

Latent Diffusion Models and “Language of Audio” in Generative Audio

Ethan Graybar

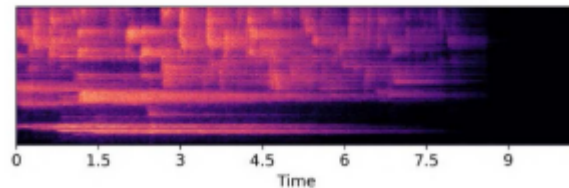
grayb031@morris.umn.edu

Introduction

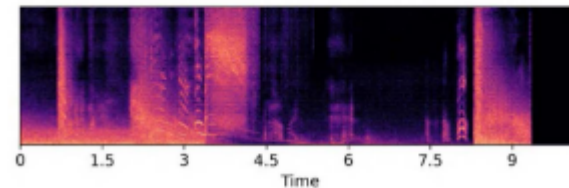
AudioLDM 2

- AI generative model designed to create audio
- Introduces “Language of Audio” as universal input translator
- Uses a latent diffusion model to generate data
- Offers high quality results at an efficient rate

Magical fairies laughter echoing through an enchanted forest.



A monkey laughs before getting hit on the head by a large atomic bomb.



Talk Outline

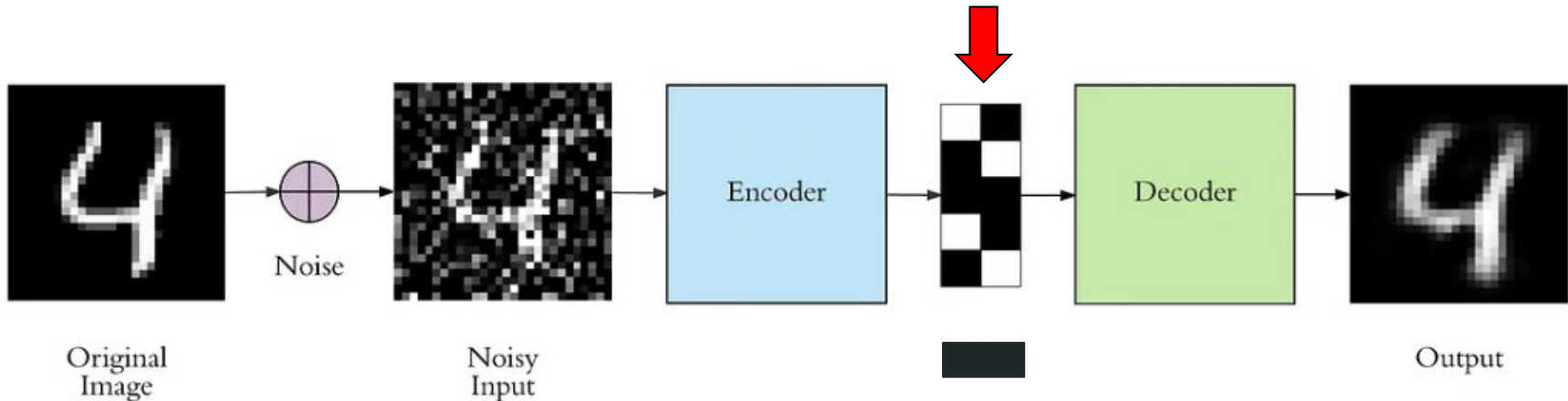
1. Background
2. Latent Diffusion Models
 - a. GPU Optimization
3. Language of Audio
4. GPT-2 Model
5. Results
6. Conclusion

Background

Autoencoder Architectures

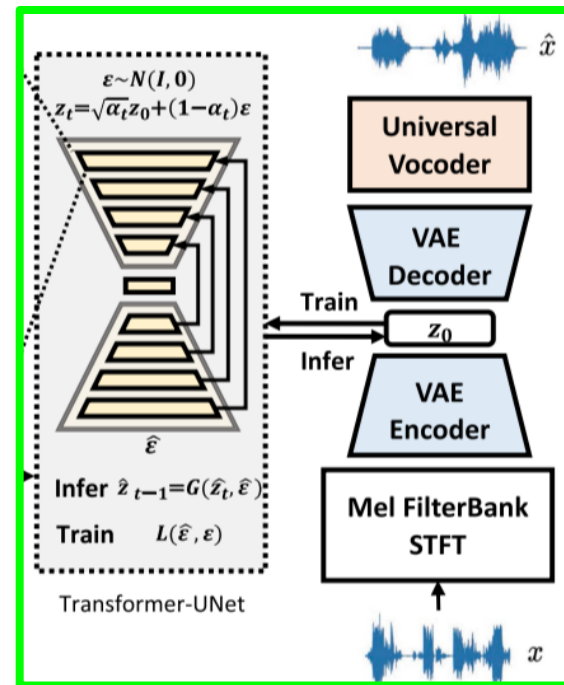
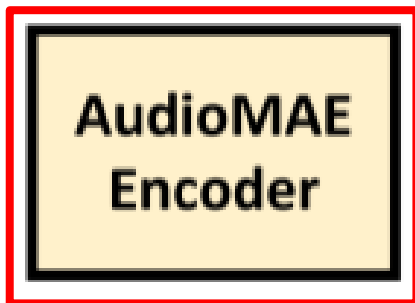
- Recreates an input without access to the input
- Learn important information about input data

Lower-dimensional latent space



Self-supervised pretrained

- Self-supervised: a system that learns without being given **labels**
- Pretrained: model is trained on a dataset in advance



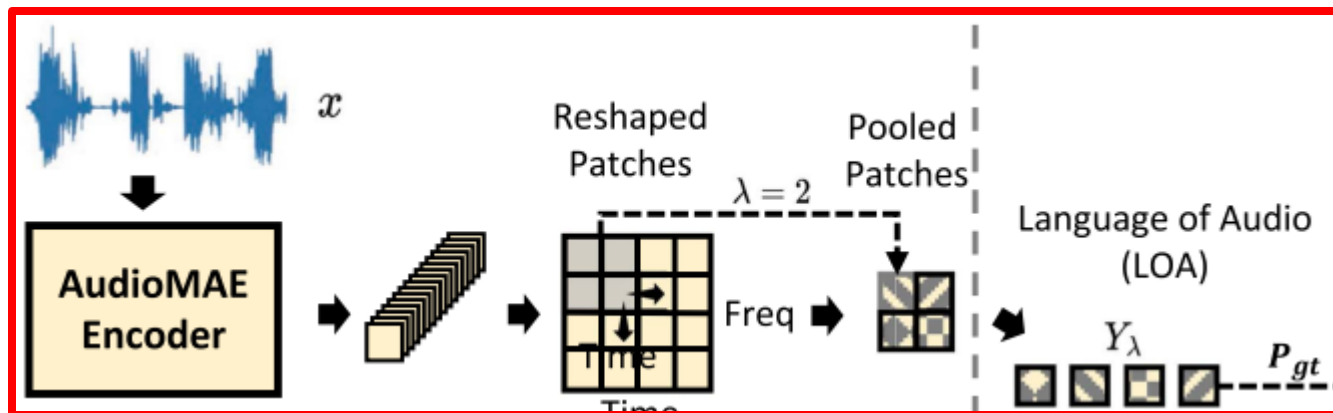
Language of Audio

Language of Audio

- Represents important information from an input audio sample
 - Acoustic: frequency (pitch), amplitude (volume), etc.
 - Semantic: the meaning of audio
- LOA is the product of **AudioMAE**

Audio to LOA Encoder

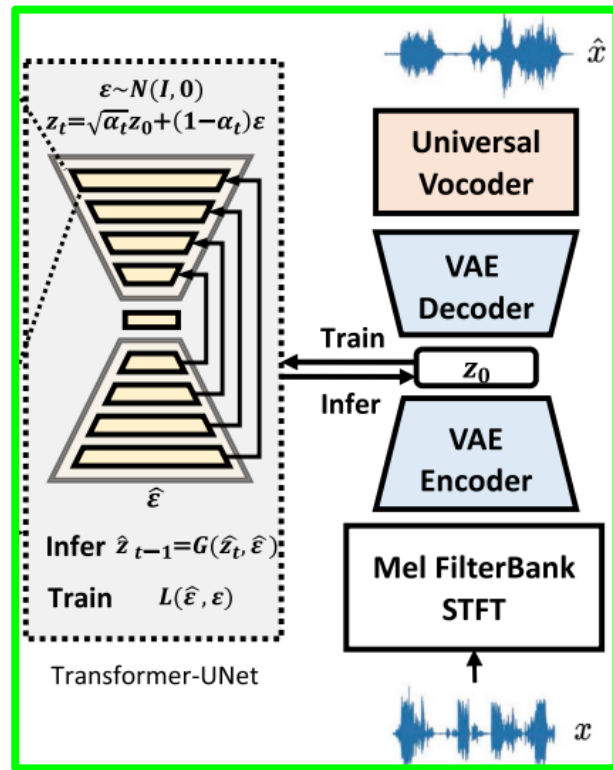
1. Takes an audio sample
2. AudioMAE Encoder converts the sample into a **mel spectrogram**
3. Mel spectrogram image is split into patches
4. Patches are encoded with noise - masked patches
5. Masked patches are decoded resulting in LOA



Latent Diffusion Models

Latent Diffusion Models

- Generate the audio output through the **diffusion process**
- Three stages of the diffusion process:
 - a. Encoder adds noise to the input data
 - b. Input data is compressed into a lower dimensional latent space
 - c. Decoder creates a representation based on the input data



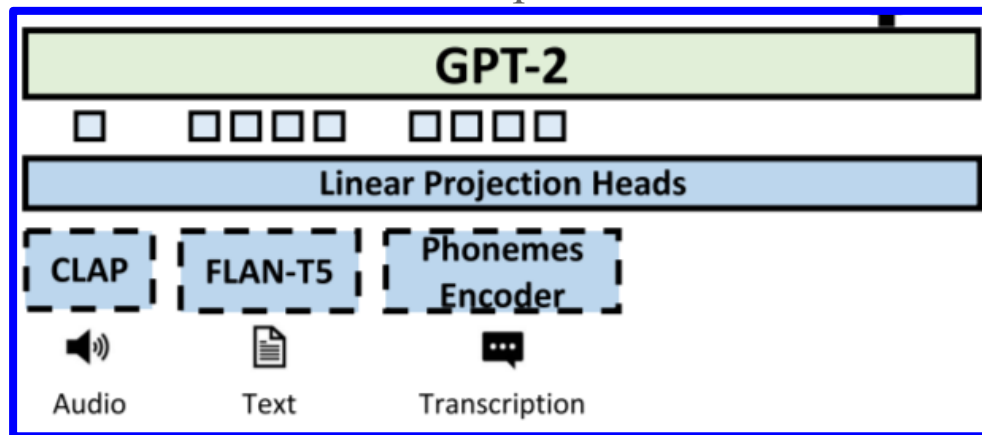
GPU Optimization

- Graphics Processing Units run LDMs
- Less efficiency results in higher energy consumption and computation time

GPT-2 Model

Generative Pretrained Transformer

- Creates **ground truth LOA** from multiple modalities
 - Ground truth LOA: most accurate representation of the semantic and acoustic information within an input dataset

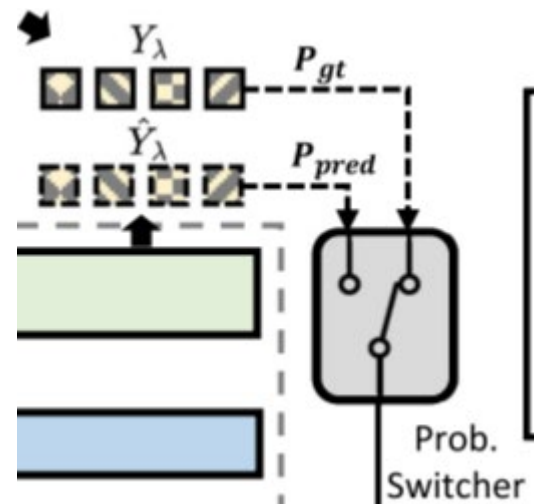


Any Modality to LOA Translator

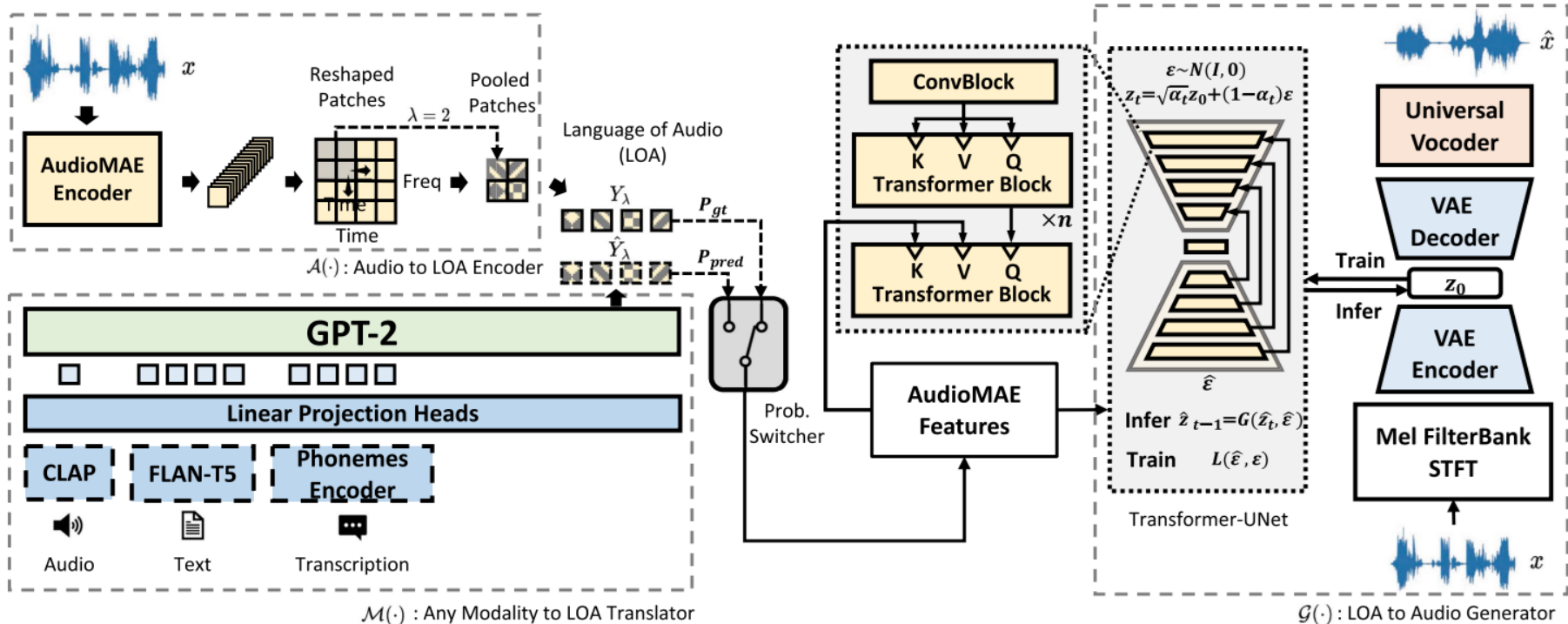
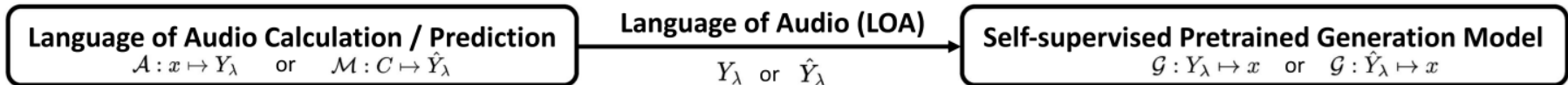
- CLAP (Contrastive Language and Audio Pretraining)
 - Text encoded conditioning information
- FLAN-T5
 - Extracts semantic information from text inputs
- Phoneme Encoder
 - Finds the smallest unit of sound that distinguishes one word from another

Probability Switcher

- Controls the probability of using LOA or ground truth LOA in the LDM
- For this experiment...
 - Probability of LOA = 0.25
 - Probability of ground truth LOA = 0.75
- Used during the training process



Results



Evaluation Metrics

- CLAP score: closeness of text prompt to audio
- FAD score: measures how close generated audio is to pre-existing music
- KLD score: measures closeness to other generated audio

Subjective Evaluation

- OVL metric: How would you rate the overall quality of this music?
- REL metric: How would you rate the relevance of music to the text description?
 - OVL and REL on a scale of *5-Excellent* to *1-Low Quality*
- MOS metric: How natural does this recording sound?
 - Given options from *completely unnatural speech* to *perfectly natural speech*

Training

- Some of the datasets used: AudioCaps, WavCaps, FMA
- LDM was trained on a random 10-second clip from the training data
- GPT-2 was trained according to prior AI models

Testing






- AudioLDM 2 was tasked with generating...
 - Text-to-audio
 - Text-to-music
 - Text-to-speech
- Each experiment was tested on three different datasets
 - AudioCaps (general audio)
 - MSD (music)
 - LJSpeech (speech)

Text-to-Audio (AudioCaps) Results Table

	FAD	KL	CLAP	OVL	REL
AudioLDM	4.53	1.99	0.141	3.61	3.55
Make-an-Audio	2.05 ↓	1.27 ↓	0.173 ↑	3.68 ↑	3.62 ↑
TANGO	1.73	1.27	0.176	3.75	3.72
AudioLDM 2-AC	1.67	1.01	0.249	3.88	3.90
AudioLDM 2-AC-Large	1.42	0.98	0.243	3.89	3.87


- Text-to-audio
 - *AudioLDM 2-AC (default)*
 - *AudioLDM 2-AC-Large*

Text-to-Music (MSD) Results Table

	FAD	KL	CLAP	OVL	REL
AudioLDM	3.20 	1.29 	0.360 	3.03 	3.25 
MusicGen	3.4	1.23	0.320	3.37	3.38
AudioLDM 2-MSD	4.47	1.32	0.294	3.41	3.30
AudioLDM 2-Full	3.13	1.20	0.301	3.34	3.54

- Text-to-music
 - *AudioLDM 2-MSD*
 - *AudioLDM 2-Full*

Text-to-Speech (LJSpeech) Results Table

	MOS
GroundTruth	4.63 
FastSpeech2	3.78
AudioLDM 2-LJS	3.65
AudioLDM 2-GIG	4.00

- Text-to-speech
 - *AudioLDM 2-LJS*
 - *AudioLDM 2-GIG*

Conclusion

Experiment Conclusions

- AudioLDM 2 is comparable to other generative AI in...
 - Text-to-audio, music, and speech
- LOA opens doors for processing general audio
- AudioLDM 2 performs best with text-to-audio and text-to-music

Sources

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. 2024. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 32 (May 2024), 2871–2883. <https://doi.org/10.1109/TASLP.2024.3399607>

Arden Dertat. 2017. Applied Deep Learning - Part 3: Autoencoders. <https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798>

Questions?