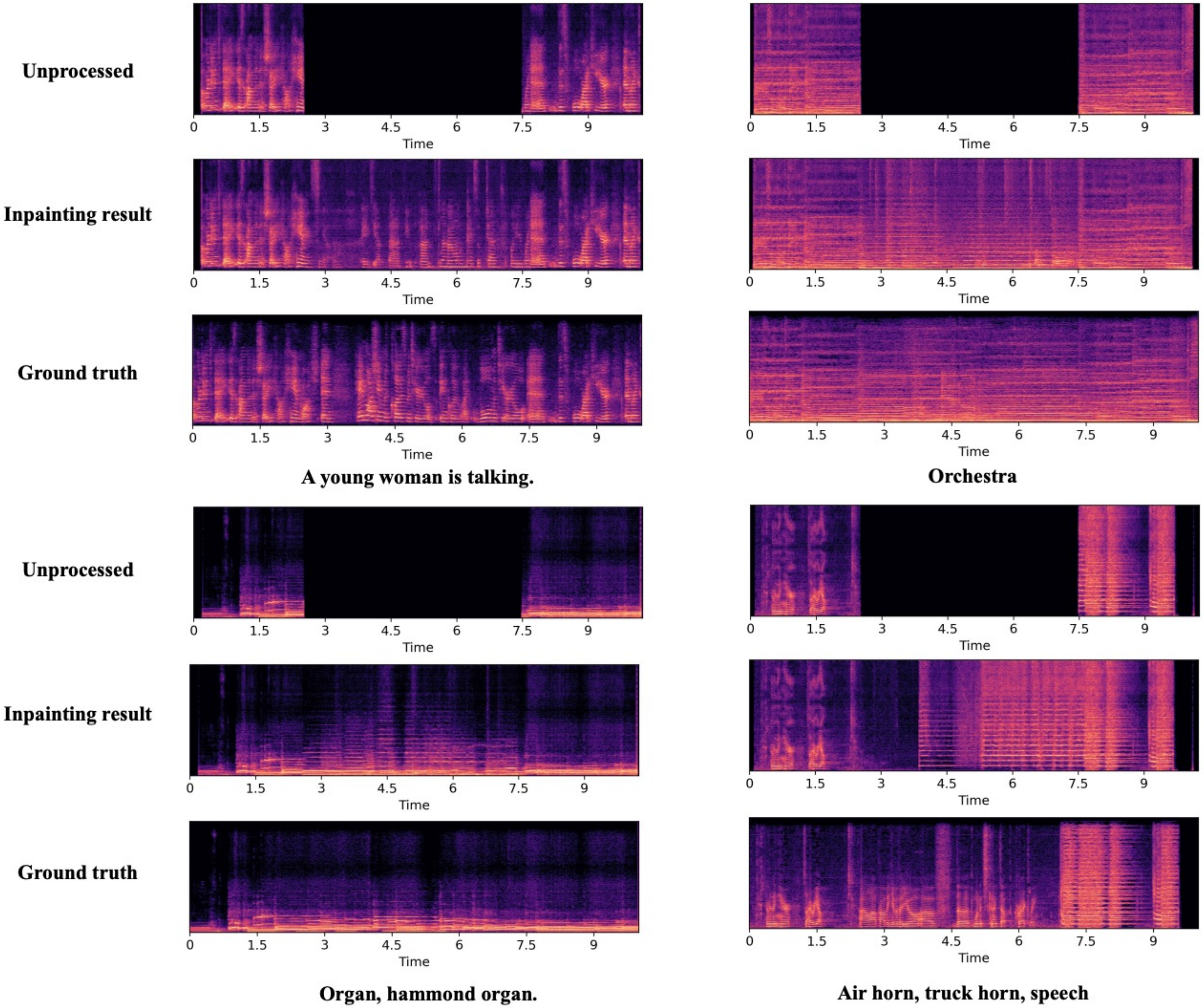
Task	Super-resolution		Inpainting
Dataset	AudioCaps	VCTK	AudioCaps
Unprocessed	2.76	2.15	10.86
Kuleshov et al. (2017)	-	1.32	-
Liu et al. (2022a)	-	**0.78**	-
AudioLDM-S	1.59	1.12	2.33
AudioLDM-L	**1.43**	0.98	**1.92**

Super-resolution: Log-spectral distance
Inpainting: Frechet audio distance



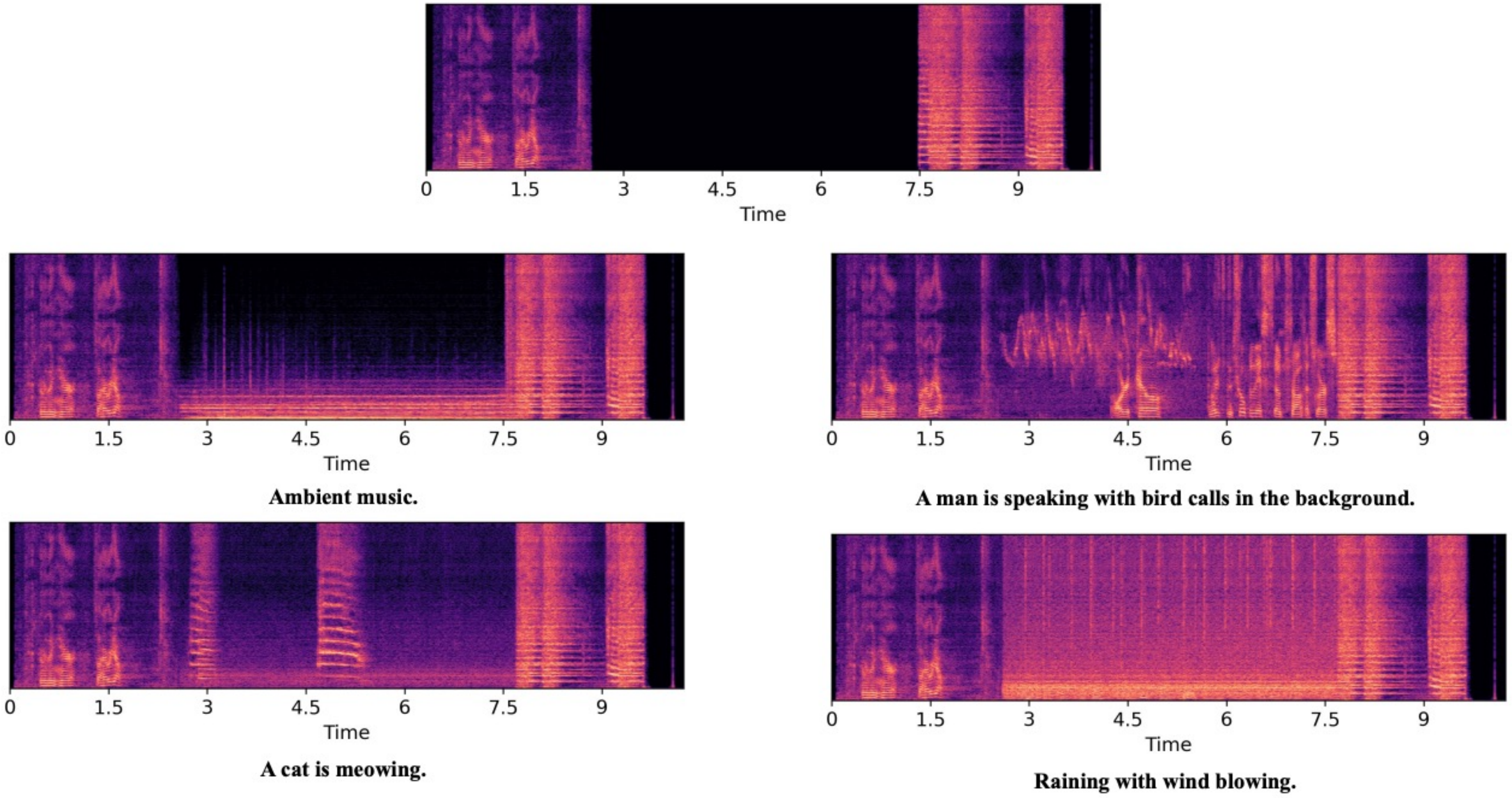
Inpainting

- Examples
  - Use matched text
  - Use un-matched text



Inpainting

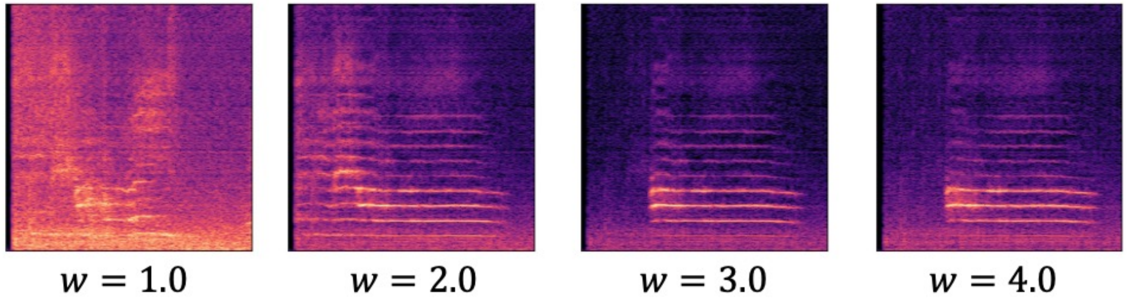
- Examples
  - Use matched text
  - Use un-matched text

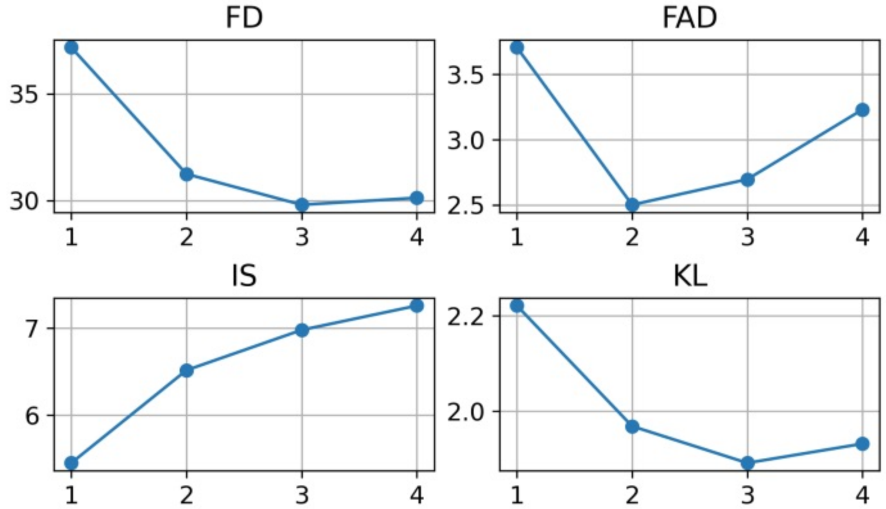


Result – Other details

- A good CFG scale is around 2.5
  - Large CFG: Less diversity
  - Small CFG: better diversity, less quality
- Different VAE compression levels.
  - 4, 8, 16
- Evaluation on AudioSet
- Sampling Steps (around 100 DDIM).
- Other ablation studies.

Effect of different classifier-free guidance scale



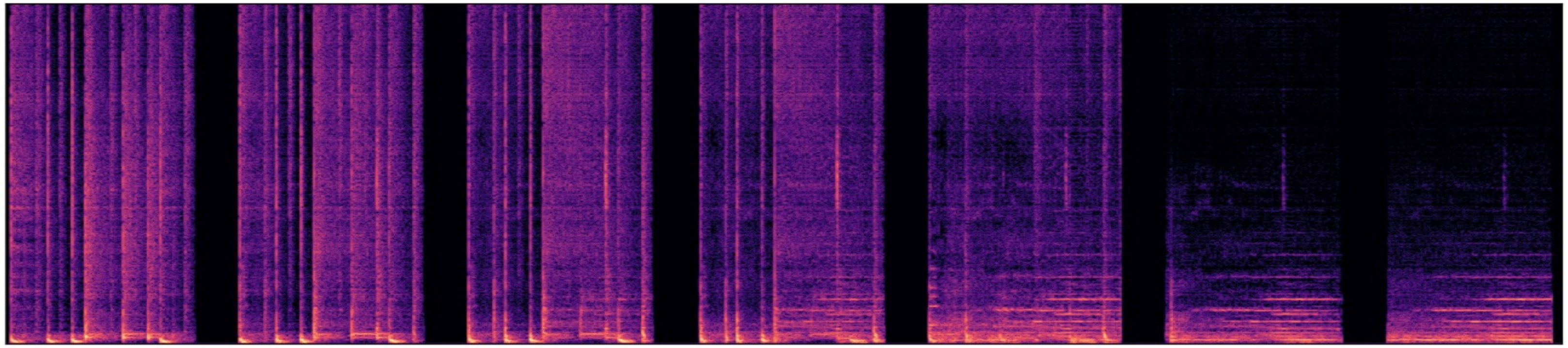


DDIM steps	10	25	50	100	200
FD	55.84	42.84	35.71	30.17	**29.48**
IS	4.21	5.91	6.51	6.85	**6.90**
KL	2.47	2.12	2.01	**1.94**	1.97

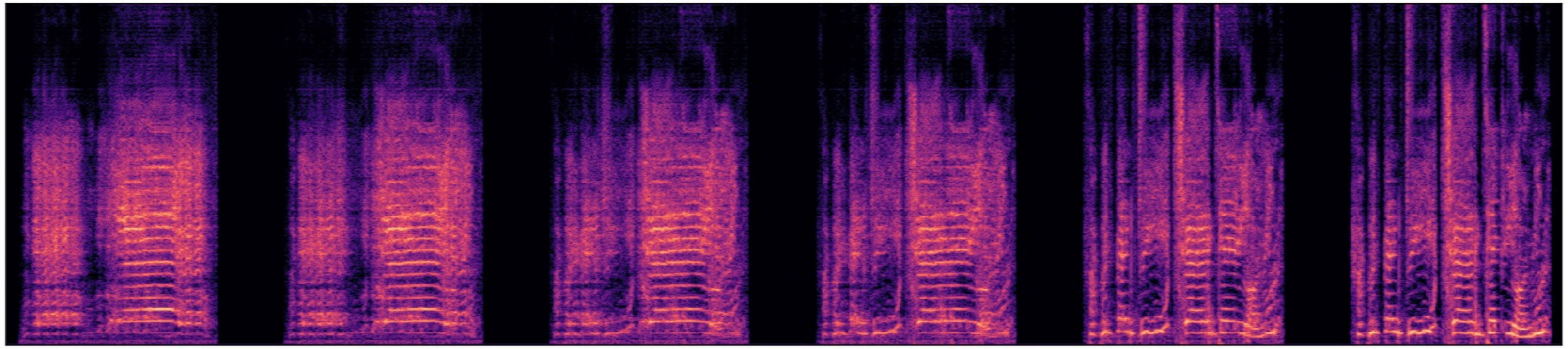
UNIVERSITY OF SURREY

Imperial College London

Audio Style Transfer



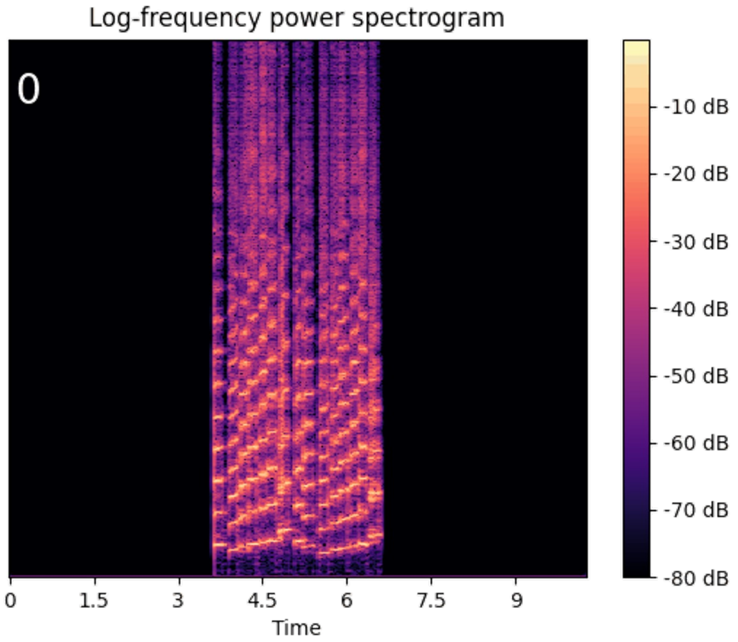
Drum beats → Ambient Music



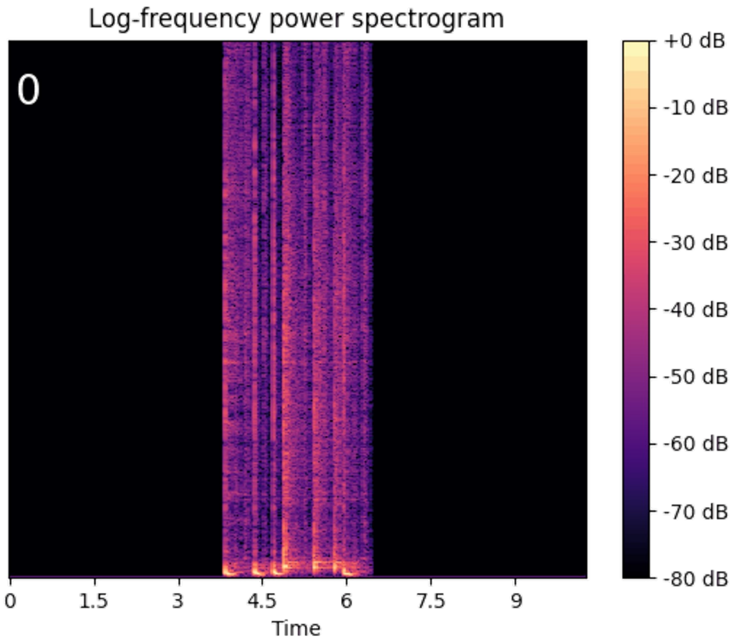
Sheep vocalization → Narration, monologue

UNIVERSITY OF SURREY Imperial College London

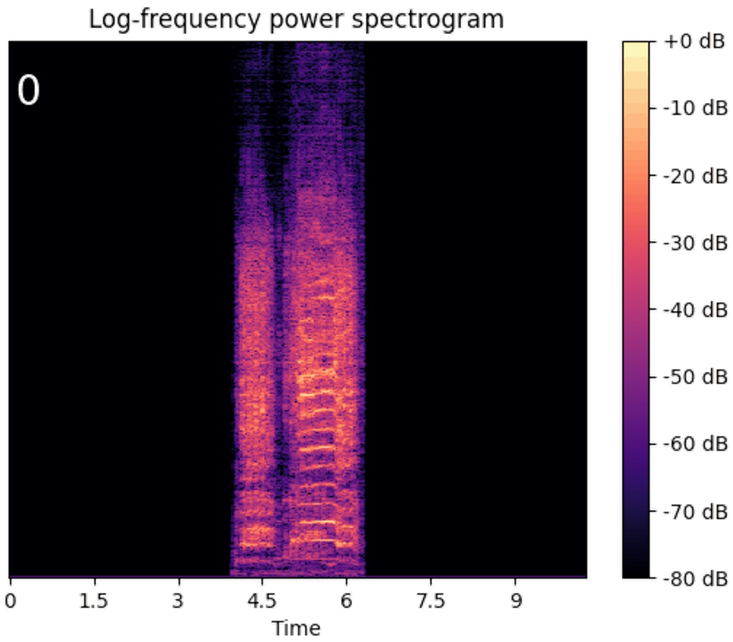
Audio Style Transfer



Trumpet
→ Children Singing



Drum beats
→ Ambient Music



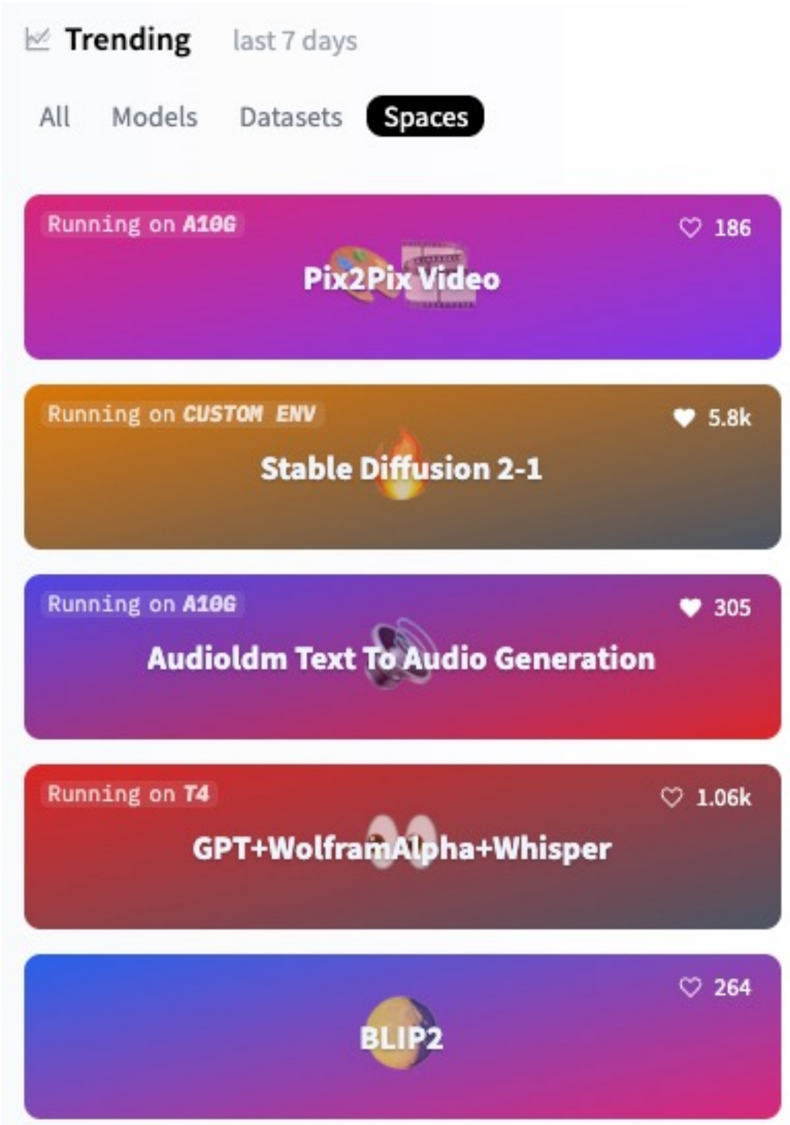
Sheep vocalization
→ Narration, monologue

UNIVERSITY OF SURREY

Imperial College London

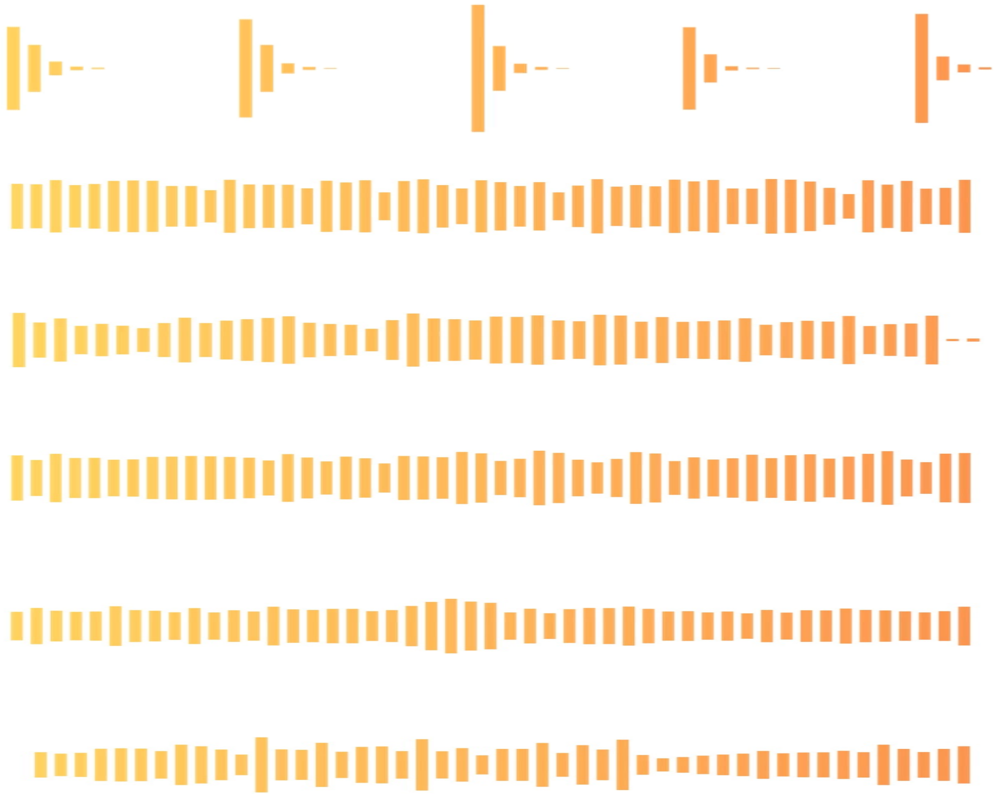
More examples

- Audio super-resolution
- Audio inpainting
- Fine-grained generation control:
  - Controls of object materials
  - Controls of acoustic environment
  - Controls of audio pitch
  - Controls of temporal orders
  - …



More examples

- A stone is hitting a metal plate
- Dance music with strong beats played by multiple instruments
- healthy deep gurgly 10 second burp
- Very windy condition, trying to fly against the wind in a parachute
- A small water steam in a forest with some bird vocalization
- somone slurping noodles long slurp



UNIVERSITY OF SURREY

Imperial College London

More examples (wired sound)

- The weirdest sound in existence

- The cry of Cthulhu the terrifying ancestral deity

- A man is speaking backwards creepily and exhaustively

More examples

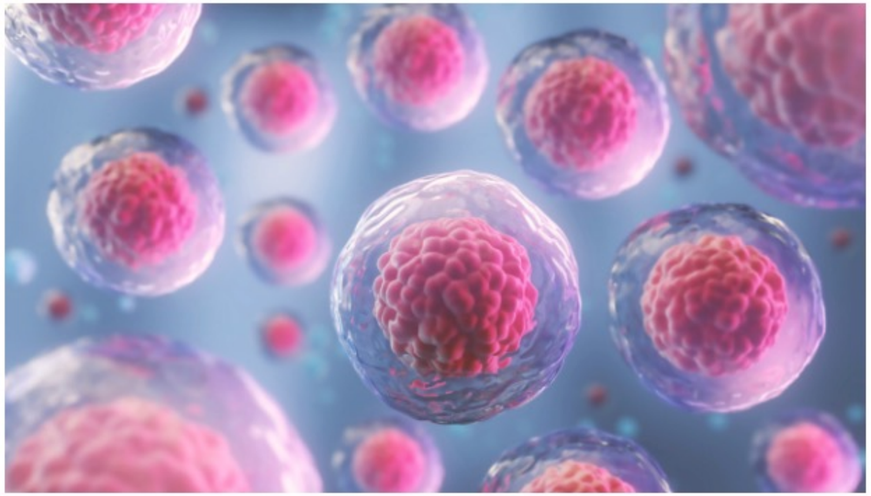
- Brain bubbles floating in primordial goo
- 漂浮在原始粘液中的脑泡

Brain bubbles floating in primordial goo #20
by evalive - opened 3 days ago

Discussion

evalive 3 days ago

Image input:



Sound Effect:

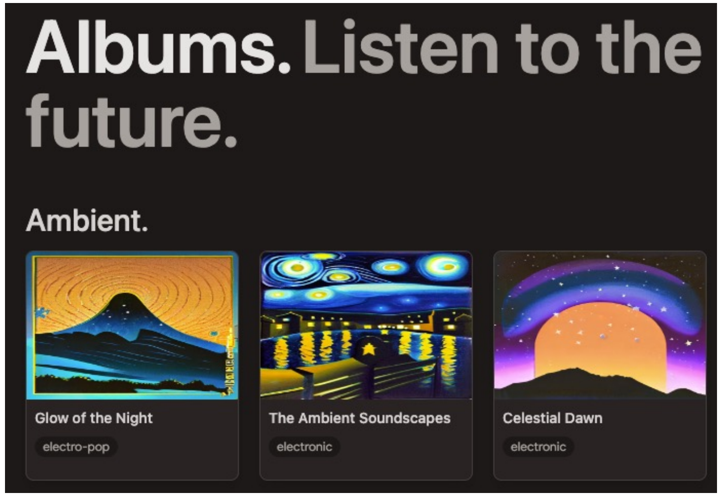
0:10 / 0:10

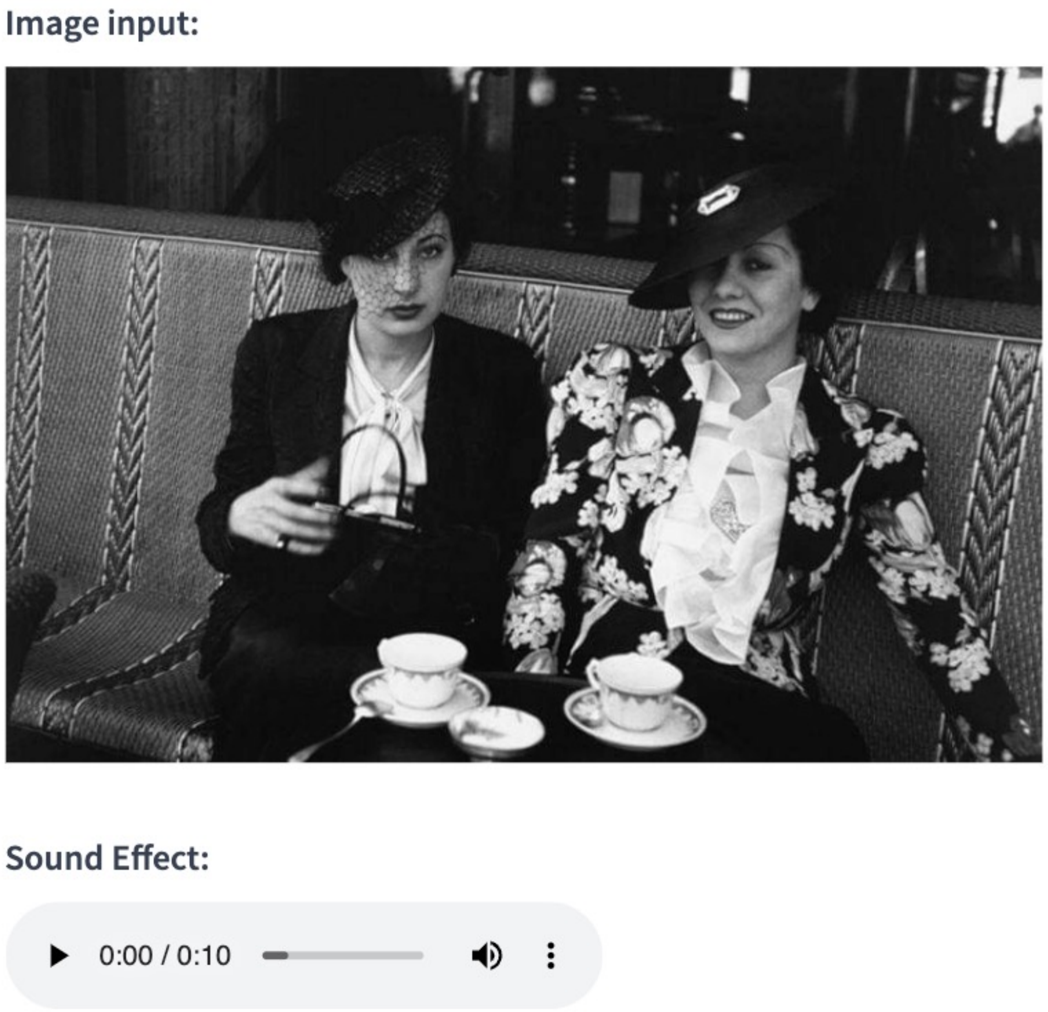
UNIVERSITY OF SURREY

Imperial College London

Interesting resources

- Image-to-Audio
  - https://huggingface.co/spaces/fffiloni/image-to-sound-fx
- AI music album:
  - https://www.latent.store/albums





UNIVERSITY OF SURREY Imperial College London

AudioLDM on Diffuser

Credit to **Sanchit Gandhi** from Hugging Face

```
from diffusers import AudioLDMPipeline

import torch import scipy

repo_id = "sanchit-gandhi/audioldm-text-to-audio"

pipe = AudioLDMPipeline.from_pretrained(repo_id, torch_dtype=torch.float16)
pipe = pipe.to("cuda")

prompt = "Techno music with a strong, upbeat tempo and high melodic riffs"

audio = pipe(prompt, num_inference_steps=10, height=512).audios[0]

# save the audio sample as a .wav file

scipy.io.wavfile.write("techno.wav", rate=16000, data=audio)
```

A few take aways here, thanks!

- Paper (https://arxiv.org/abs/2301.12503):
  - AudioLDM: Text-to-Audio Generation with Latent Diffusion Models
- Project Page: https://audioldm.github.io/
- Hugging Face Space:
  - https://huggingface.co/spaces/haoheliu/audioldm-text-to-audio-generation
- Github:
  - Pretrained model: https://github.com/haoheliu/AudioLDM
  - Evaluation tools: https://github.com/haoheliu/audioldm_eval
- Interesting demo website:
  - https://www.latent.store/albums

@LiuHaohe

UNIVERSITY OF SURREY
Imperial College London