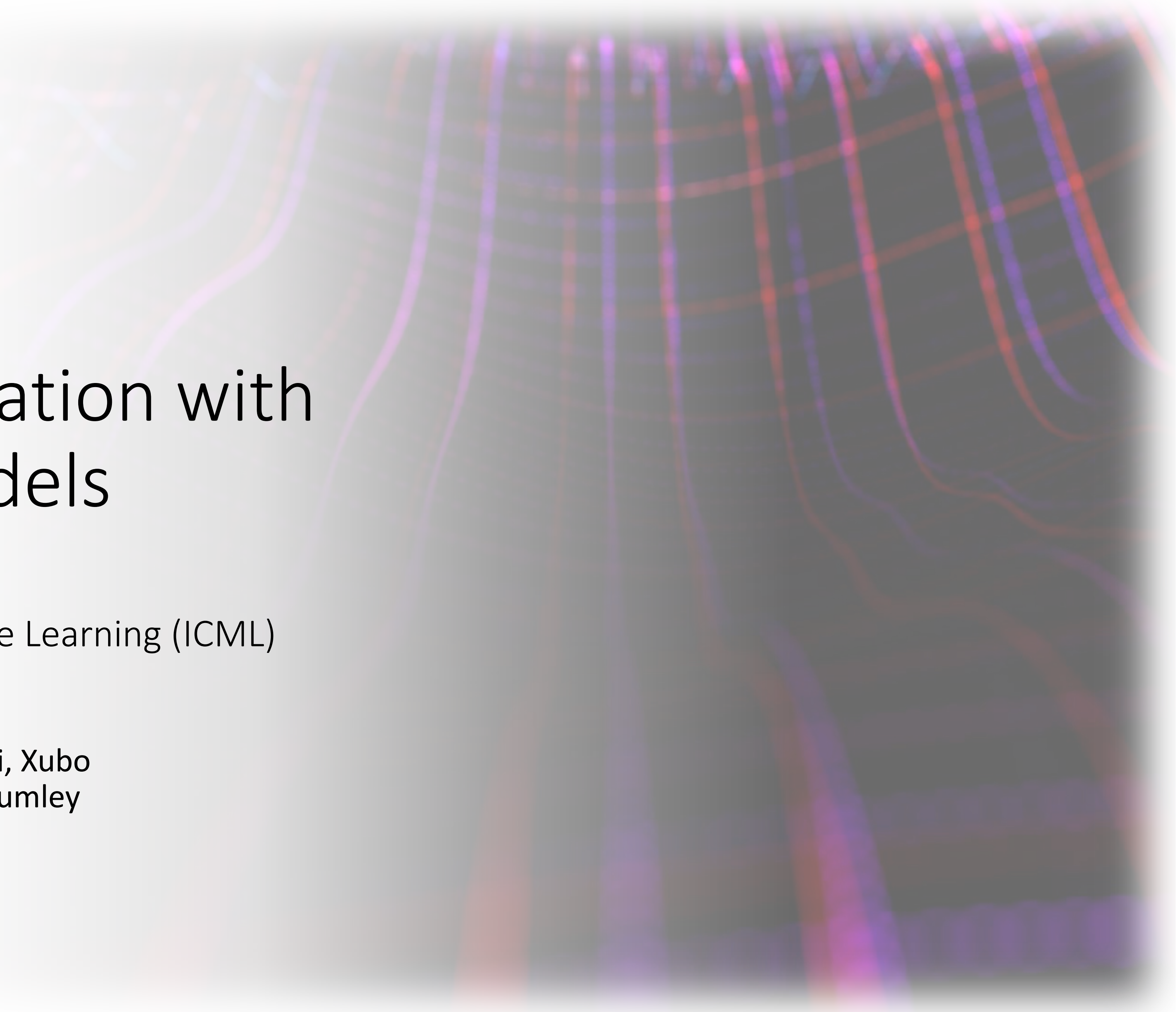


AudioLDM:
Text-to-Audio Generation with Latent Diffusion Models

The International Conference on Machine Learning (ICML)

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

UNIVERSITY OF SURREY

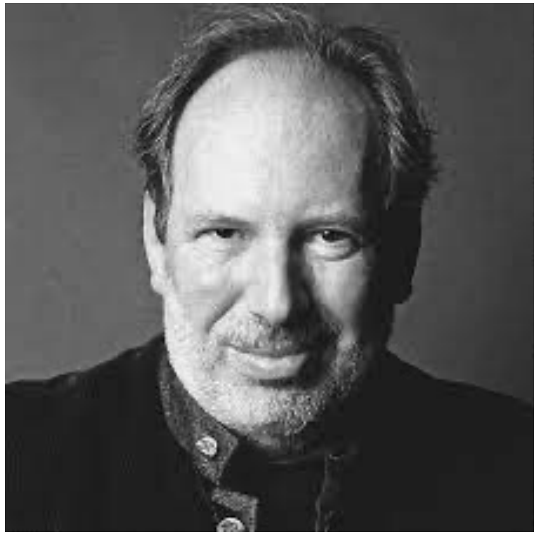
Imperial College London

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)

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
  - …
- In the Academia



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

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

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),…

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!

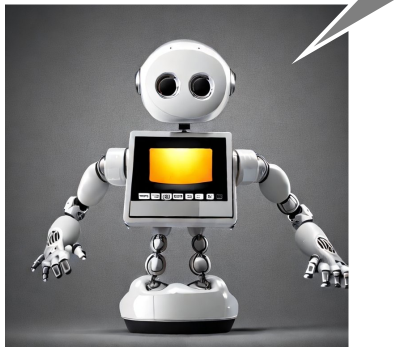
Self-supervised Audio Generation

Step 1

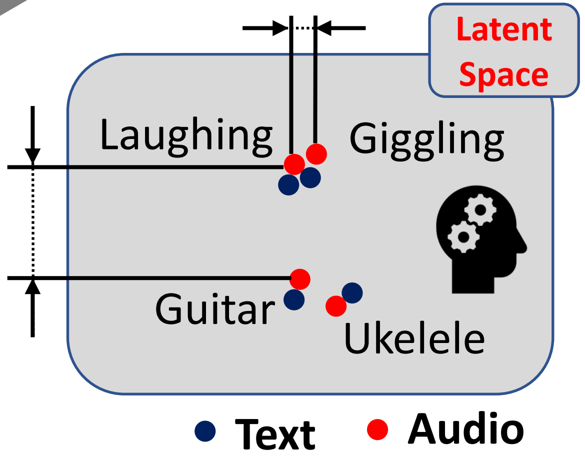
Audio ⇆ Text

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

AI



OK Got it

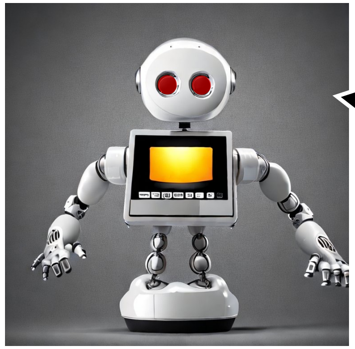


Step 2

Audio

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

AI



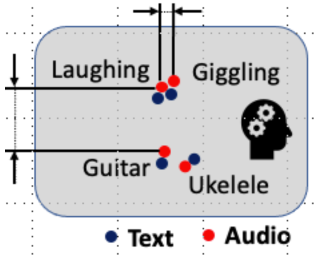
Nobody knows audio better than me!!!

Tell me what you want!

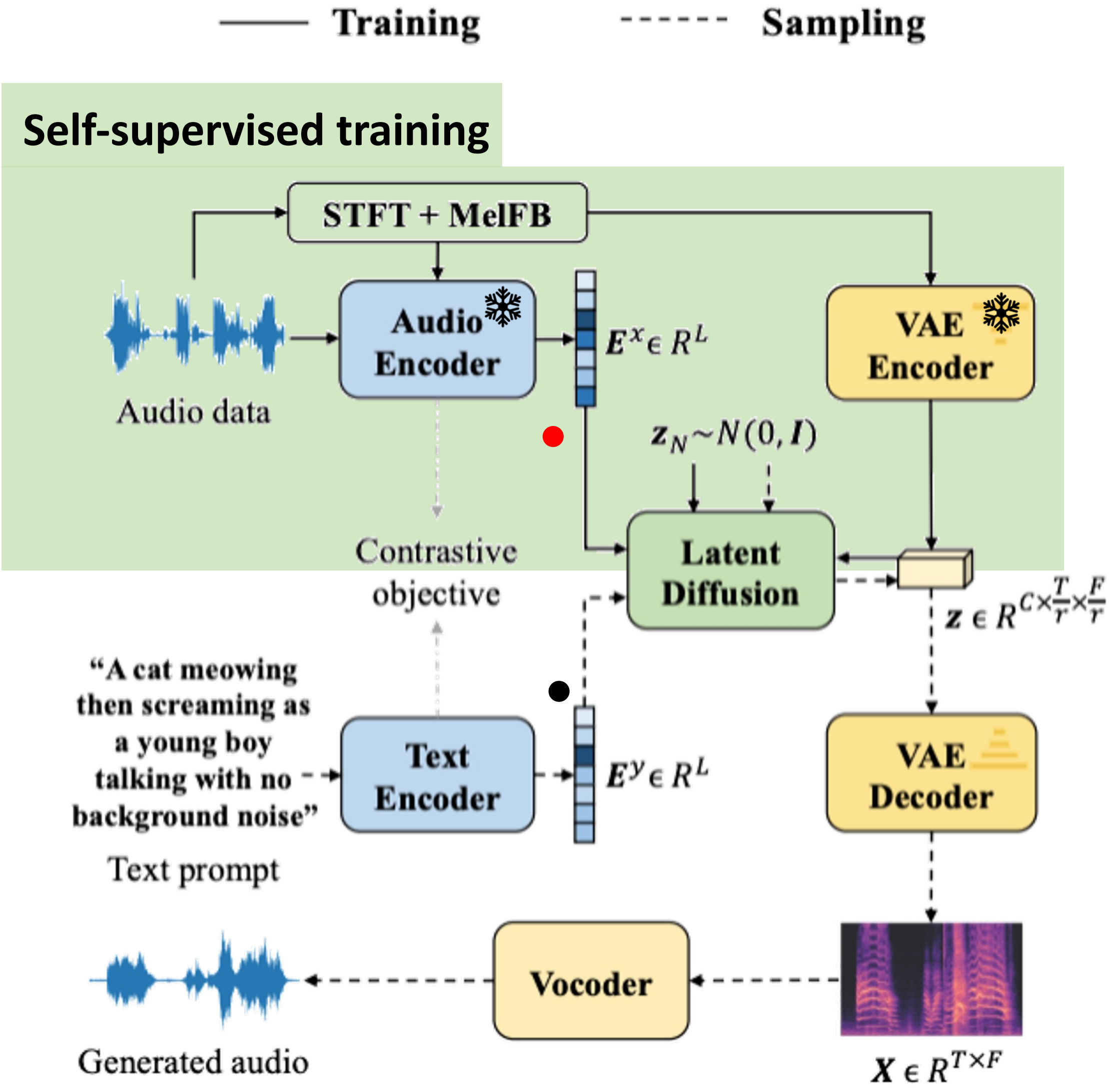
UNIVERSITY OF SURREY

Imperial College London

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

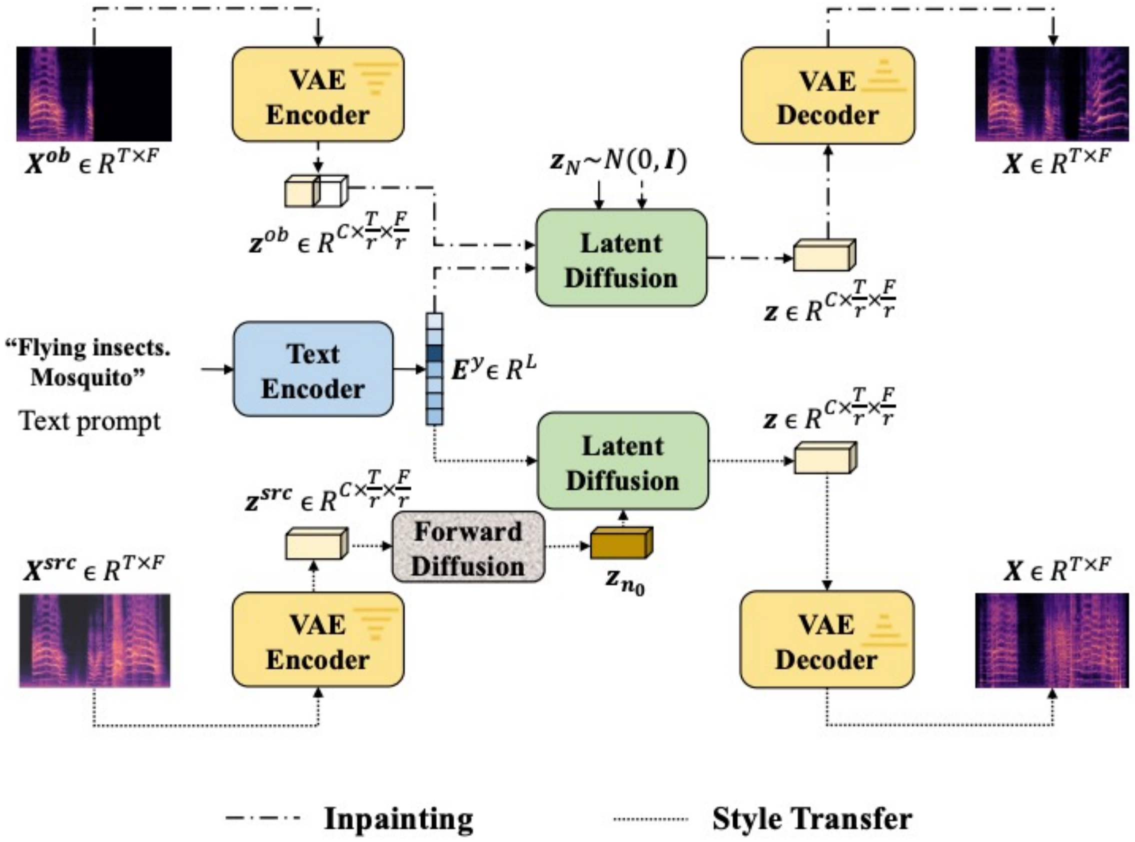


UNIVERSITY OF SURREY

Imperial College London

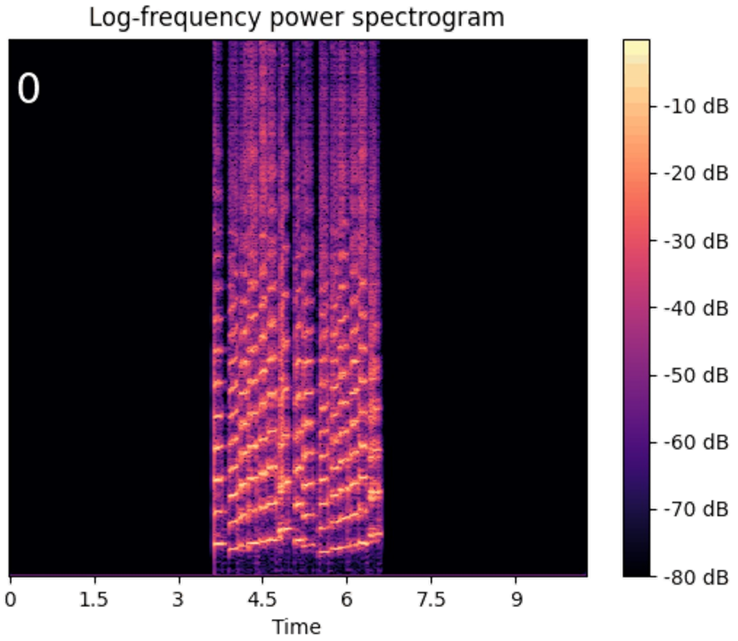
Zero-shot down stream tasks

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

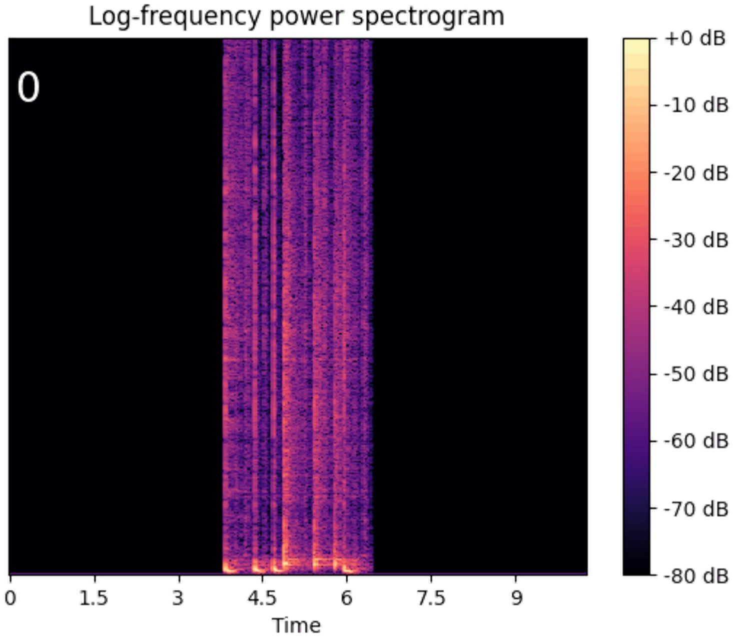


$X^{ob} \in R^{T\times F}$
VAE Encoder
$z^{ob} \in R^{C\times\frac{T}{r}\times\frac{F}{r}}$
$z_N \sim N(0, I)$
Latent Diffusion
VAE Decoder
$X \in R^{T\times F}$
$z \in R^{C\times\frac{T}{r}\times\frac{F}{r}}$
"Flying insects. Mosquito"
Text prompt
Text Encoder
$E^y \in R^L$
Latent Diffusion
$z \in R^{C\times\frac{T}{r}\times\frac{F}{r}}$
$z^{src} \in R^{C\times\frac{T}{r}\times\frac{F}{r}}$
Forward Diffusion
z_{n_0}
$X^{src} \in R^{T\times F}$
VAE Encoder
VAE Decoder
$X \in R^{T\times F}$
Inpainting
Style Transfer

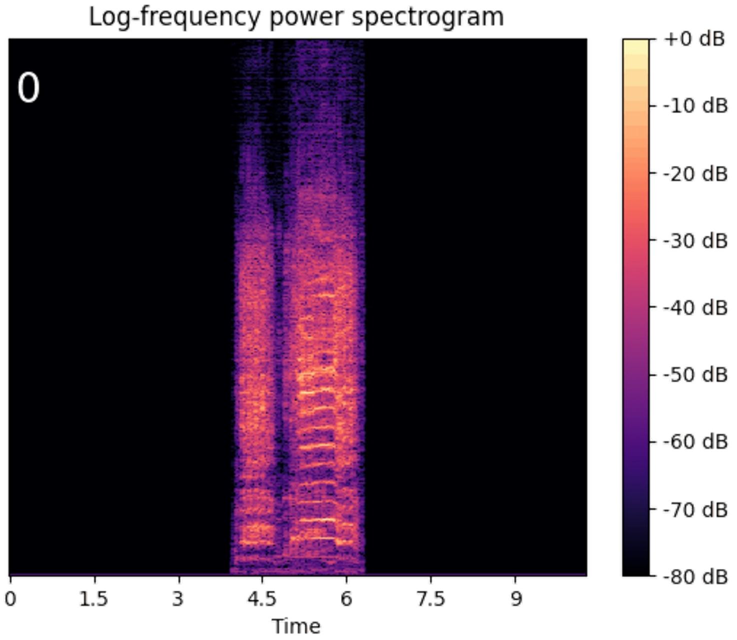
Audio Style Transfer



Trumpet
→ Children Singing

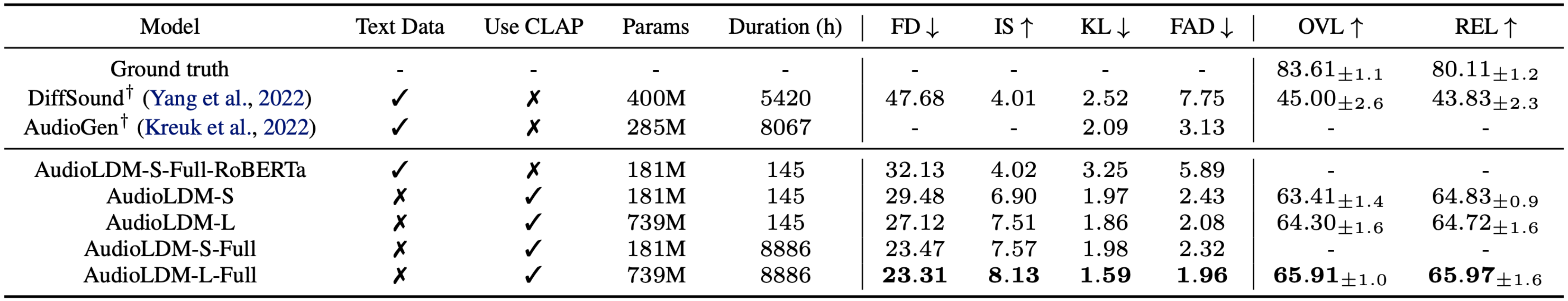


Drum beats
→ Ambient Music

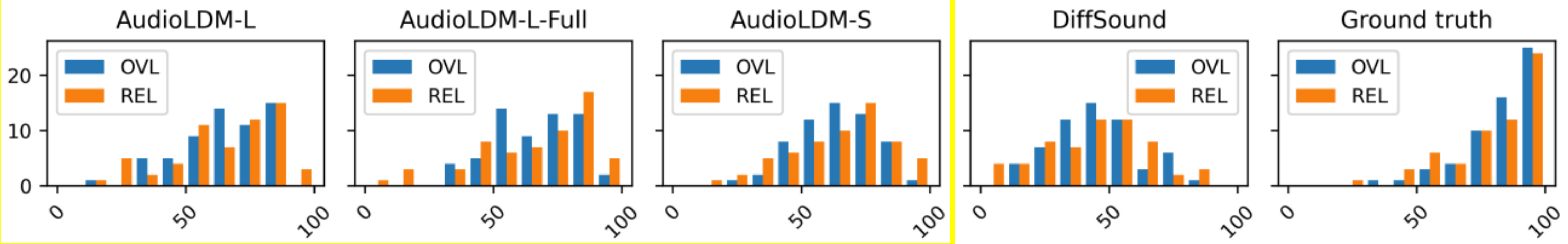


Sheep vocalization
→ Narration, monologue

Result – SOTA comparison



Model	Text Data	Use CLAP	Params	Duration (h)	FD ↓	IS ↑	KL ↓	FAD ↓	OVL ↑	REL ↑
Ground truth	-	-	-	-	-	-	-	-	$83.61_{\pm 1.1}$	$80.11_{\pm 1.2}$
DiffSound† (Yang et al., 2022)	✓	✗	400M	5420	47.68	4.01	2.52	7.75	$45.00_{\pm 2.6}$	$43.83_{\pm 2.3}$
AudioGen† (Kreuk et al., 2022)	✓	✗	285M	8067	-	-	2.09	3.13	-	-
AudioLDM-S-Full-RoBERTa	✓	✗	181M	145	32.13	4.02	3.25	5.89	-	-
AudioLDM-S	✗	✓	181M	145	29.48	6.90	1.97	2.43	$63.41_{\pm 1.4}$	$64.83_{\pm 0.9}$
AudioLDM-L	✗	✓	739M	145	27.12	7.51	1.86	2.08	$64.30_{\pm 1.6}$	$64.72_{\pm 1.6}$
AudioLDM-S-Full	✗	✓	181M	8886	23.47	7.57	1.98	2.32	-	-
AudioLDM-L-Full	✗	✓	739M	8886	**23.31**	**8.13**	**1.59**	**1.96**	$\mathbf{65.91}_{\pm 1.0}$	$\mathbf{65.97}_{\pm 1.6}$



Trained on a single 3090 or A100 GPU!

A few take aways here, thanks!

- Project Page: https://audioldm.github.io/
- Hugging Face Space (Listen to the samples contributed by the community!):
  - 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