Learning Audio Pattern with Latent Diffusion Model

Talk at the Spoken Language System (SLS) group, CSAIL, MIT

Haohe Liu

Final-year PhD Student

Centre for Vision, Speech and Signal Processing (CVSSP)

University of Surrey

Haohe Liu

Final year PhD Student

University of Surrey

Centre for Vision, Speech and Signal Processing

Supervisors:
Prof. Mark D. Plumbley
Prof. Wenwu Wang

In 2014, I aspired to become a pianist.

- Build audio technology that **inspires creativity** and **enhances communications**
- 🔬 Research
  - Audio and Music Generation; Text-to-Speech; Audio Recognition; Audio Quality Enhancement; Audio Source Separation, etc.
- 📊 Stats
  - 7000+ GitHub stars; 900+ citations; 100,000+ checkpoint downloads
  - ICML, AAAI, NeurIPS, TPAMI, TASLP, ICASSP, INTERSPEECH, etc.

https://haoheliu.github.io/

Overview

- Background (10 mins):
  - Diffusion Model
  - Latent Diffusion Model
  - Conditional Latent Diffusion Model
- Applications (20 mins):
  - AudioLDM 1&2, MusicLDM: Text-to-Audio Generation
  - AudioSR: Audio Super-resolution
  - SemantiCodec: Ultra-low bitrate semantic audio codec

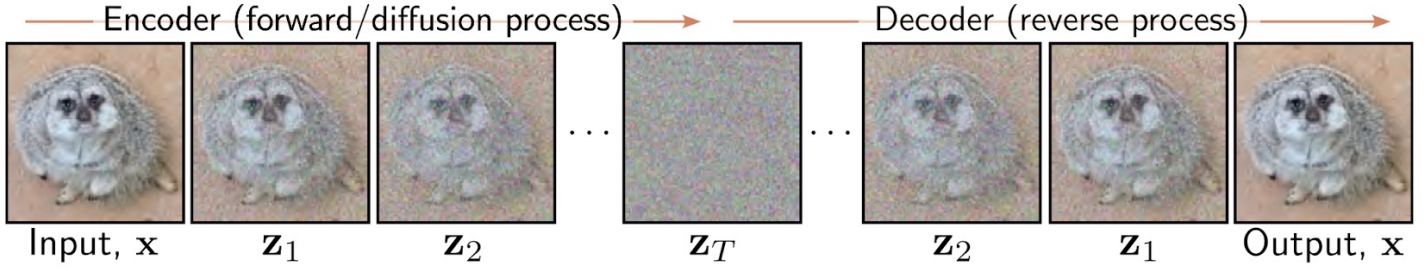
Background

Diffusion Model

Latent Diffusion Model

Conditional Latent Diffusion Model

Diffusion Model



- Forward Diffusion: Start with real data x (or z_0).
- Reverse Diffusion: Train a model to predict z_{t-1} from z_t: $q(\mathbf{z}_{t-1}|\mathbf{z}_t)$

markov chain $\mathbf{z}_t = \sqrt{1-\beta_t}\mathbf{z}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_t$

z_t in closed form $\mathbf{z}_t = \sqrt{\alpha_t}\cdot\mathbf{x} + \sqrt{1-\alpha_t}\cdot\boldsymbol{\epsilon},$

$$
\begin{aligned}
q(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x}) &= \frac{q(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{x})q(\mathbf{z}_{t-1}|\mathbf{x})}{q(\mathbf{z}_t|\mathbf{x})} \quad \leftarrow \text{Constant} \\
&\propto q(\mathbf{z}_t|\mathbf{z}_{t-1})q(\mathbf{z}_{t-1}|\mathbf{x}) \\
&= \text{Norm}_{\mathbf{z}_t}\left[\sqrt{1-\beta_t}\cdot\mathbf{z}_{t-1}, \beta_t\mathbf{I}\right] \text{Norm}_{\mathbf{z}_{t-1}}\left[\sqrt{\alpha_{t-1}}\cdot\mathbf{x}, (1-\alpha_{t-1})\mathbf{I}\right] \\
&= \text{Norm}_{\mathbf{z}_{t-1}}\left[\frac{1}{\sqrt{1-\beta_t}}\mathbf{z}_t, \frac{\beta_t}{1-\beta_t}\mathbf{I}\right] \text{Norm}_{\mathbf{z}_{t-1}}\left[\sqrt{\alpha_{t-1}}\cdot\mathbf{x}, (1-\alpha_{t-1})\mathbf{I}\right]
\end{aligned}
$$

Gaussian Identity ↓

$$
q(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x}) = \text{Norm}_{\mathbf{z}_{t-1}}\left[\frac{(1-\alpha_{t-1})}{1-\alpha_t}\sqrt{1-\beta_t}\mathbf{z}_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\mathbf{x}, \frac{\beta_t(1-\alpha_{t-1})}{1-\alpha_t}\mathbf{I}\right]
$$

Conditional diffusion distribution in closed form

$$
\begin{aligned}
Pr(\mathbf{z}_T) &= \text{Norm}_{\mathbf{z}_T}[\mathbf{0}, \mathbf{I}] \\
Pr(\mathbf{z}_{t-1}|\mathbf{z}_t,\boldsymbol{\phi}_t) &= \text{Norm}_{\mathbf{z}_{t-1}}\left[\mathbf{f}_t[\mathbf{z}_t,\boldsymbol{\phi}_t], \sigma_t^2\mathbf{I}\right] \\
Pr(\mathbf{x}|\mathbf{z}_1,\boldsymbol{\phi}_1) &= \text{Norm}_{\mathbf{x}}\left[\mathbf{f}_1[\mathbf{z}_1,\boldsymbol{\phi}_1], \sigma_1^2\mathbf{I}\right],
\end{aligned}
$$

noise schedule

Diffusion Model Decoder

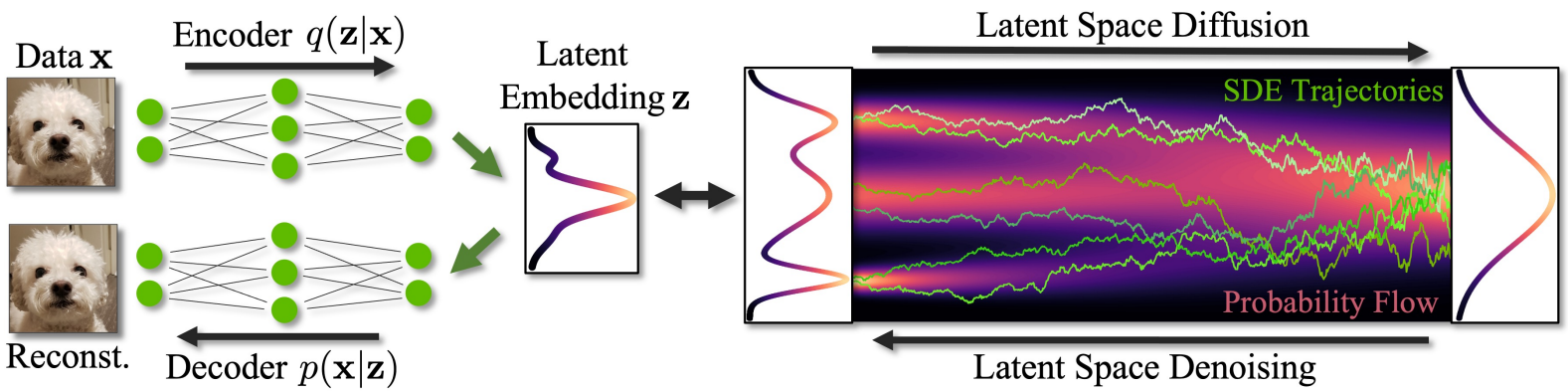
$$
D_{KL}\left[Pr(\mathbf{z}_{t-1}|\mathbf{z}_t,\boldsymbol{\phi}_t)||q(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})\right] = \frac{1}{2\sigma_t^2}\left\|\frac{(1-\alpha_{t-1})}{1-\alpha_t}\sqrt{1-\beta_t}\mathbf{z}_t + \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\mathbf{x} - \mathbf{f}_t[\mathbf{z}_t,\boldsymbol{\phi}_t]\right\|^2 + C,
$$

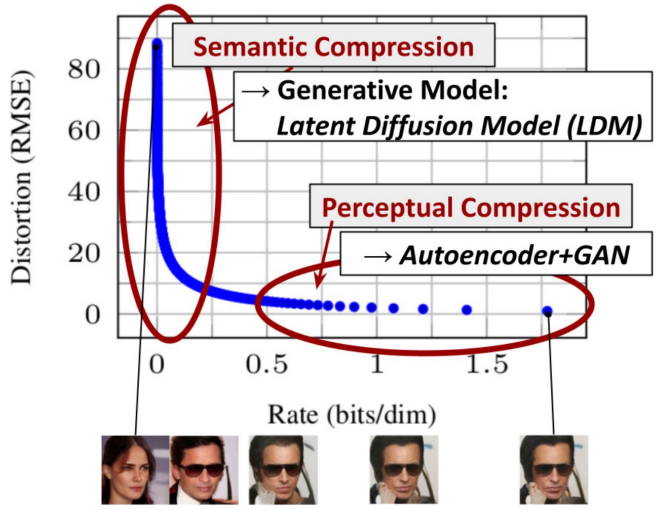
Derived from the Evidence Lower Bound of $\log\left[Pr(\mathbf{x},\boldsymbol{\phi}_{1\ldots T})\right]$

Diffusion Model Decoder Training

Latent Diffusion Model (Rombach et al. 2021)

- Diffusion modeling in a compressed space
  - Less computation
  - Better generative modeling

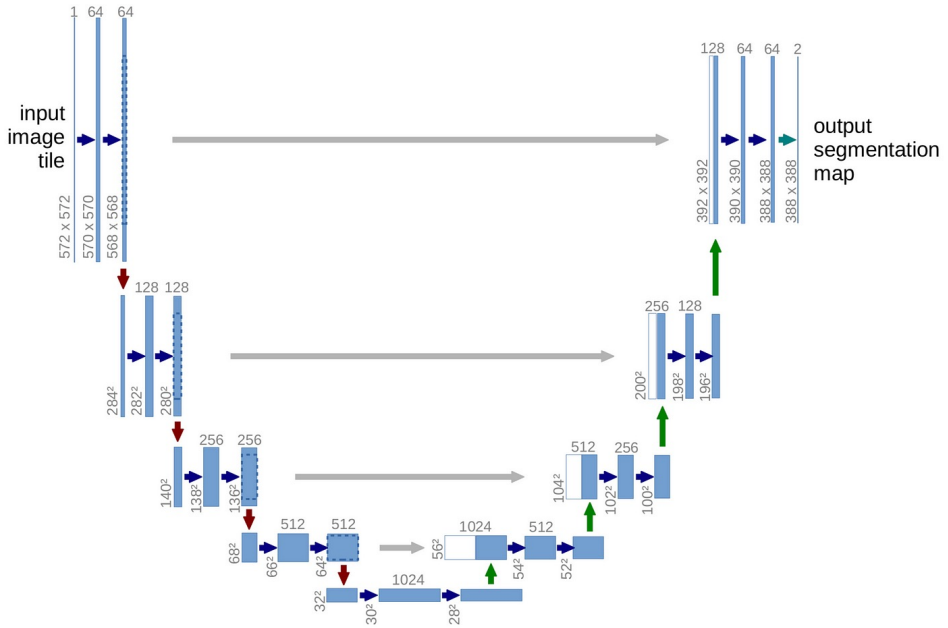




Credit: Latent Diffusion Models: Is the Generative AI Revolution Happening in Latent Space? (NeurIPS 2023 Tutorial)

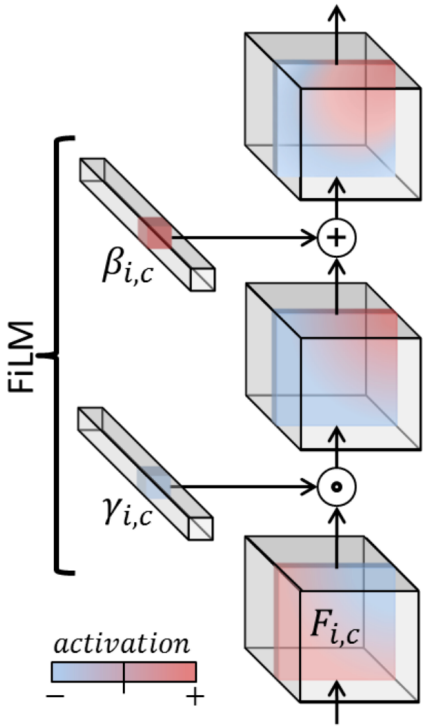
Backbone for the LDM

- Unet: A U-Shape encoders and decoders with skip-connection on the same level.
- Each layer of the encoder contains:
  - CNNs
  - Self-attention
  - (Optional) Cross-attention



Conditional LDM

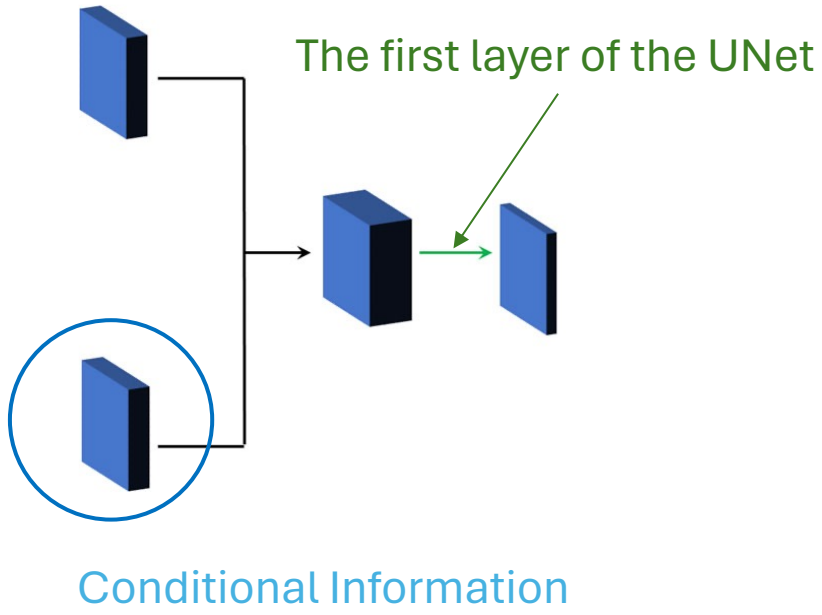
- FiLM: Feature-wise Linear Modulation
  - + 1D vector - Global conditional feature
  - + Marginal computational cost
  - - No temporal information
- Channel concatenation
- Cross-attention
- ControlNet



FiLM (Perez et al. 2017)

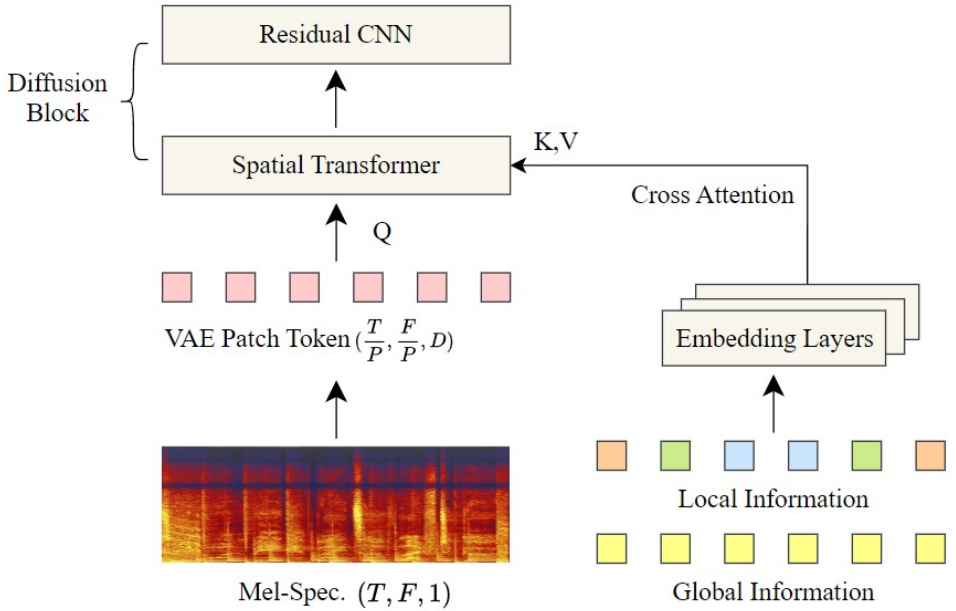
Conditional LDM

- FiLM: Feature-wise Linear Modulation
- Channel concatenation
  - + 2D tensor – Temporal Information Included
  - + Marginal computation introduced
  - - Fixed shape input
- Cross-attention
- ControlNet



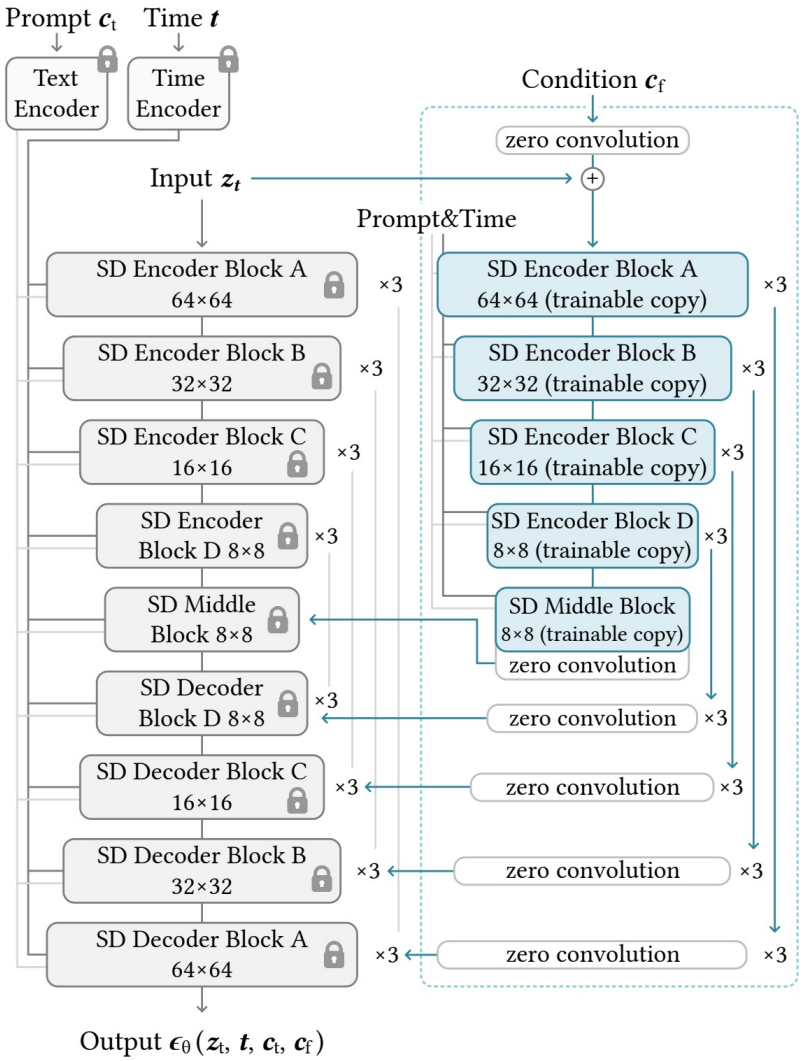
Conditional LDM

- FiLM: Feature-wise Linear Modulation
- Channel concatenation
- Cross-attention
  - + Support variable length and embed dimension
  - - Need extra attention layers
  - - Need positional encoding
- ControlNet



Conditional LDM

- FiLM: Feature-wise Linear Modulation
- Channel concatenation
- Cross-attention
- ControlNet
  - + Plug and play
  - + Configurable control strength
  - - Extra parameters
  - - Not always works well



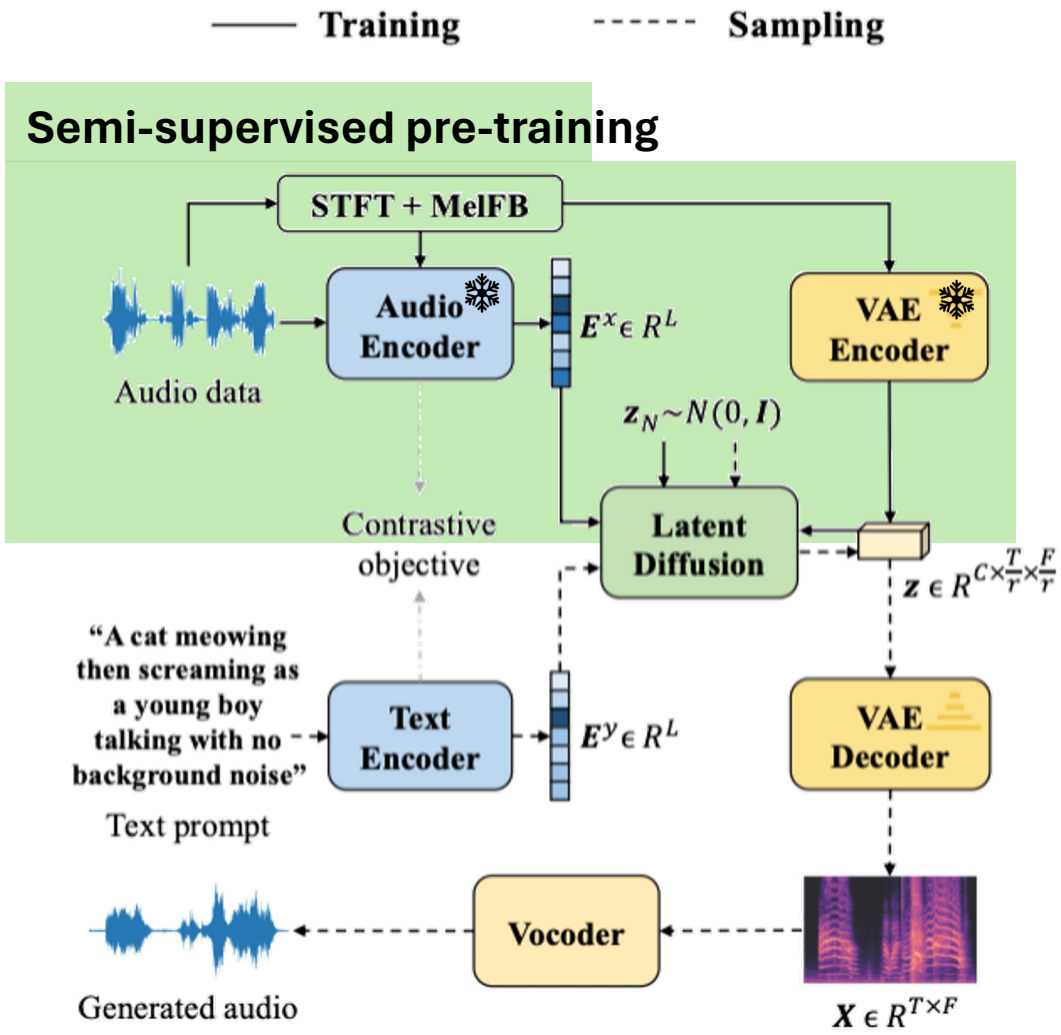
ControlNet (Zhang et al. 2023)

Text-to-Audio Generation

AudioLDM, AudioLDM 2

AudioLDM

- Enable the audio-only semi-supervised training approach
- Alleviate issues in LM approach such as slow inference speed, error propagation, etc.
- Related works
  - AudioLM, AudioGen, MusicGen, Make-an-Audio, DiffSound, MusicLM

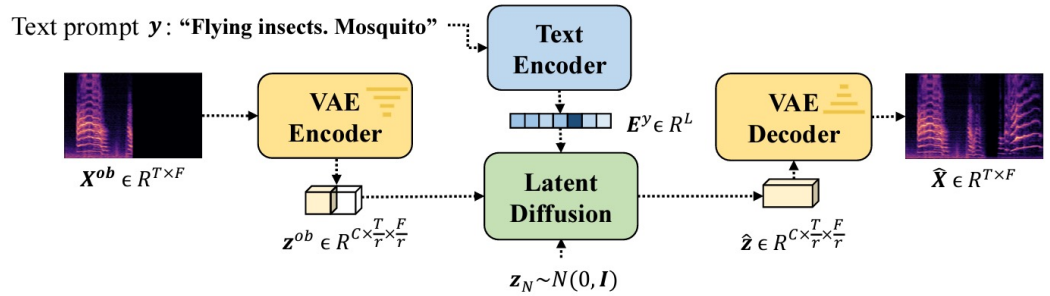


Liu, H.*, Chen, Z.*, Yuan, Y., Mei, X., Liu, X., Mandic, D., ... & Plumbley, M. D. (2023). AudioLDM: Text-to-audio generation with latent diffusion models. Proceedings of the International Conference on Machine Learning, 2023

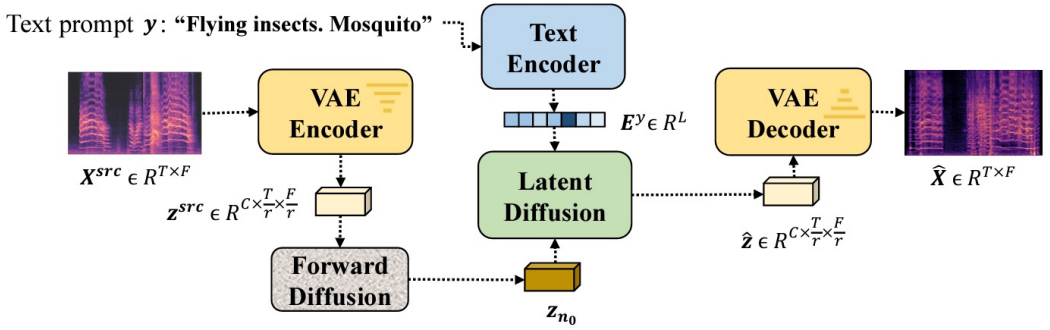
https://audioldm.github.io/

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.



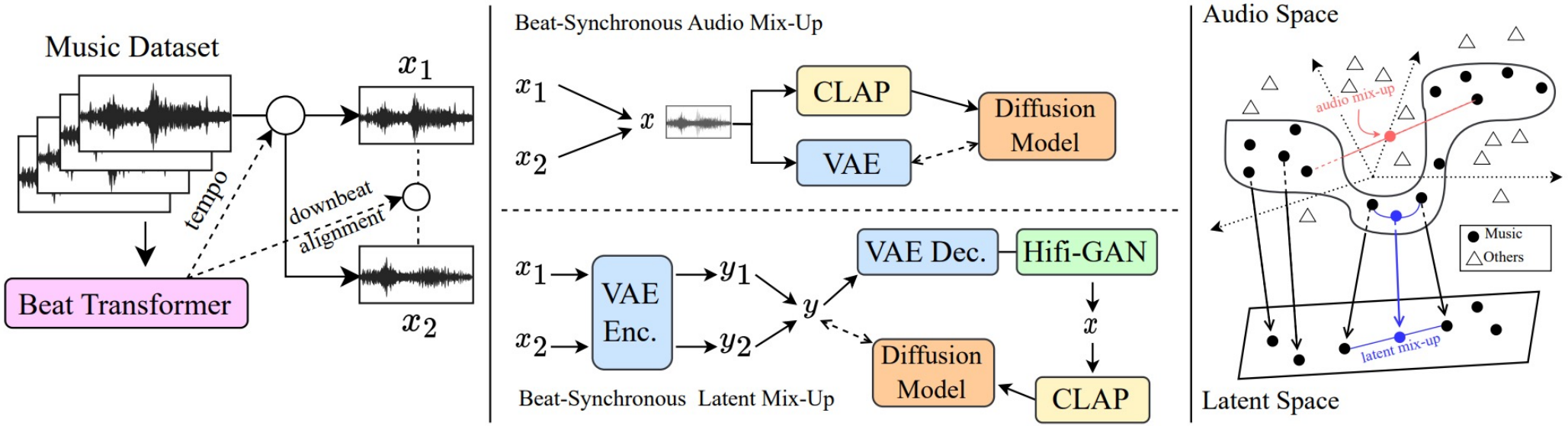
(b) Audio inpainting with AudioLDM



(c) Audio style transfer with AudioLDM

Plagiarism issue in audio generation

- Does audio generative model copy the training data?
- How to make model copy as less as possible?



Chen, K.*, Wu, Y.*, Liu, H.*, Nezhurina, M., Berg-Kirkpatrick, T., & Dubnov, S. (2023). MusicLDM: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. *arXiv preprint arXiv:2308.01546*.

https://musicldm.github.io/

LLM+Diffusion > Diffusion?

- Auto-regressive modeling:
  - ✅ Explicit modeling of temporal dependencies.
  - ✅ Enjoy the advance of recent LLM development.
  - ✅ Good in-context learning performance.
  - ❌ Long generation sequence/ lack of parallelism
  - ❌ Long range dependencies
  - ❌ Error propagation
- Diffusion-based approach:
  - ✅ Stable
  - ✅ State-of-the-art generation quality
  - ✅ Flexible formulation for manipulation, interpolation, etc.
  - ❌ Do not explicitly model temporal dependencies
  - ❌ Less flexible on duration
- Can we utilize both advantages from LLM and Diffusion?

How to combine LLM with Diffusion

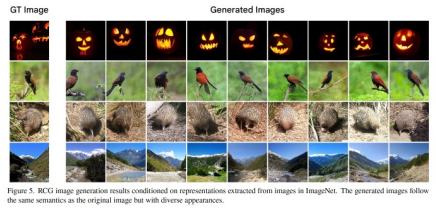
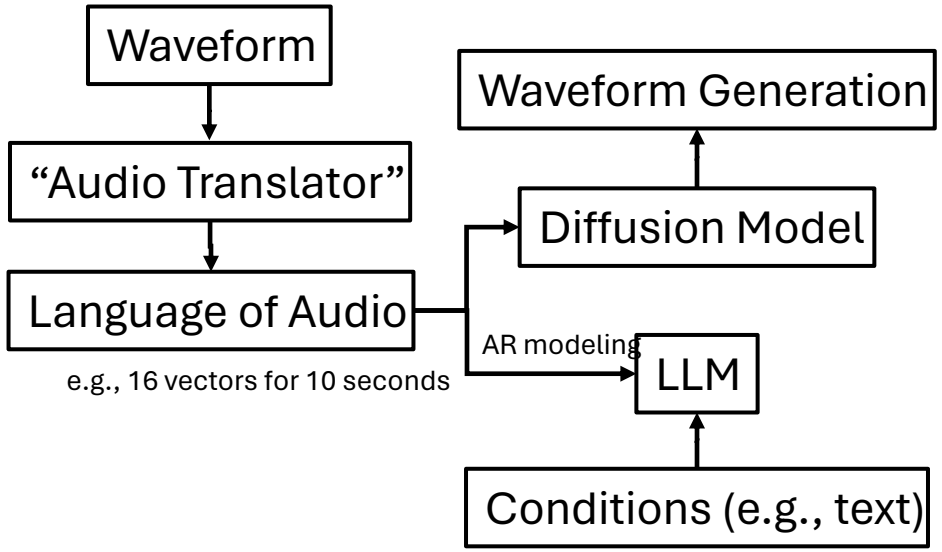


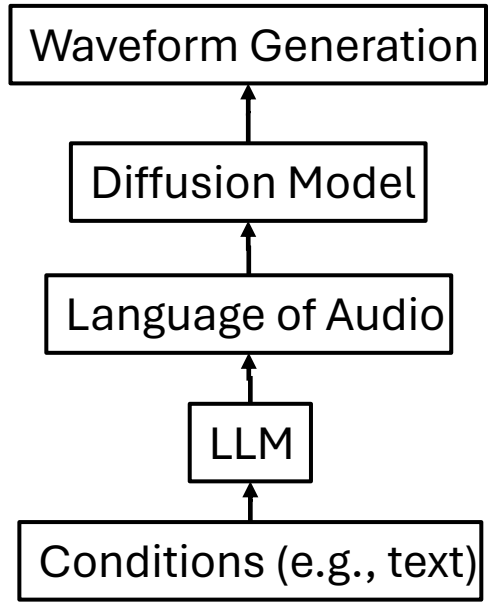
Figure 5. RCG image generation results conditioned on representations extracted from images in ImageNet. The generated images follow the same semantics as the original image but with diverse appearances.

Self-conditioned Image Generation via Generating Representations

1. AR modeling of the semantic audio sequence
2. Reconstruct semantic audio sequence to waveform



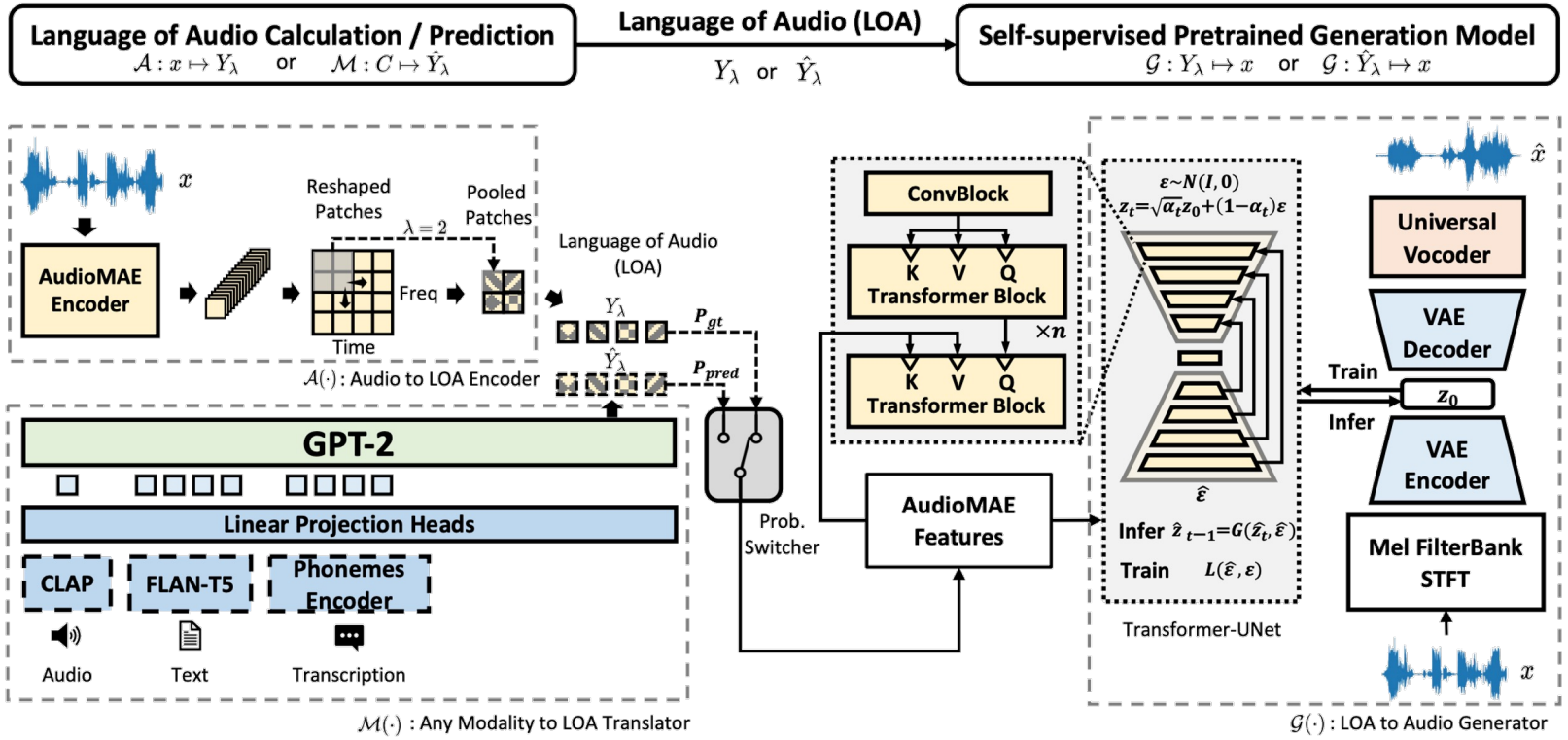
Training



Inference

Combining LLM and Diffusion

- AudioLDM 2: Combining LLM with Diffusion

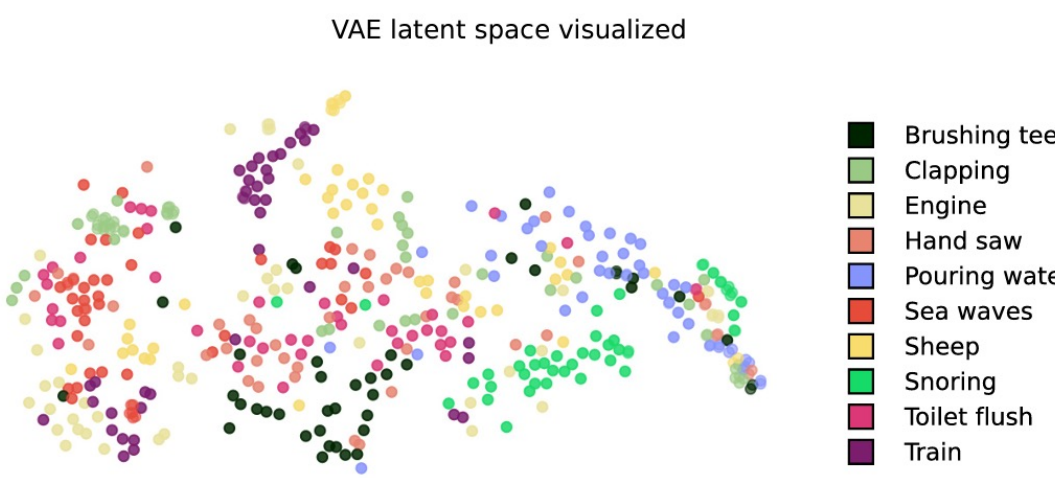


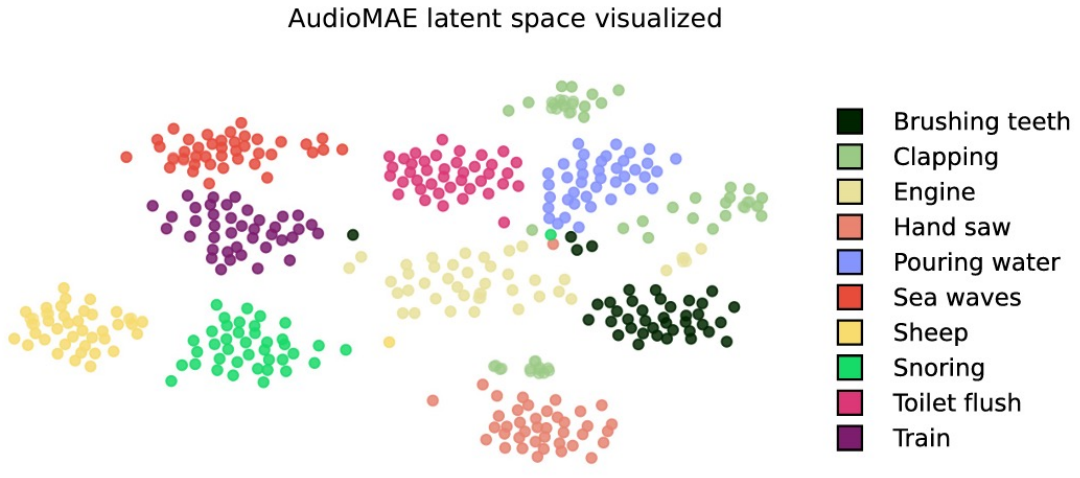
H. Liu et al., "AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, doi: 10.1109/TASLP.2024.3399607.

https://audioldm.github.io/audioldm2/

Why AudioMAE but not VAE?

- AudioMAE has a better structured latent space.
- Better represent the semantic of audio for LLM training



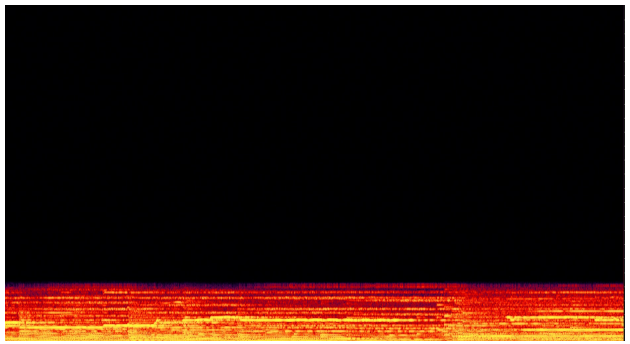


Downside of AudioLDM 2

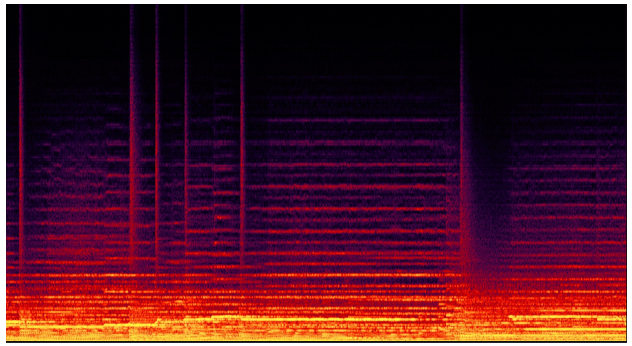
- Using GPT-2 model as “LLM”, which could be improved, e.g., LLaMA.
- Language modeling part seems to make model less creative.
  - Model tend to copy the previous token in rare cases.
- The joint finetuning of LLM and Diffusion is slow.
- Preserve the downside of LLM as well
  - On generating long audio signal, model could be unstable

Are we good enough?

- *No, at least for audio quality we are far from good.*
- Not all audio generation model works on CD quality!
  - e.g., AudioLDM-16kHz, MusicGen-32kHz, Fastspeech 2-22.05 kHz
- Not all generated samples can cover full frequency band!
- Can we build a plug-and-play module to enhance the audio quality?



✨🔮✨
Magic Module

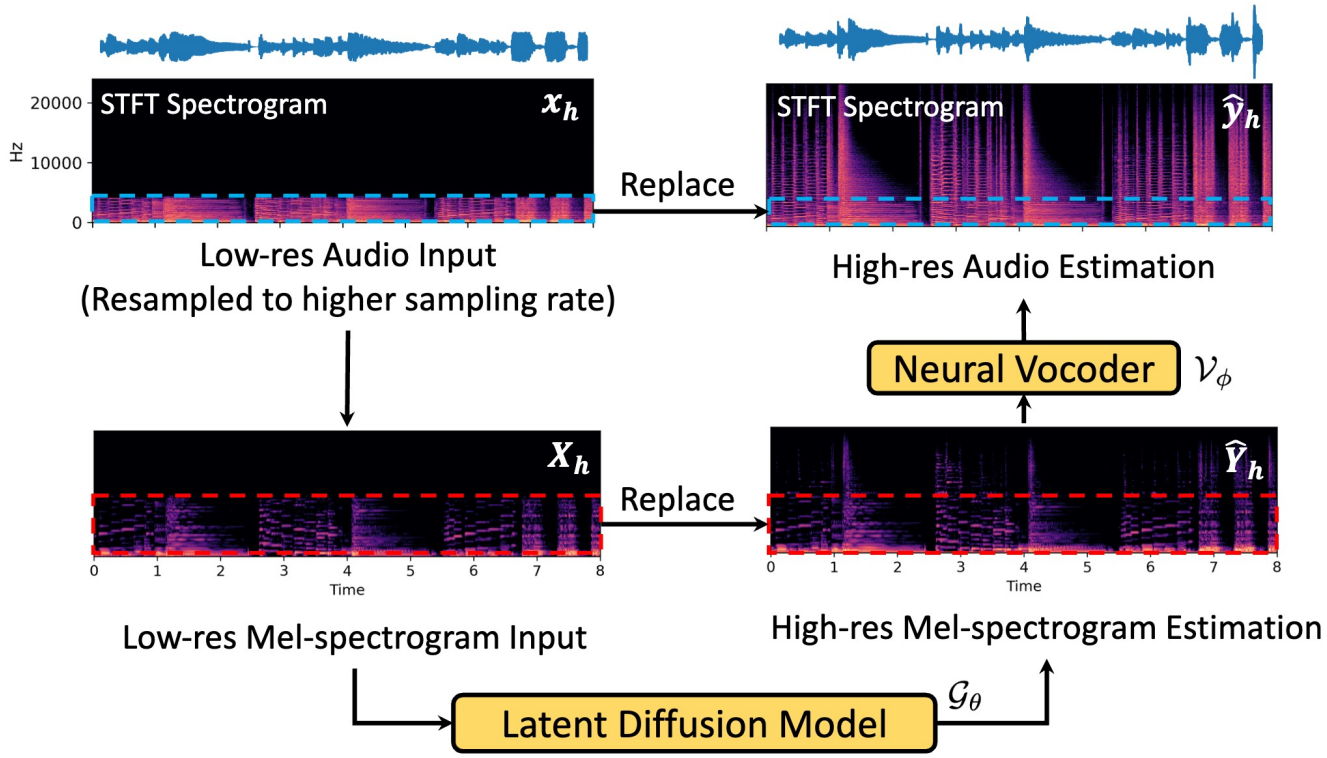


Audio Quality Enhancement

AudioSR: Versatile Audio Super-resolution

Enhance Audio Quality

- AudioSR: Versatile audio super resolution



Channel-Concatenation

Liu, H., Chen, K., Tian, Q., Wang, W., & Plumbley, M. D. (2023). AudioSR: Versatile Audio Super-resolution at Scale. *arXiv preprint arXiv:2309.07314*.

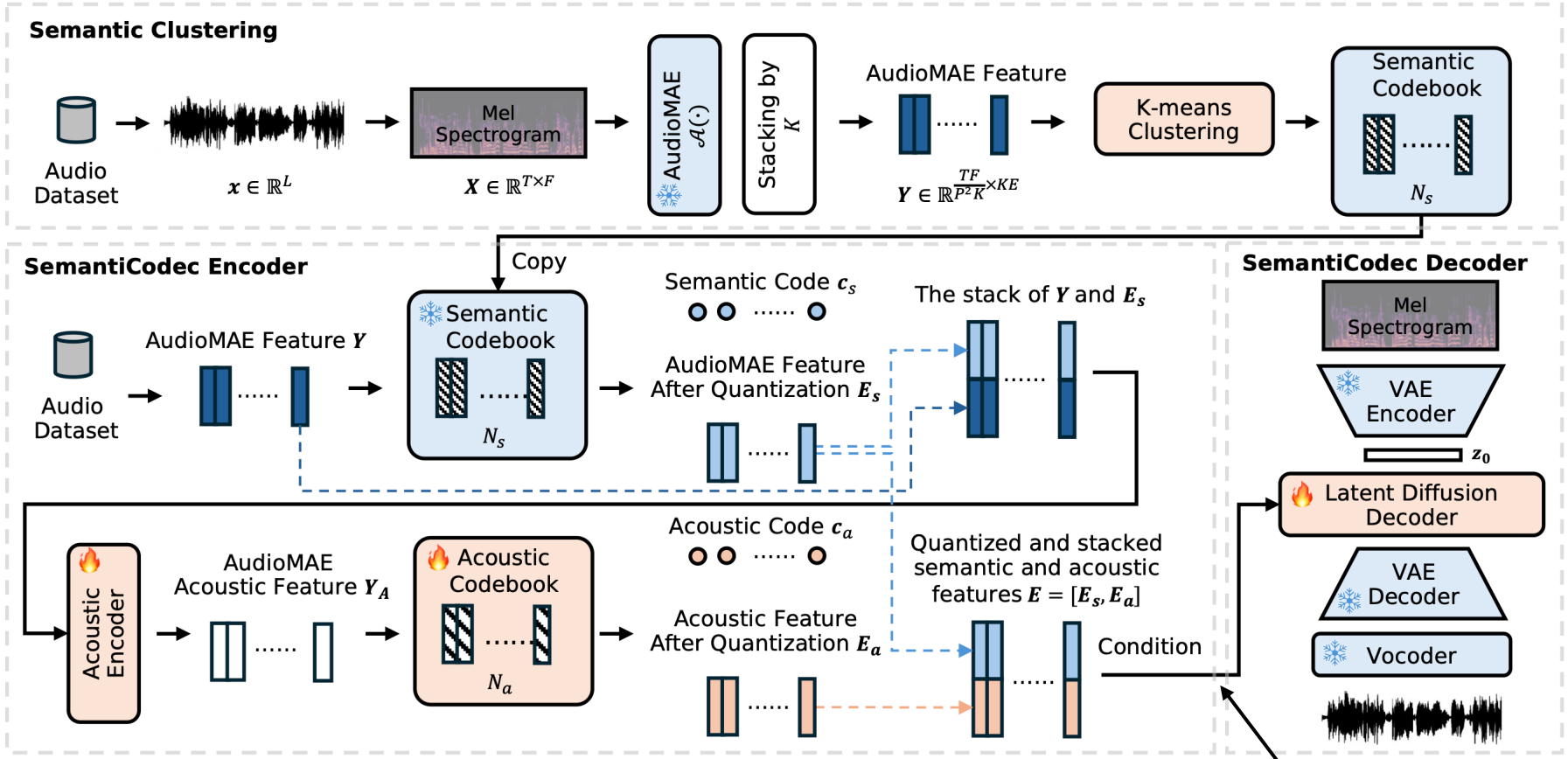
https://audioldm.github.io/audiosr/

Neural Audio Codec

Can we utilize the generation capabilities of DDPM for ultra-low rate codec?

SemantiCodec

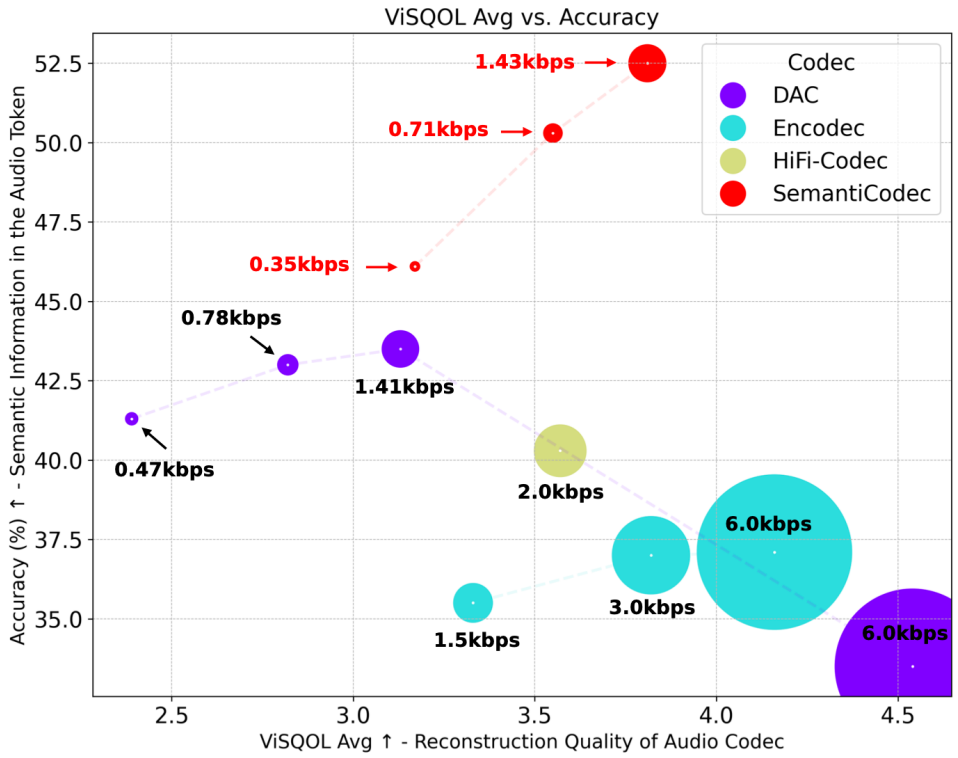
- Ultra-low bit rate (0.31 kbps ~1.40 kbps, token rate 25, 50, or 100 per second) & Strong semantics in the token & Variable vocabulary sizes

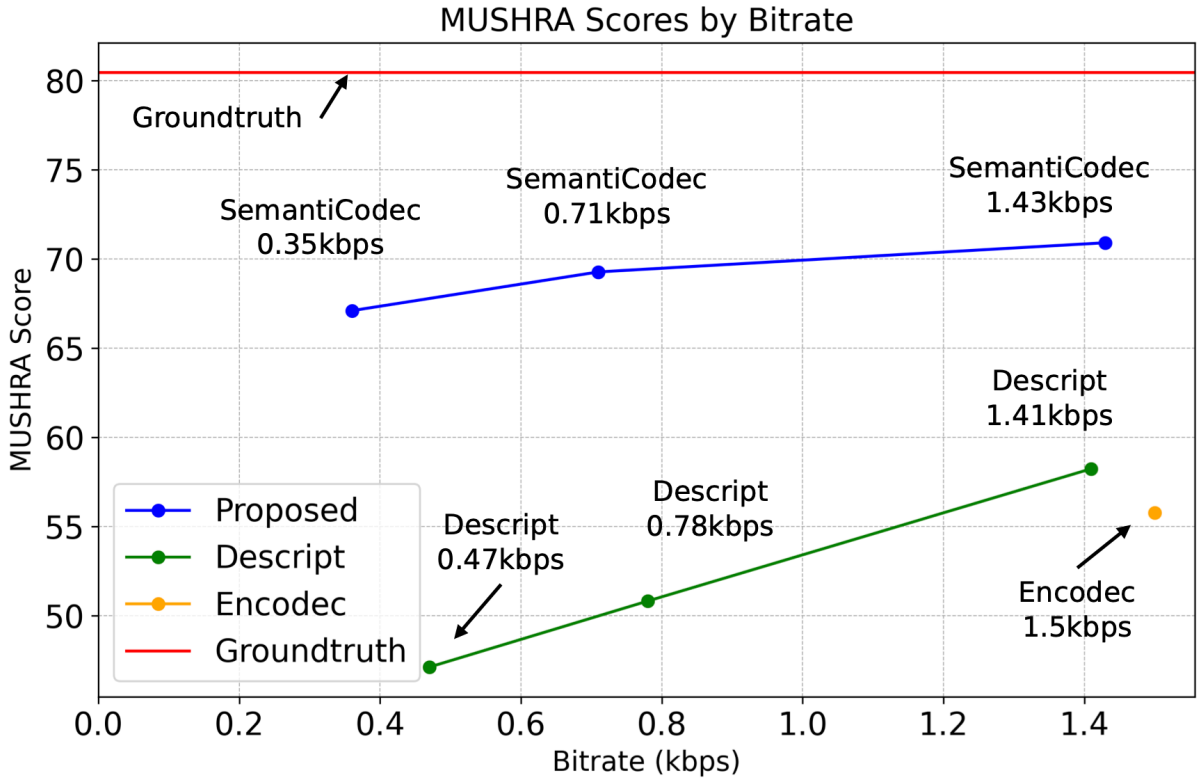


https://haoheliu.github.io/SemantiCodec/

SemantiCodec

- Better reconstruction with a lower bit rate
- Better semantic in the audio token (Potentially Better Audio LLM?)





Take aways

- Three primary types of LDM condition:
  - FiLM, Concat, CrossAtten
- Applications of conditional LDM
  - AudioLDM, AudioLDM 2, MusicLDM, AudioSR, SemantiCodec
- For more information, please visit:
  - https://haoheliu.github.io/
  - Or drop me an email haohe.liu@surrey.ac.uk