Latent Diffusion Model as a Versatile Coarse-to-Fine Audio Decoder

Haohe Liu

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

- 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.

https://haoheliu.github.io/

Outline

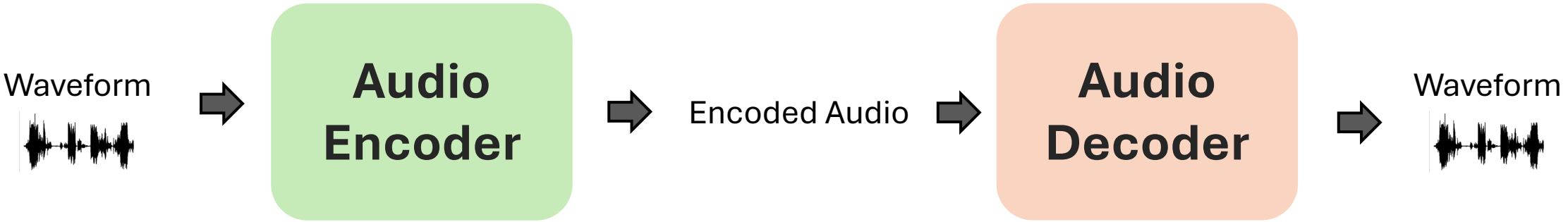
- What is an “Audio Decoder”?
- Why is Versatile Audio Decoding a difficult task?
- Latent Diffusion Model (LDM) for Versatile Audio Decoding.
- Case studies:
  - AudioLDM ½
  - AudioSR
  - SemantiCodec

  } Only different in the encoding and conditioning process.
- Other highlighted recent works
- Conclusions

What is an “Audio Decoder”

An **audio decoder** converts **encoded digital audio data** (e.g., compressed formats) back into a raw audio signal for playback or further processing.

Used in streaming services, video playback, telecommunication, etc.



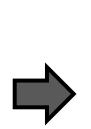
The term “Audio Decoder” in this talk is conceptually similar to this but different in how the audio is encoded

(Generalized) Audio Decoding Problems

X-to-Audio

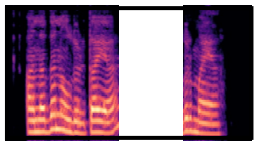
Text Descriptions
e.g., A man is giving a speech

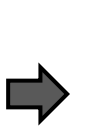
Audio Decoder



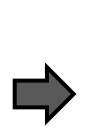
Waveform

Degraded Mel





Audio Decoder

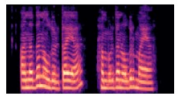


Waveform

...

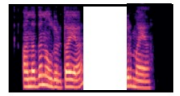
Potential Coarse Level Audio Representations

Mel Spectrogram



Text Descriptions
e.g., A man is giving a speech

Degraded Mel

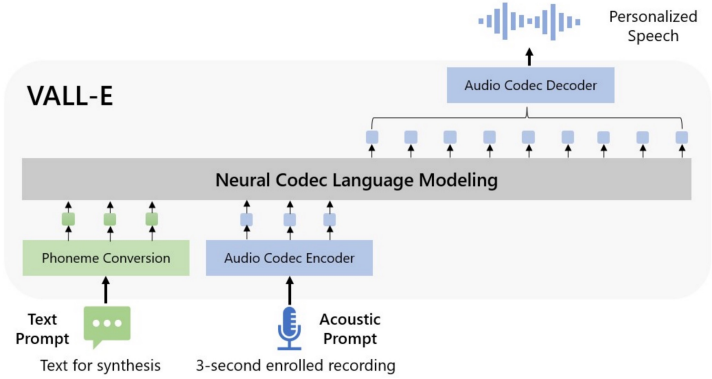


Audio Encoder
e.g., AudioMAE, Quantized Encoder

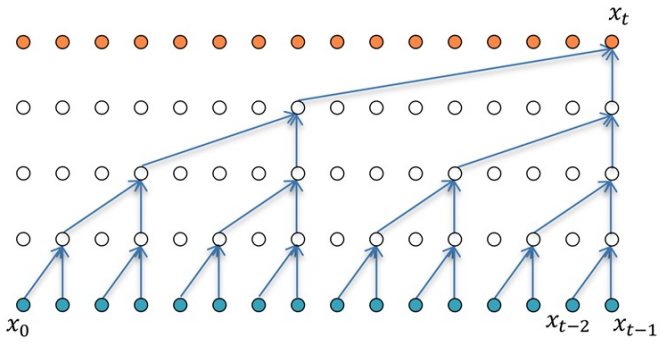
Modality Contrastive Encoder
e.g., CLAP, ImageBind

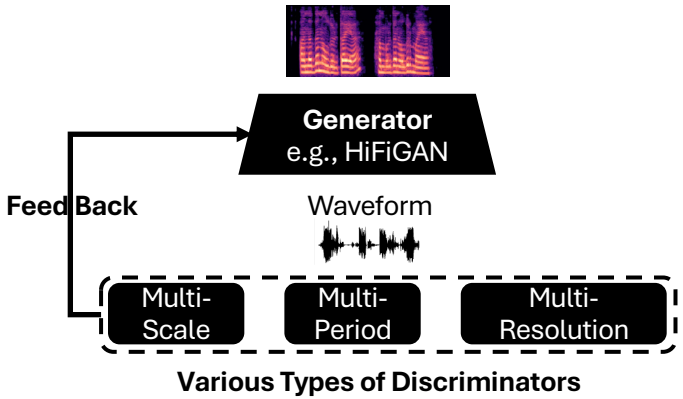
Chromagram **Pitch** **Visual Features** **RMS** **Phonemes** ...

Examples - Audio Decoder



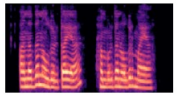
- LSTM/Transformer Decoder: Need quantization
  - VALL-E (Wang et al. 2023), AudioLM (Borsos et al. 2023), MusicGen (Copet et al. 2024), etc.
- GAN-based Decoder: Unstable training
  - For example: HiFiGAN (Kong et al. 2020), BigVGAN (Lee et al. 2022)
- VAE-based Decoder: Not diverse enough, hard to control
  - Usually for representation learning purposes
- Waveform Decoder: Extremely Slow
  - WaveNet (van den Oord et al. 2016)





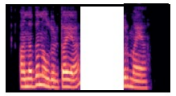
Potential Coarse Level Audio Representations

Mel Spectrogram



Text Descriptions
e.g., A man is giving a speech

Degraded Mel



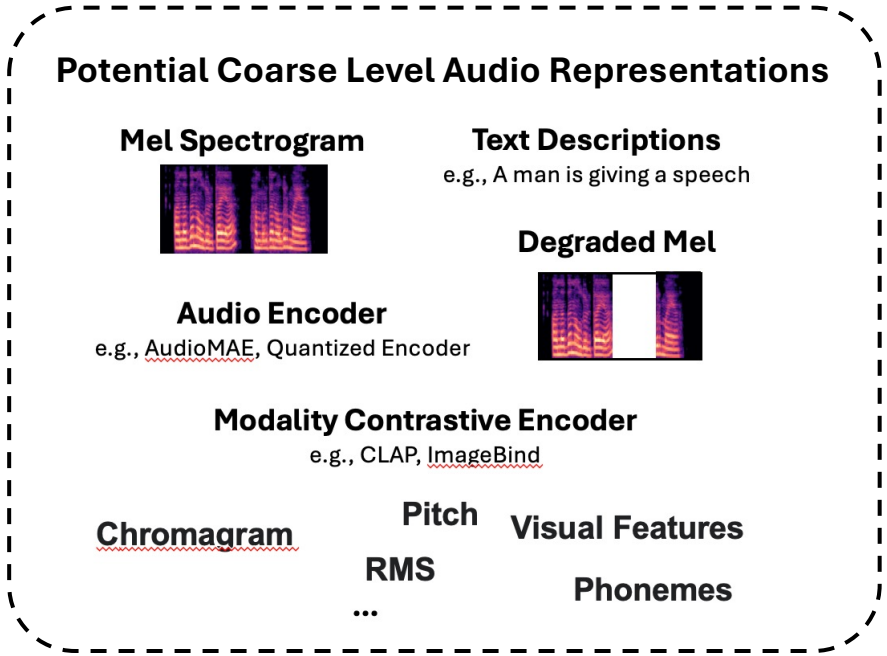
Audio Encoder
e.g., AudioMAE, Quantized Encoder

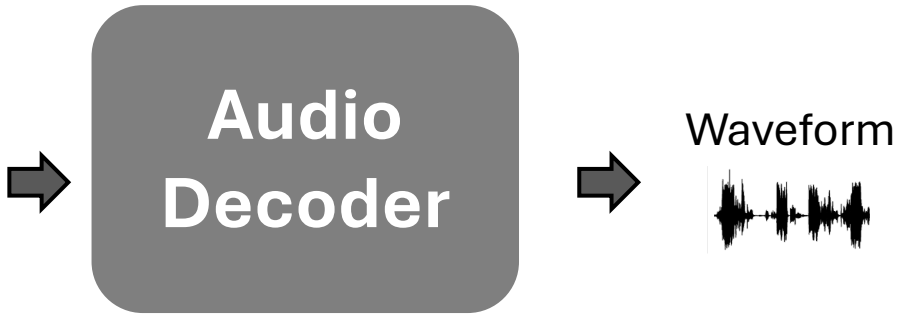
Modality Contrastive Encoder
e.g., CLAP, ImageBind

Chromagram
Pitch
Visual Features
RMS
Phonemes
...

Why is Versatile Audio Decoding Challenging?

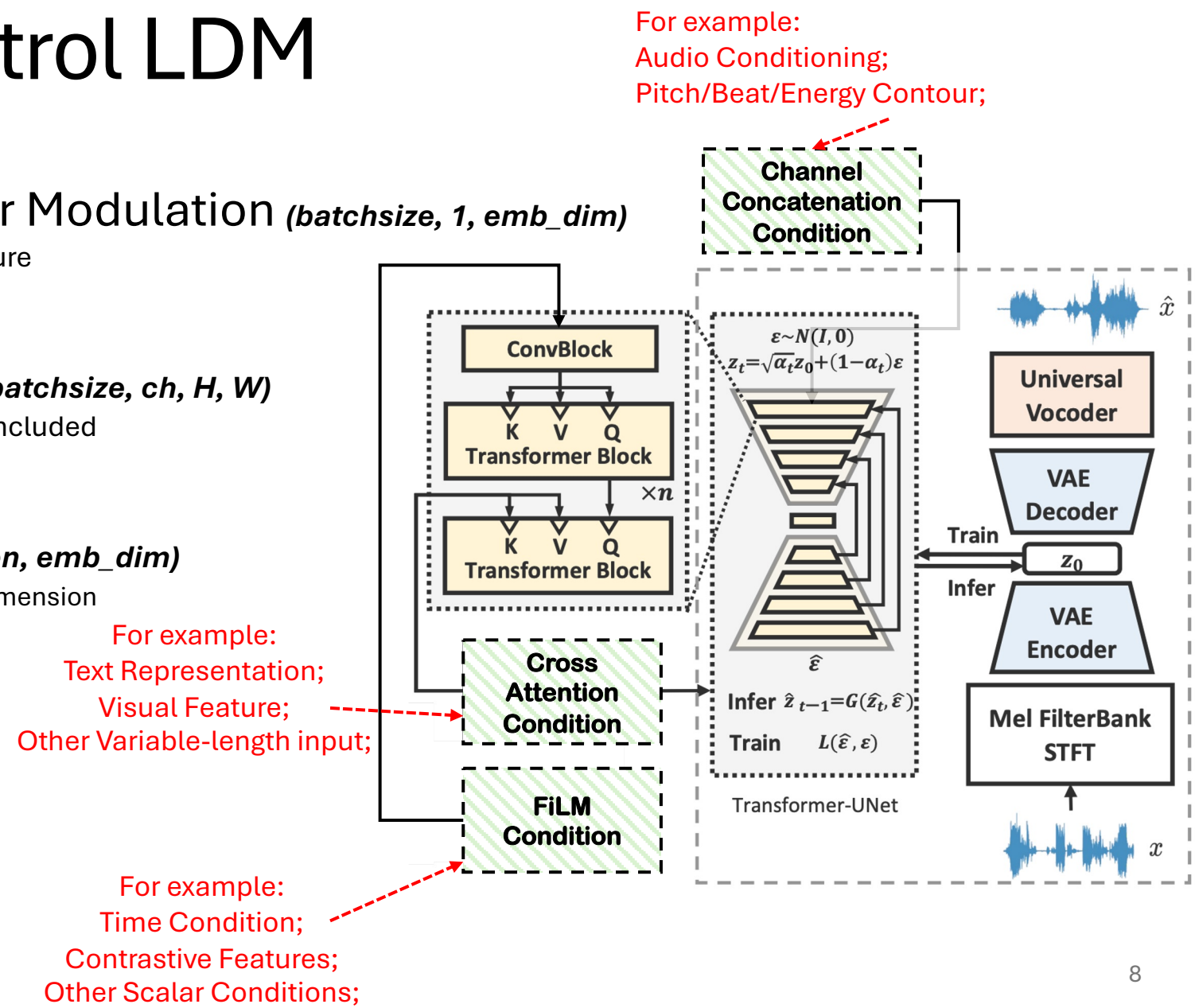
- **Need to accommodate various conditioning types**
- **Need strong modelling Capacity**





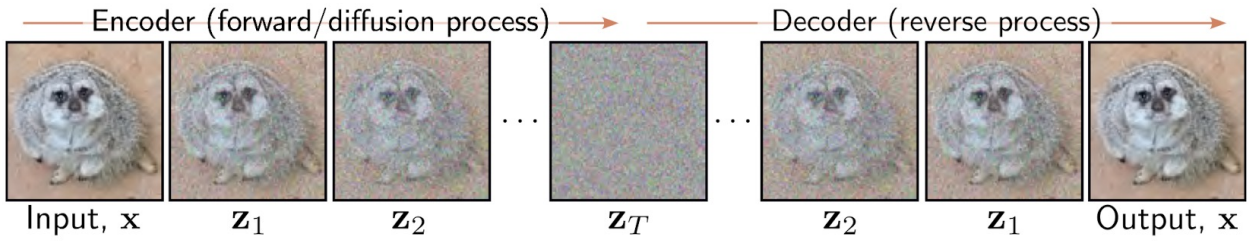
Flexibility - Control LDM

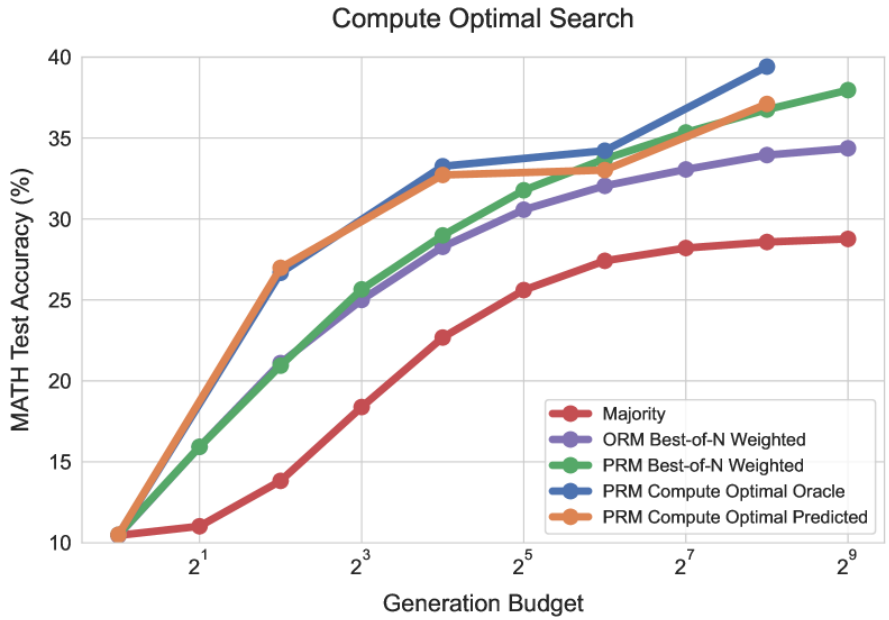
- FiLM: Feature-wise Linear Modulation ***(batchsize, 1, emb_dim)***
  - + 1D vector - Global conditional feature
  - + Marginal computational cost
  - - No temporal information
- Channel concatenation ***(batchsize, ch, H, W)***
  - + 2D tensor – Temporal Information Included
  - + Marginal computation introduced
  - - Fixed shape input
- Cross-attention ***(batchsize, len, emb_dim)***
  - + Support variable length and embed dimension
  - - Need extra attention layers
  - - Need positional encoding
- ControlNet
  - + Plug and play
  - + Configurable control strength
  - - Extra parameters
  - - Not always works well



Capacity - 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: $q(\mathbf{z}_{t-1}|\mathbf{z}_t)$
- Conceptually similar to test time scaling law (Snell et al. 2024)





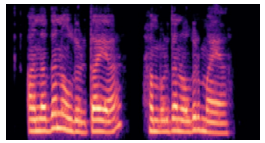
Why LDM for Versatile Audio Decoding

- **Flexibility**
  - Works for most types of conditioning
  - *FiLM, Cross Attention, Channel Concatenations*
- **Modelling Capacity**
  - Stable training & sampling on modelling data distributions
  - Not that expensive to train

Versatile Audio Decoding with LDM

Coarse Level Audio Representations

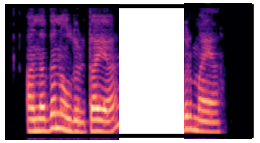
Mel Spectrogram



Text Descriptions

e.g., A man is giving a speech

Degraded Mel



Audio Encoder

e.g., AudioMAE, Quantized Encoder

Modality Contrastive Encoder

e.g., CLAP, ImageBind

Chromagram **Pitch** **Visual Features** **RMS** **Phonemes** **...**

Control the generation process

Diffusion Decoder

VAE Encoder

Diffusion Modeling

VAE Decoder

Waveform

Advantages

- **Do not need quantization**
- **Reasonably fast**
- **Stable training & inference**
- **Diverse**
- **Flexible for conditions**

Use Cases of LDM-based Audio Decoder

- **AudioLDM 1&2: Text-to-Audio Generation**
  - Decode Text to Audio
- **AudioSR: Audio Super-resolution**
  - Decode low-quality Mel-Spec to Audio
- **SemantiCodec: Neural Audio Codec**
  - Decode quantized audio codes to Audio

AudioLDM

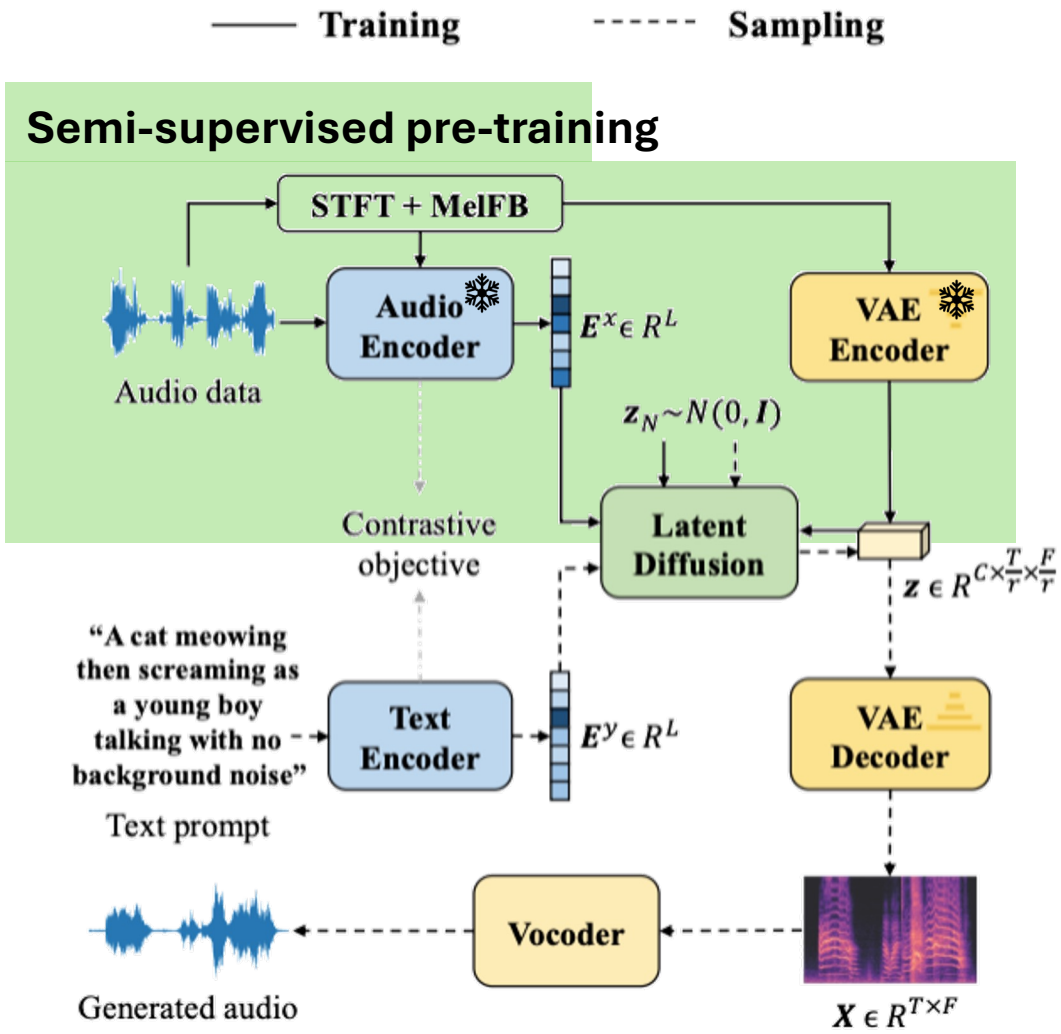
*Mode 1: CLAP Audio Emb -> **LDM -> Audio***

*Mode 2: CLAP Text Emb -> **LDM -> Audio***

*Mode 3: Text **-> LDM -> Audio***

AudioLDM

- Model the latent data distribution with the diffusion model.
- Enable the audio-only semi-supervised training approach
- Alleviate issues in LM approach such as slow inference speed, error propagation, etc.
- *CLAP Audio Emb -> LDM -> Audio*
- *Text -> LDM -> Audio*

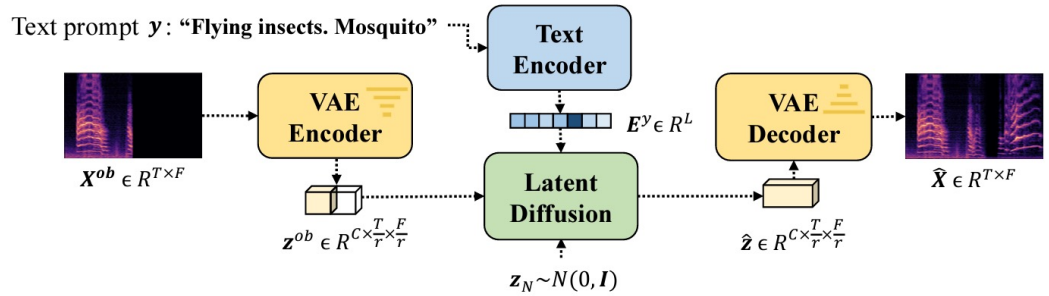


Liu, Haohe, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models." In International Conference on Machine Learning, pp. 21450-21474. PMLR, 2023.

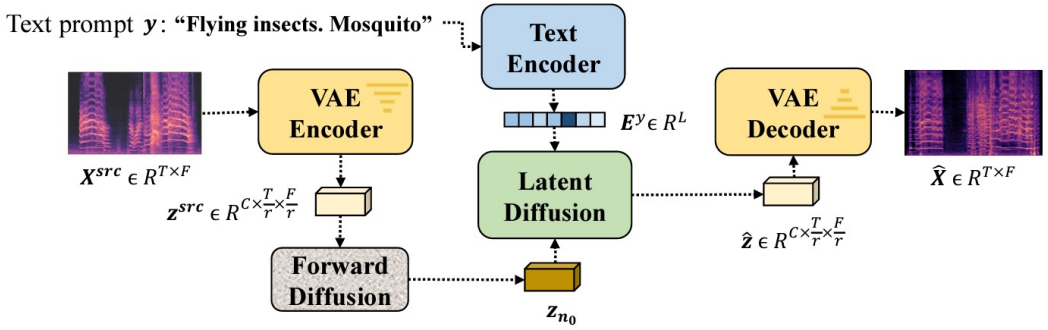
https://audioldm.github.io/

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.

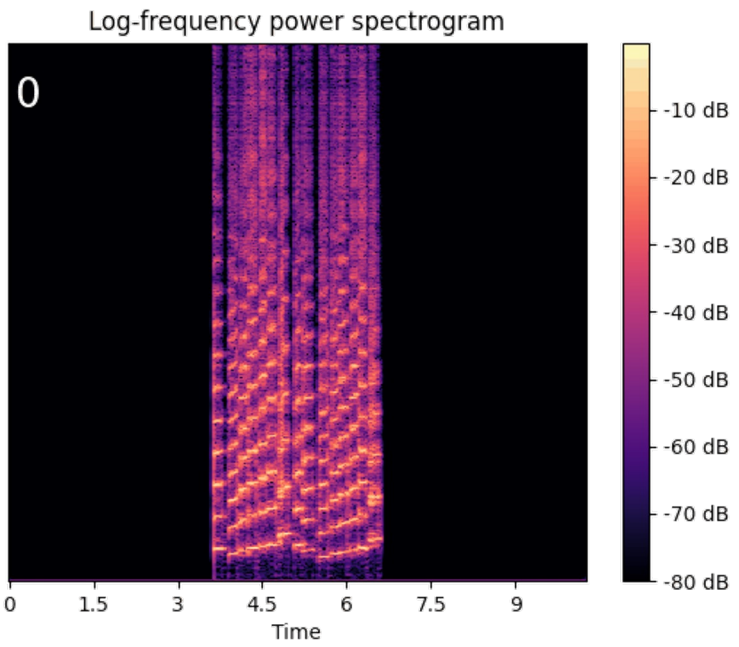


(b) Audio inpainting with AudioLDM

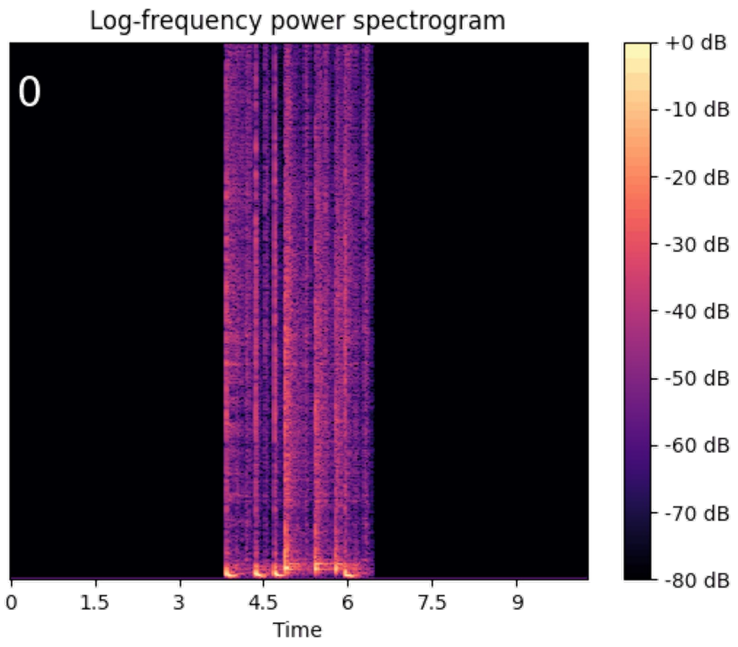


(c) Audio style transfer with AudioLDM

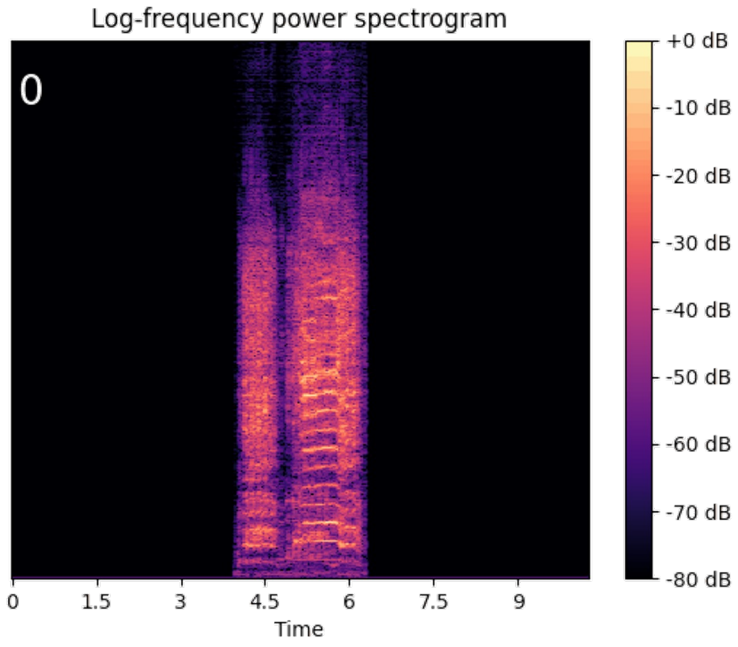
Audio Style Transfer with AudioLDM



Trumpet
→ Children Singing



Drum beats
→ Ambient Music



Sheep vocalization
→ Narration, monologue

AudioLDM 2

*Mode 1: Audio -> AudioMAE -> **LDM -> Audio***

*Mode 2: Other conditions -> AudioMAE -> **LDM -> Audio***

How to combine LLM with Diffusion

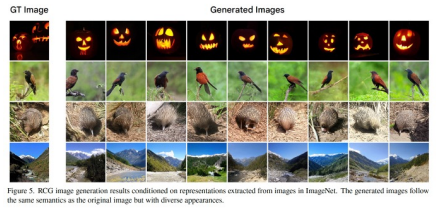
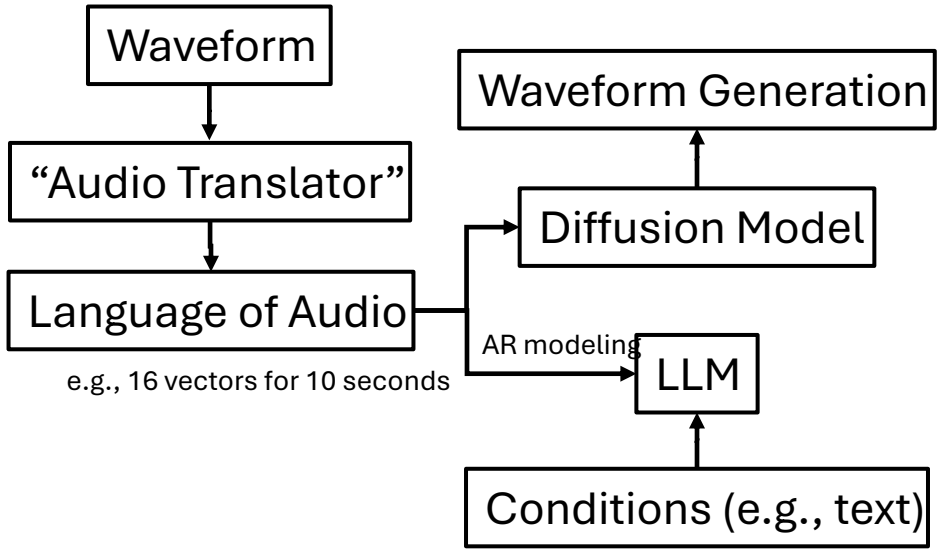


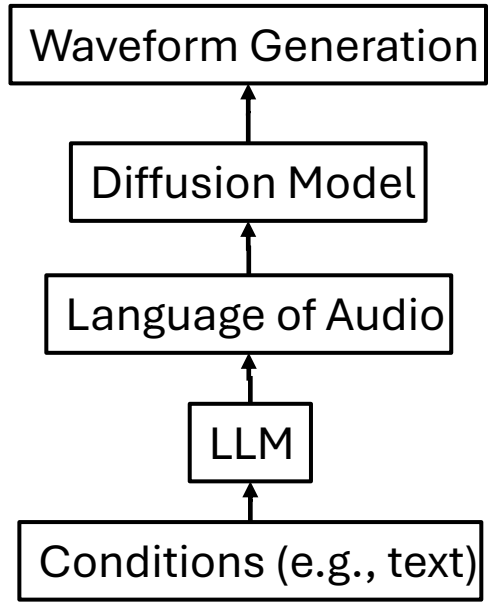
Figure 5. RCG image generation results conditioned on representations extracted from images in ImageNet. The generated images follow the same semantics as the original image but with diverse appearances.

Self-conditioned Image Generation via Generating Representations

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



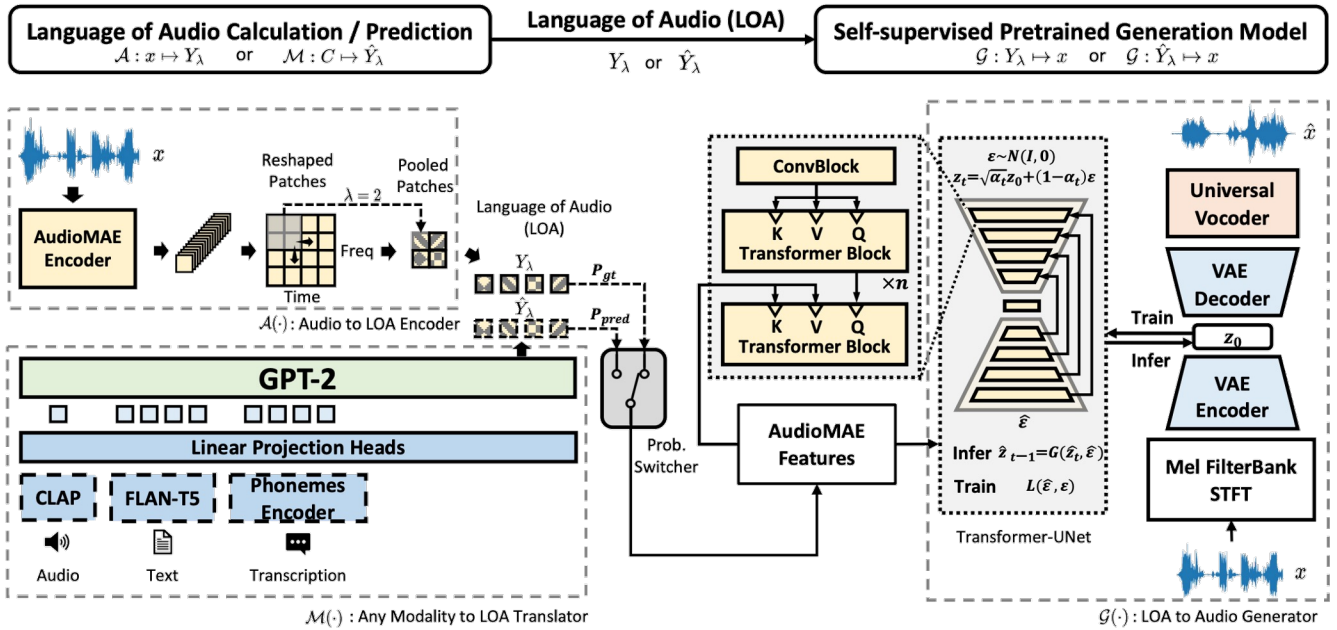
Training



Inference

Combining LLM and Diffusion

- AudioLDM 2: Combining LLM with Diffusion
- *Audio/Other conditions -> AudioMAE -> LDM -> Audio*

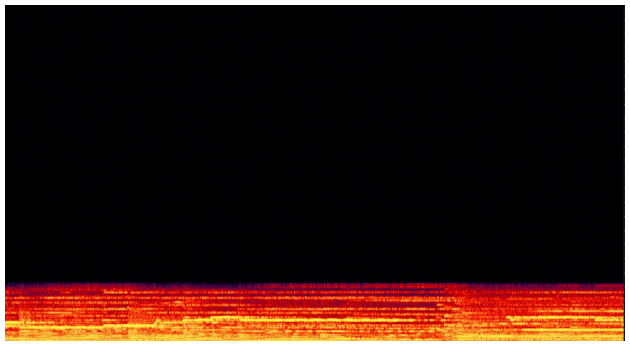


Liu, Haohe, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining." IEEE/ACM Transactions on Audio, Speech, and Language Processing (2024).

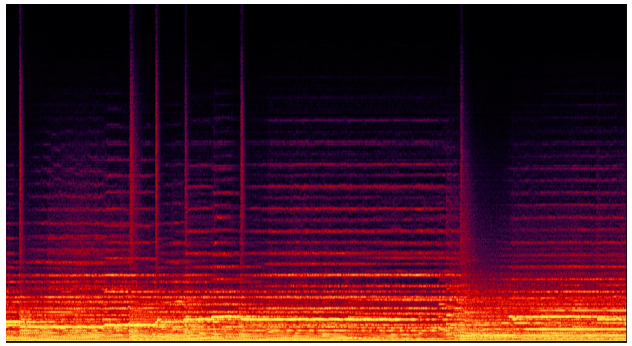
https://audioldm.github.io/audioldm2/

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



AudioSR

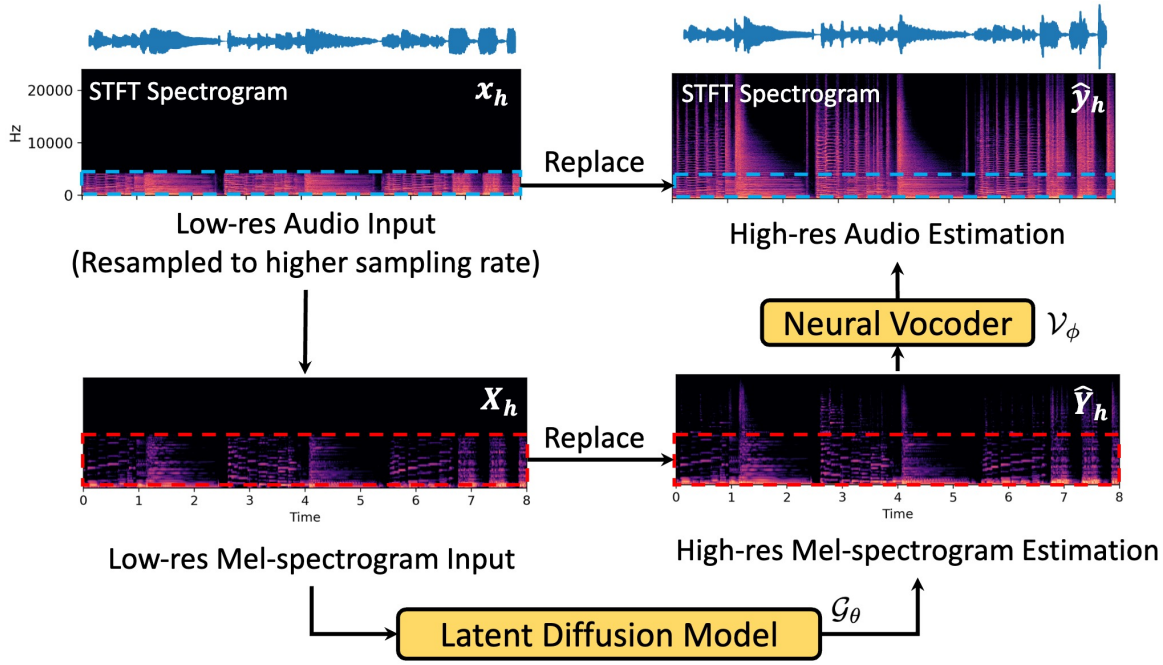
*Low-resolution Mel -> **LDM -> High-resolution Audio***

Enhance Audio Quality

Moshi: a speech-text foundation model for real-time dialogue (Défossez et al., 2024):
"The original audio is sampled at 8kHz, and we use AudioSR (Liu et al., 2023a) to upsample it to 24kHz."

- AudioSR: Versatile audio super resolution
- Low-resolution Mel -> LDM -> High-resolution Audio

Channel-Concatenation



Liu, Haohe, Ke Chen, Qiao Tian, Wenwu Wang, and Mark D. Plumbley. "AudioSR: Versatile audio super-resolution at scale." In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1076-1080. IEEE, 2024.

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

SemantiCodec

*Audio -> Compressor & Quantizer -> **LDM -> Audio***

Motivation of Building SemantiCodec

- **Long sequence**
  - e.g., 6kbps DAC has 600 tokens per second
  - Make auto-regressive modeling challenging and computational expensive
- **Bad reconstruction at low bit rate (e.g., 0.6 kbps).**
  - Most previous studies work on bit rate > 2kbps
  - Can we go further under 1.0 kbps?
- **Reconstruction quality != Easy to perform LM modeling**
  - *How to properly evaluate the easiness for LM modeling?*

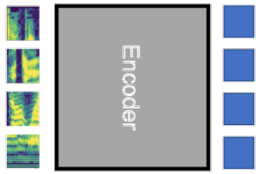
SemantiCodec

- **Limitations of previous studies:**
  - Long sequence
  - Bad reconstruction at lower bit rate (e.g., 0.6 kbps).
  - Reconstruction quality != Easy to perform LM modeling
- **Contributions**
  - Shorter sequence: Token rate 25, 50, or 100 per second.
  - Better reconstruction at lower bit rate: 0.3~1.4 kbps
  - Assumption: Semantic in the token can contribute to LM modeling
    - Achieve better semantic in the codec token
- *Audio -> Compressor & Quantizer -> LDM -> Audio*

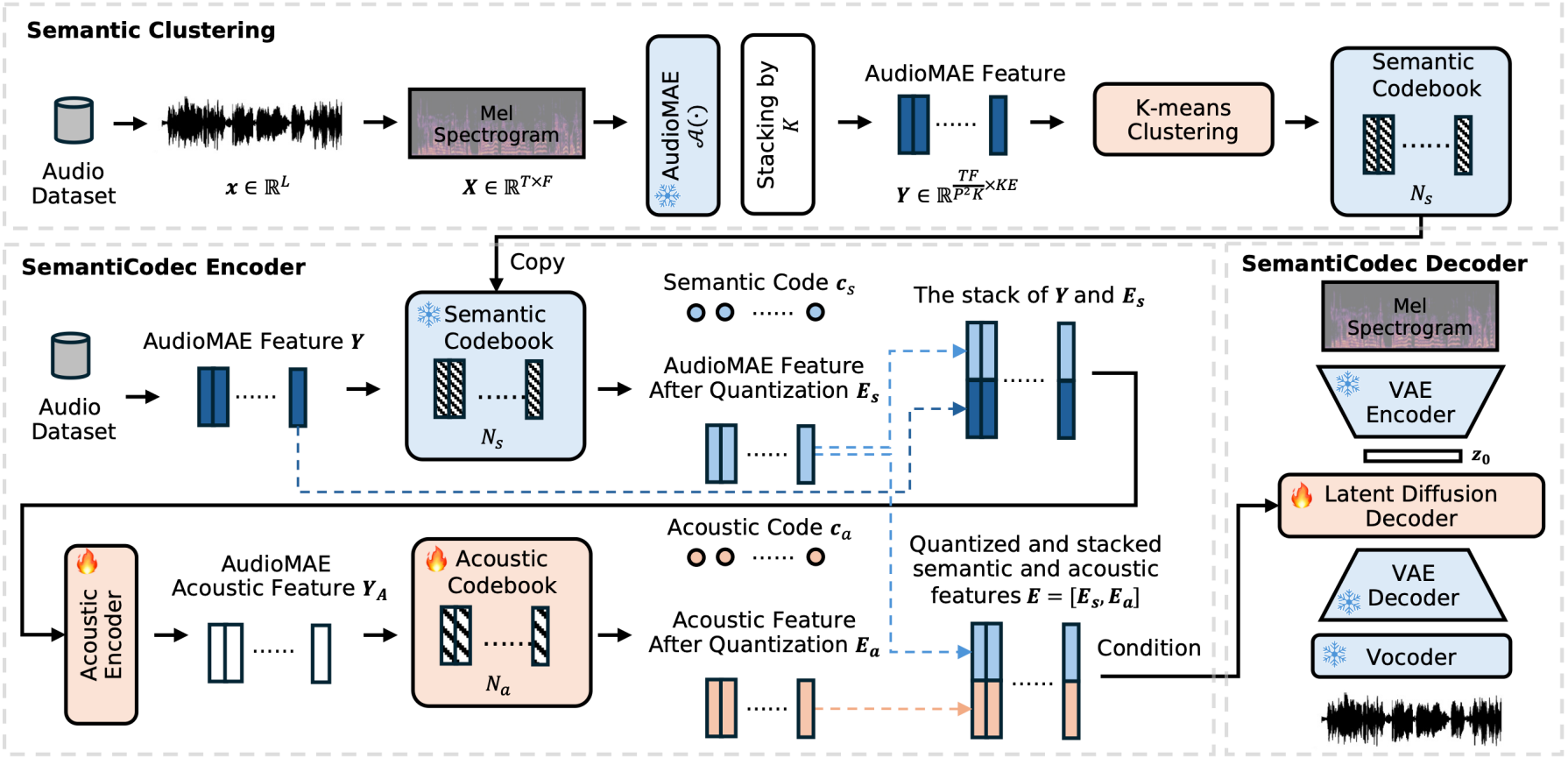
SemantiCodec

Large scale k-means is challenging
AudioSet + Million Song Dataset + GigaSpeech
https://github.com/haoheliu/kmeans_pytorch

- 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



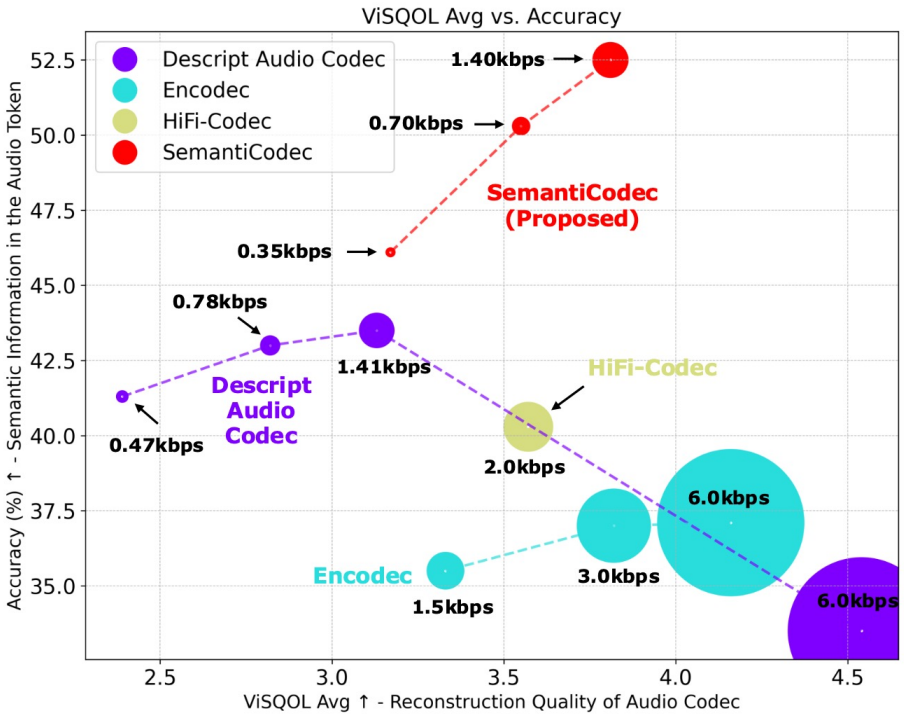
AudioMAE Encoder

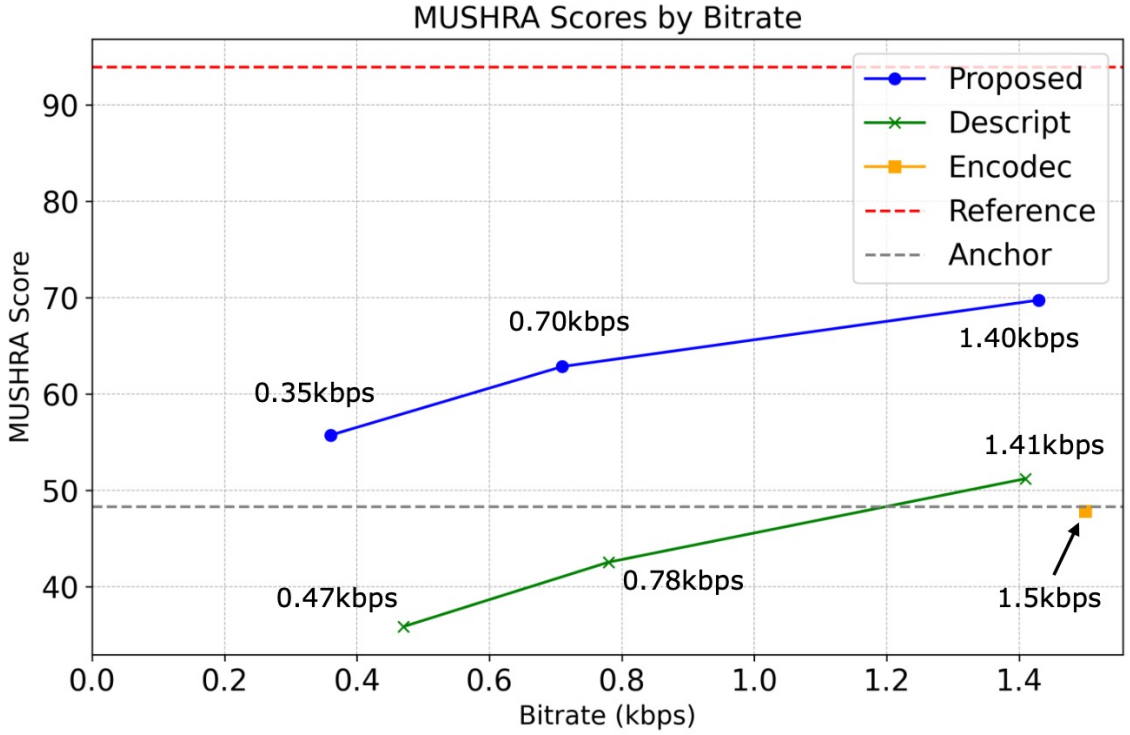


Liu, Haohe, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D. Plumbley. "SemantiCodec: An Ultra Low Bitrate Semantic Audio Codec for General Sound." IEEE Journal of Selected Topics in Signal Processing (Accepted), 2024

Visual Comparison

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





Better Reconstruction Quality

TABLE I

OBJECTIVE EVALUATION OF SEMANTICODECS AND COMPETING BASELINE CODECS AT VARIOUS BITRATES ON SPEECH, MUSIC AND GENERAL AUDIO. VIS STANDS FOR THE VISQOL METRIC.

			General Audio			Music			Speech				**Average**
Model	kBit/Sec	Token/Sec	MEL↓	STFT↓	VIS↑	MEL↓	STFT↓	VIS↑	MEL↓	STFT↓	VIS↑	WER↓	VIS↑
GroudTruth	–	-	0.0	0.0	4.99	0.0	0.0	4.99	0.0	0.0	4.99	2.09	4.99
SemantiCodec w. GT AudioMAE	–	-	3.78	3.89	4.58	3.79	3.40	4.56	3.77	3.18	4.71	2.7	4.61
SemantiCodec w.o. Acoustic VQ	0.71	50	7.29	4.93	2.43	7.67	4.44	2.61	8.68	4.58	2.78	55.6	2.61
DAC	6.00	600	2.91	3.03	4.36	2.83	2.90	4.54	2.79	2.92	4.71	3.0	4.54
Encodec	6.00	600	4.38	4.10	4.00	4.17	2.90	4.14	4.54	3.19	4.35	3.3	4.16
	3.00	300	4.84	4.26	3.58	4.67	3.11	3.78	5.06	3.40	4.10	3.7	3.82
	1.50	150	5.39	4.47	3.04	5.30	3.33	3.27	5.83	3.67	3.67	5.0	3.33
HiFi-Codec	2.00	200	4.35	3.61	3.11	4.37	3.11	3.42	3.93	2.99	4.18	3.6	3.57
DAC	0.47	47	7.56	4.58	2.12	7.80	4.24	2.33	8.62	4.70	2.73	28.2	2.39
	0.78	78	6.73	4.41	2.47	6.44	3.88	2.80	6.76	4.13	3.19	11.6	2.82
	1.41	141	6.56	4.78	2.85	6.30	3.72	3.10	6.71	3.91	3.44	5.0	3.13
SemantiCodec	0.36	25	5.06	4.02	2.84	5.22	3.76	3.18	5.77	3.72	3.49	19.6	3.17
	0.71	50	4.67	3.97	3.20	4.74	3.62	3.54	4.95	3.49	3.92	5.1	3.55
	1.43	100	4.39	3.79	3.48	4.44	3.56	3.80	4.54	3.38	4.17	3.4	3.81

Better Semantic in the Tokens

TABLE II
SEMANTIC EVALUATION OF SEMANTICODECS AND COMPETING BASELINE CODECS USING THE HEAR BENCHMARK.

VQ setting	Model	kBit/Sec	Token/Sec	ESC-50	NSPitch	SPC	LbCount	CRM-D	VoImit	Average ACC
Unquantized	AudioMAE	-	-	79.5	82.0	48.0	69.4	67.3	17.4	60.6
All VQ layers	Encodec	6.00	600	40.7	60.8	27.3	45.0	44.2	4.4	37.1
	HiFi-Codec	2.00	200	36.3	71.0	26.5	58.2	45.6	4.3	40.3
	DAC	6.00	600	33.4	56.1	21.0	46.4	39.9	3.4	33.4
		1.41	141	41.1	**80.9**	30.3	59.5	44.7	4.8	43.5
		0.78	78	39.5	78.5	30.3	59.5	45.7	5.0	43.0
		0.47	47	36.7	75.7	26.4	59.6	44.5	4.9	41.3
	SemantiCodec	1.43	100	**63.8**	73.3	**43.6**	67.0	**57.9**	**9.6**	**52.5**
		0.71	50	60.9	64.9	41.9	**71.6**	53.2	9.3	50.3
		0.35	25	56.4	61.3	33.7	70.4	46.9	8.0	46.1
First VQ Layer	Encodec	0.75	75	32.0	45.3	23.0	44.8	40.7	4.2	31.6
	HiFi-Codec	1.00	100	33.7	58.9	25.9	58.3	44.3	4.1	37.5
	DAC (6.00k)	0.5	50	23.8	23.3	15.9	43.1	38.4	3.1	24.6
	DAC (1.41k)	0.16	16	29.0	44.2	19.9	57.0	39.5	4.0	32.3
	DAC (0.78k)	0.16	16	27.6	39.9	17.9	56.6	40.9	4.0	31.1
	DAC (0.36k)	0.16	16	29.7	46.9	18.3	58.2	43.0	4.0	33.3
	SemantiCodec	0.71	50	**66.6**	**73.9**	**42.7**	**66.7**	**57.5**	**11.1**	**53.1**
		0.36	25	64.4	70.2	36.0	65.0	54.1	10.7	50.1
		0.18	13	59.6	66.3	30.7	61.3	45.8	9.8	45.6

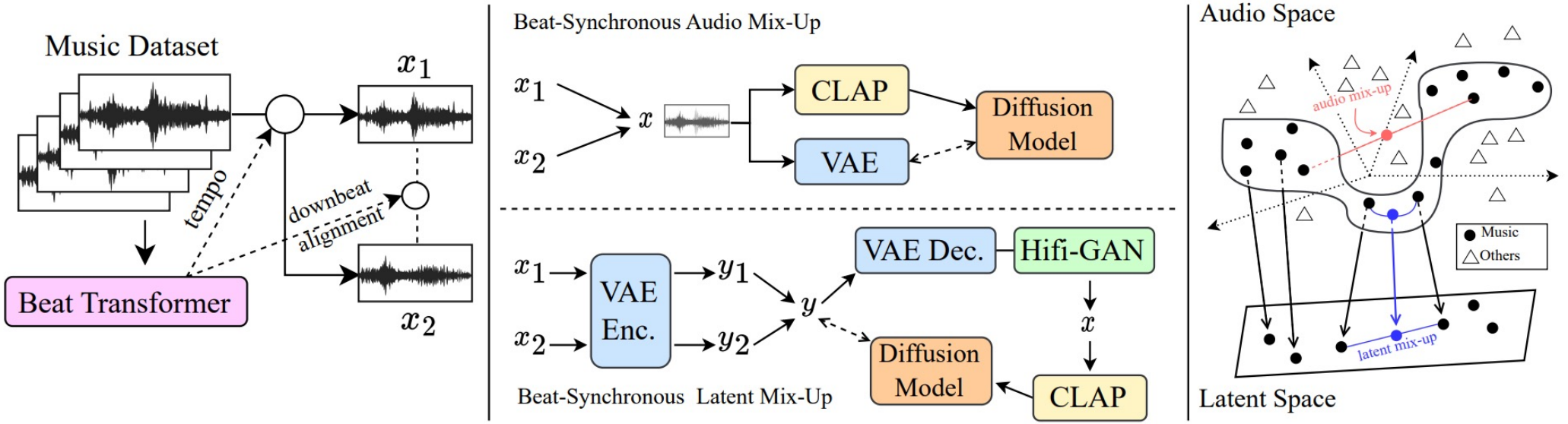
Significantly better when using only a single codebook

Other Highlighted Recent Works

1. Similar Audio Decoder Problems: MusicLDM, FlowSep, FlashSpeech
2. Effort on Data Curation: WavCaps, AudioSetCaps

Plagiarism issue in Text-to-Music generation

- Music generative model does copy the training data.
- Beat-Synchronous Mix-up make model copy less.

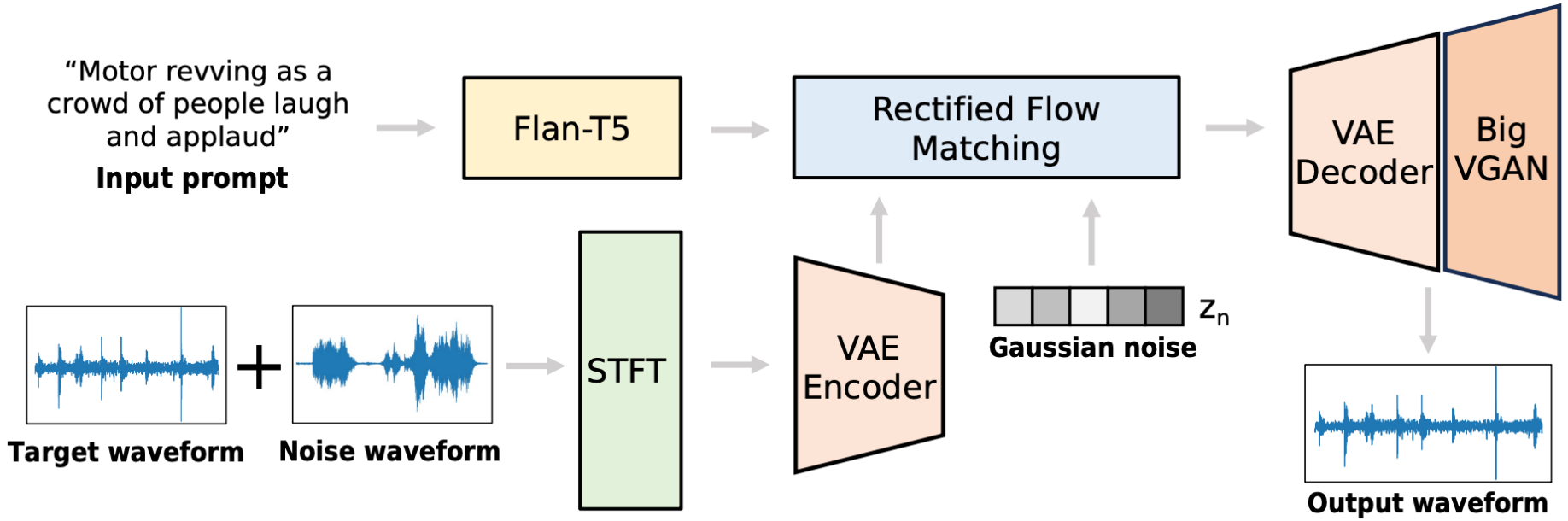


Chen, Ke*, Yusong Wu*, Haohe Liu*, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov.
"MusicLDM: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies."
In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1206-1210, 2024.

https://musicldm.github.io/

FlowSep: Language-Queried Source Separation

- Conditioning Audio and Text Query to the RFM.

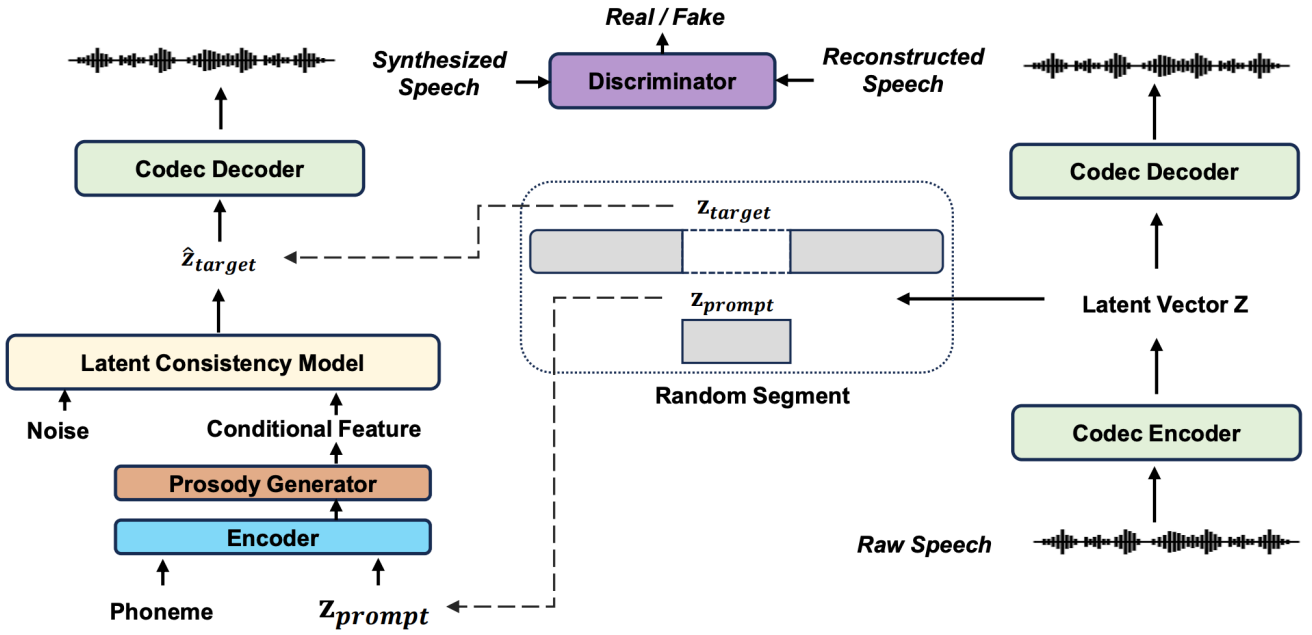


Yuan, Yi, Xubo Liu, Haohe Liu, Mark D. Plumbley, and Wenwu Wang.
"FlowSep: Language-Queried Sound Separation with Rectified Flow Matching."
arXiv preprint arXiv:2409.07614 (2024).

https://audio-agi.github.io/FlowSep_demo/

FlashSpeech: Latent Consistency Model

- Zero-shot text-to-speech with latent consistency model

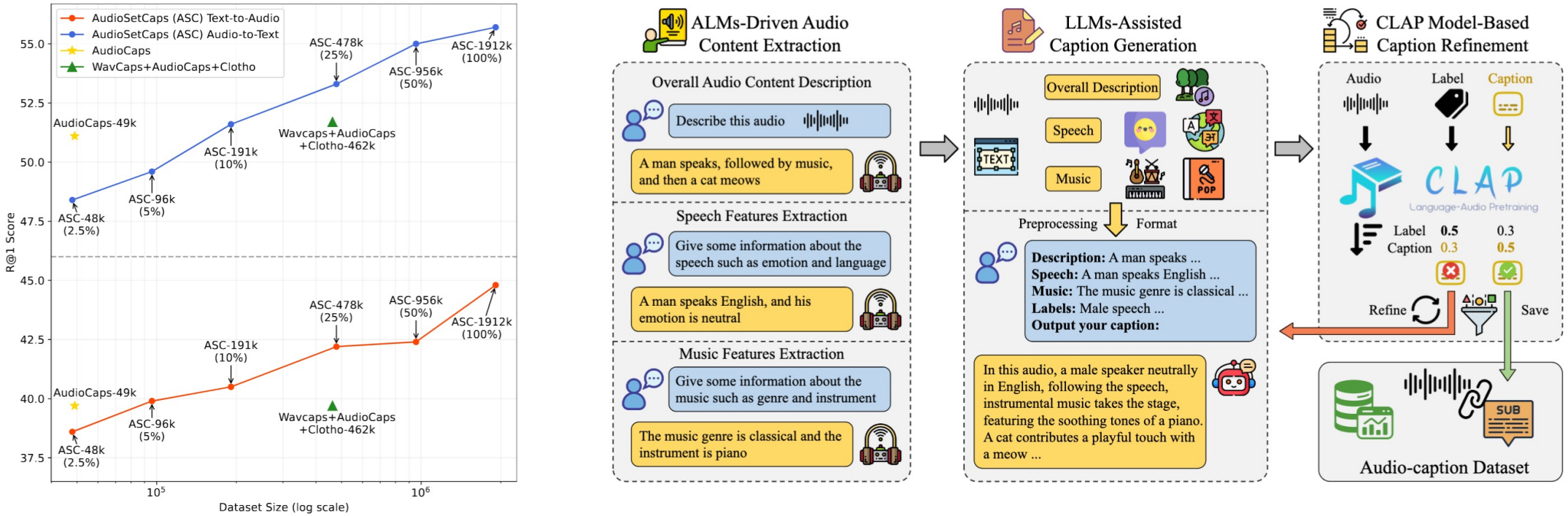


Ye, Zhen, Zeqian Ju, Haohe Liu, Xu Tan, Jianyi Chen, Yiwen Lu, Peiwen Sun et al. "Flashspeech: Efficient zero-shot speech synthesis." In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 6998-7007. 2024.

https://flashspeech.github.io/

AudioSetCaps: 6M Audio-Language Pairs

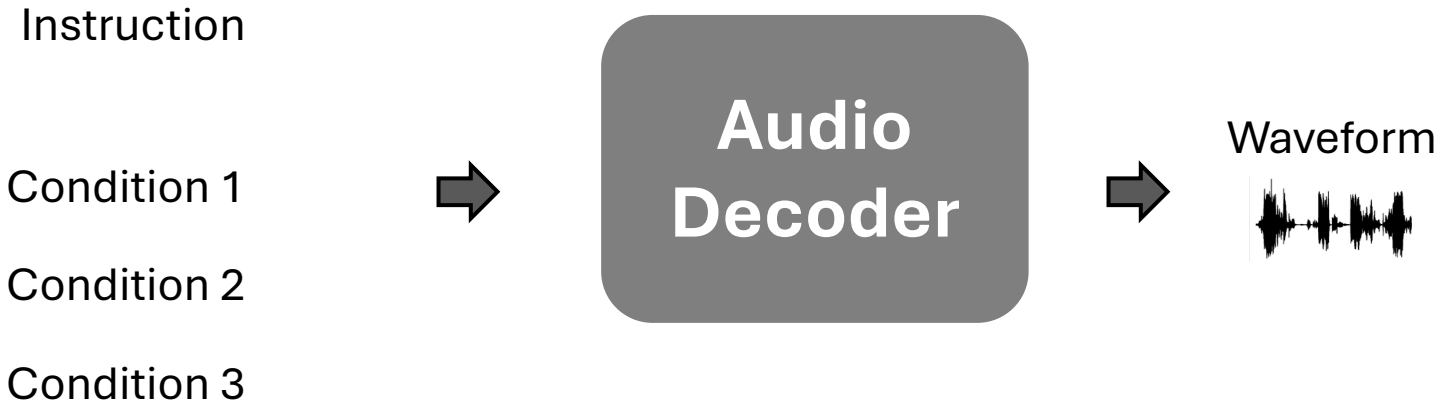
- Build Synthetic Caption for 6M audio with ALM and LLM



Bai, Jisheng*, Haohe Liu*, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D. Plumbley, Woon-Seng Gan, and Jianfeng Chen.
"AudioSetCaps: Enriched Audio Captioning Dataset Generation Using Large Audio Language Models."
Paper presented at Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation, 2024.

https://github.com/JishengBai/AudioSetCaps

Look Ahead – *Large Audio Diffusion Model?*



Scaling Law?
SSL Pretraining? (e.g., unconditional generation)
Instruction Finetuning?
Zero-shot Abilities?

Thanks

- If your model target space is audio, consider using LDM.
- Training LDM as audio decoder:
  - https://github.com/haoheliu/AudioLDM-training-finetuning
- Evaluate LDM audio decoder performance:
  - https://github.com/haoheliu/audioldm_eval
- Open-sourced AudioLDM, AudioSR, and SemantiCodec:
  - https://github.com/haoheliu

https://haoheliu.github.io/