AudioLDM

Text-to-Audio Generation with Latent Diffusion Models

By Haohe Liu et al.

Presentation by Yumi Kim
on 12. March 2024

Text-to-audio Generation

Demo

Prompt : sound of birds in the forest

Demo

Prompt: A woman in pointed high heels walking on a cracked wooden floor

Prompt: A woman with pointed high heels walking down a large corridor

Diffusion model

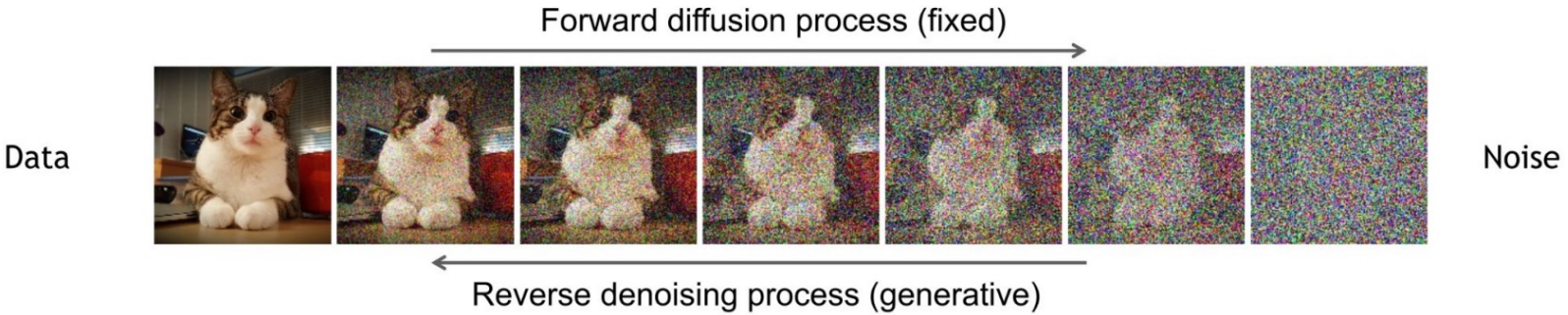


Image Source: https://uwaterloo.ca/data-analytics/sites/default/files/uploads/documents/lec17_from_stat940-allslidespdf_1-17.pdf

Latent Diffusion Model (LDM)

By Stable Diffusion(Rombach et al. 2022)

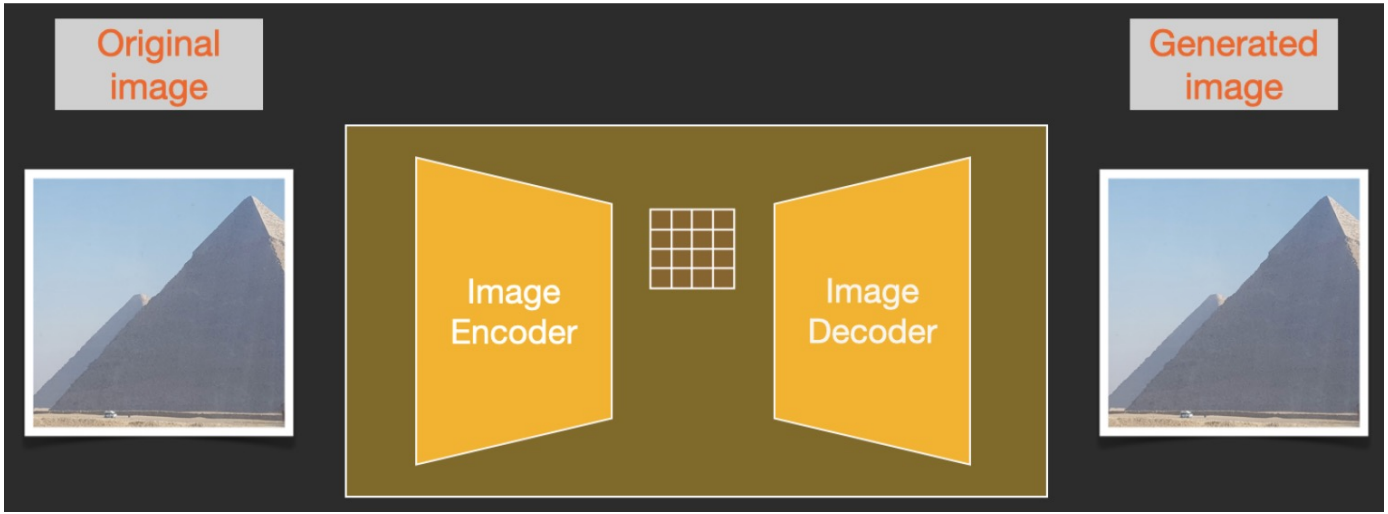


Image Source : https://jalammar.github.io/illustrated-stable-diffusion/

Latent Diffusion Model

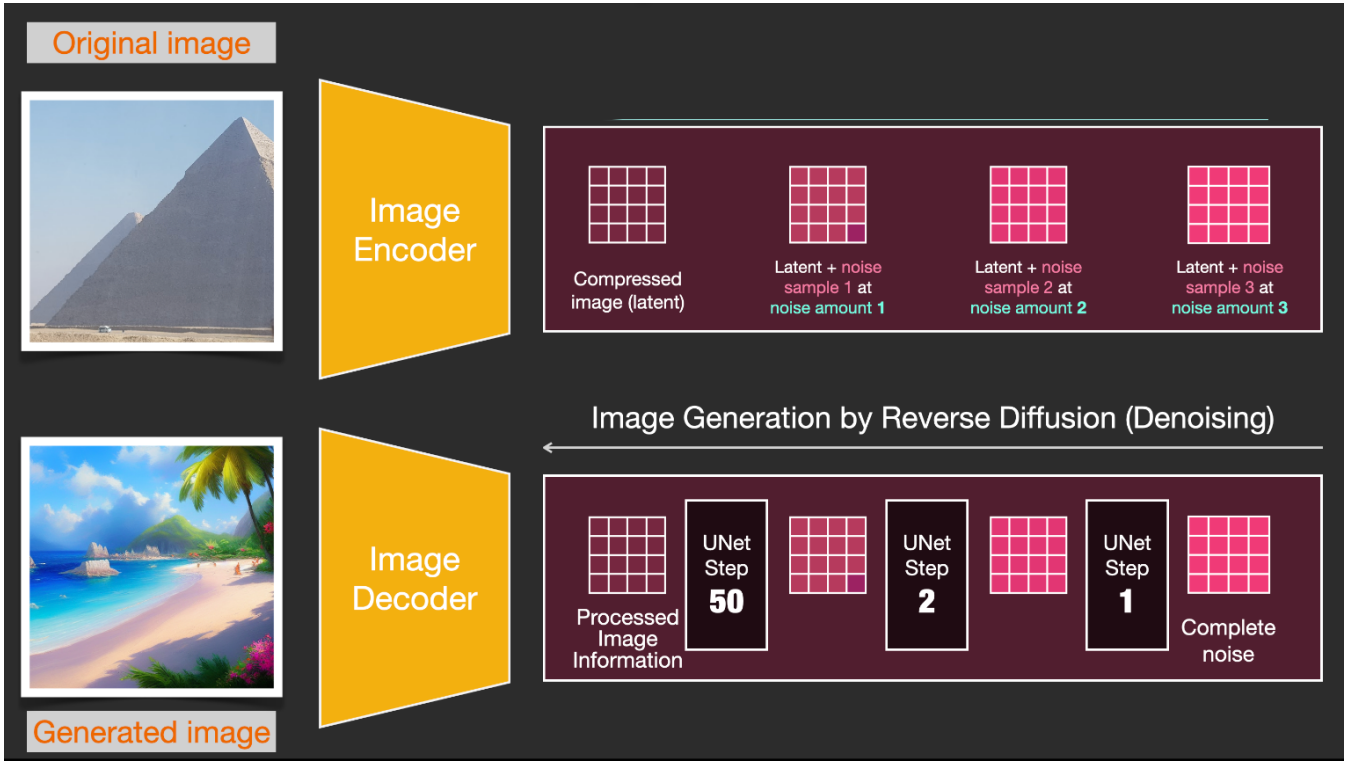
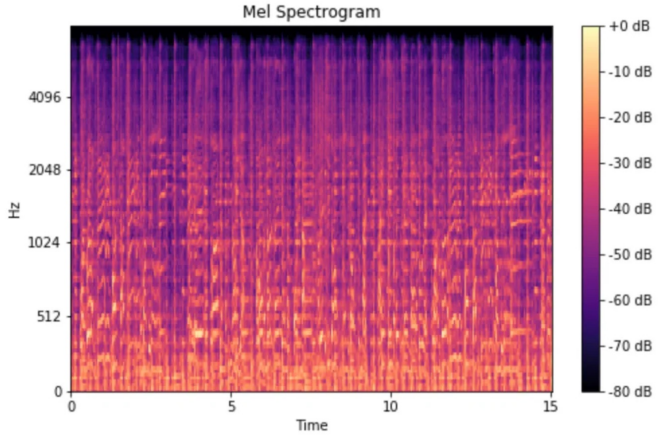


Image Source : https://jalammar.github.io/illustrated-stable-diffusion/

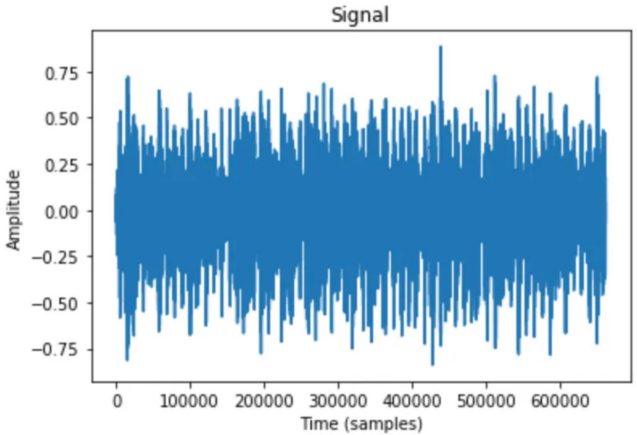
Mel Spectogram



Mel Spectogram

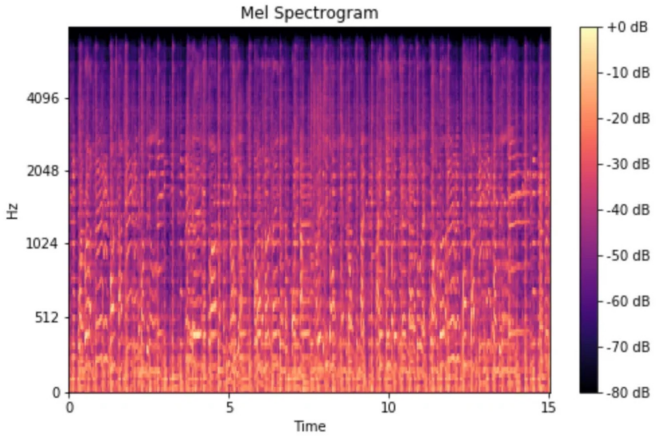
Image source :https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

Mel Spectogram



Short FFT

Mel Scale



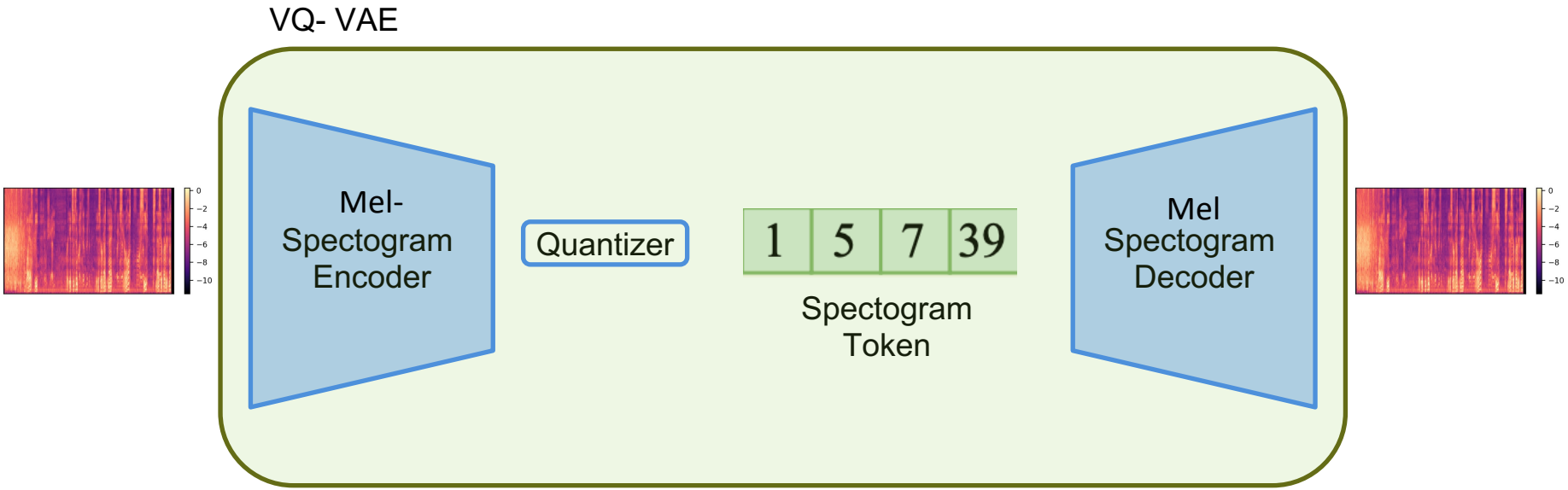
Waveform

Mel Spectogram

Image source :https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

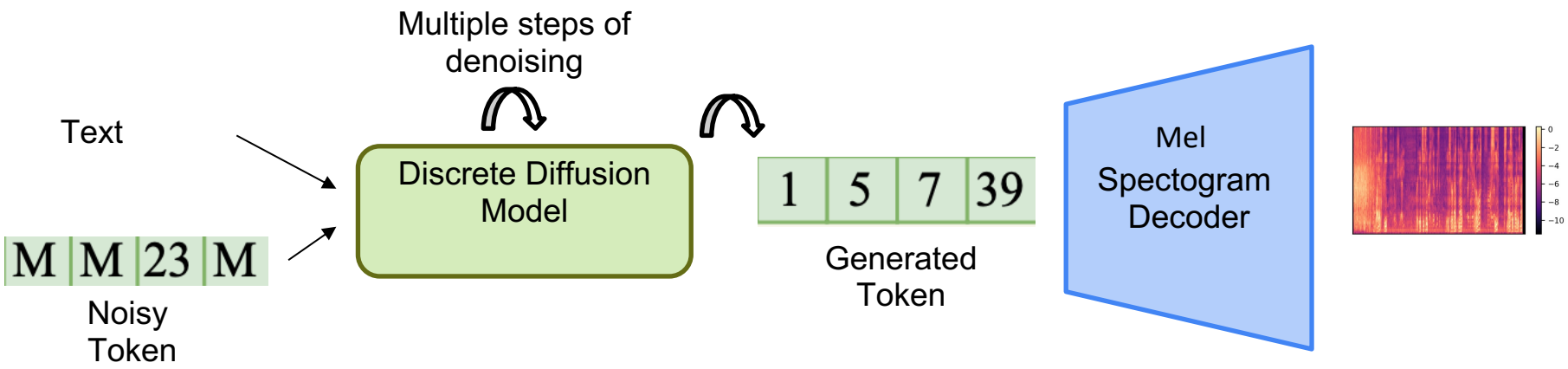
Related Work

- text-to-audio generation system : DiffSound *(Yang et.al 2021)*



Related Work

- non-autoregressive token-decoder based on the discrete diffusion model



Related Work

- Text to audio Generation: AudioGen
  - **Discrete** audio tokens
  - **Autoregressive** Transformer-based decoder
    - generates Audio tokens based on text

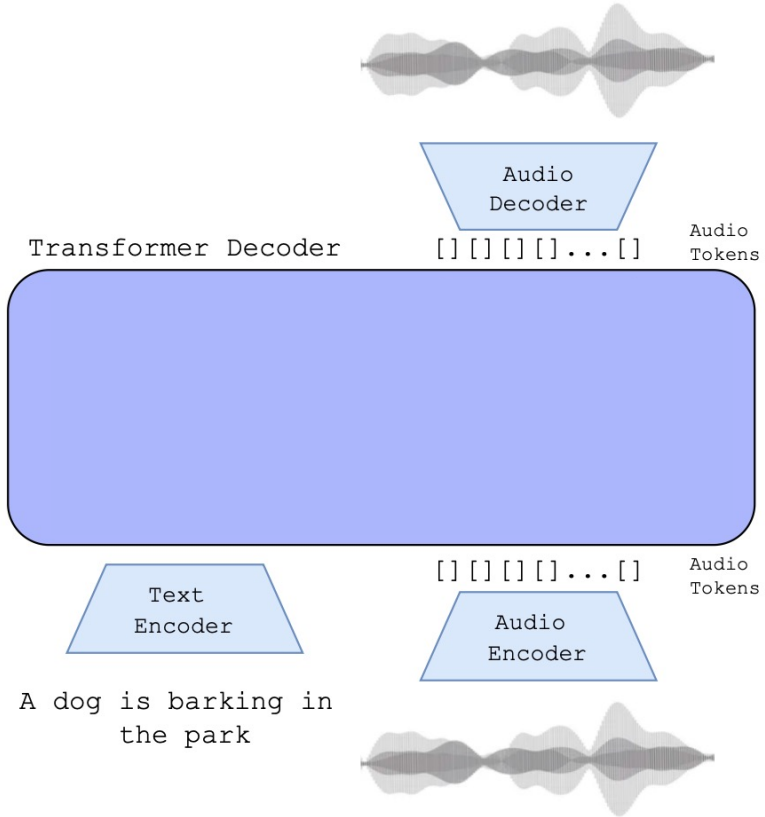


Image Source : AudioGen

Previous works

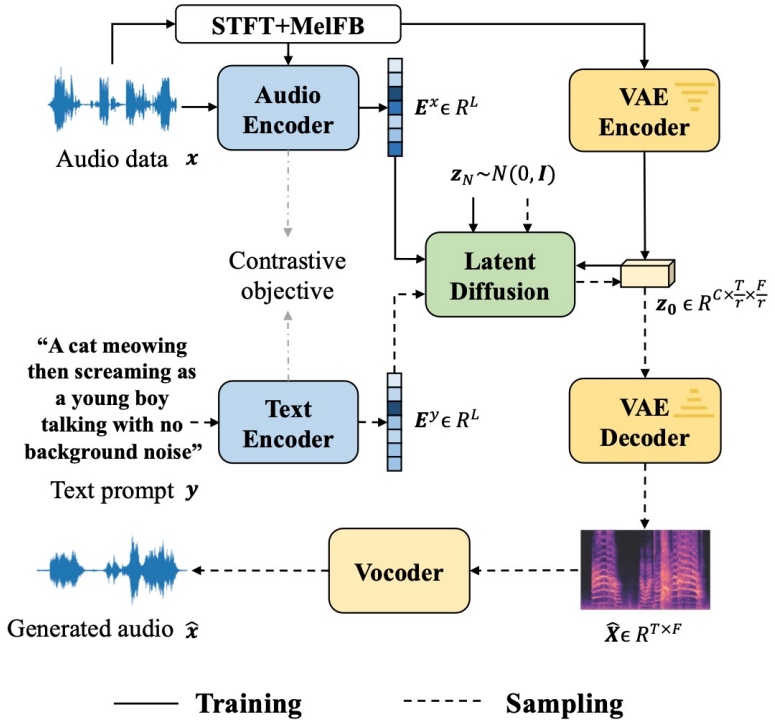
- Discrete tokens
- Paired Audio-text training data
  - Error prone text captions
  - Some captions don’t describe the audio well
    - E.g. highly abstracted “Boats : Battleships-5.25 conveyor space”

AudioLDM

- No need for audio-text paired data for LDM training
- Continuous audio representation
- text-guided Audio manipulation in zero-shot fashion
    - Style Transfer and Inpainting
- Computationally efficient
    - trained on 1 GPU

Key components

1. Contrastive Language-Audio Pretraining (CLAP) Encoders
2. Mel-spectrogram VAE
3. Latent Diffusion Models (LDM)
4. Mel-to-Waveform Vocoder



(a) Training and sampling process of AudioLDM

Image Source : AudioLDM

1. Contrastive Language-Audio Pretraining (CLAP)

By Wu et.al

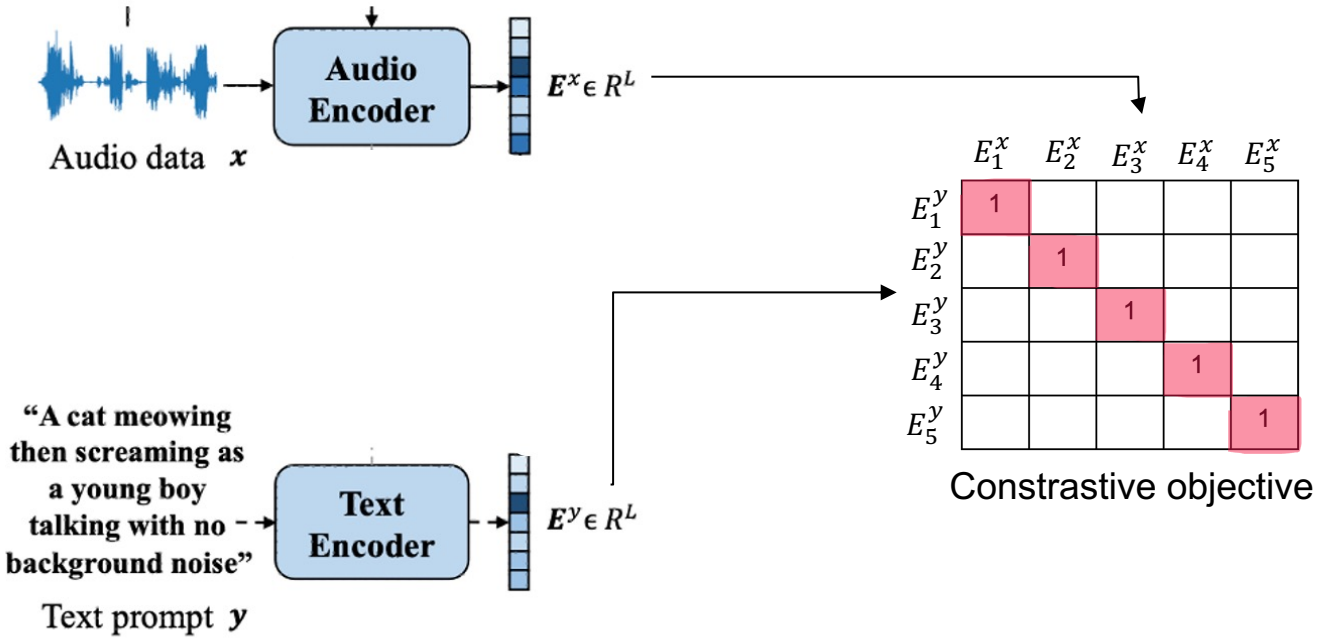
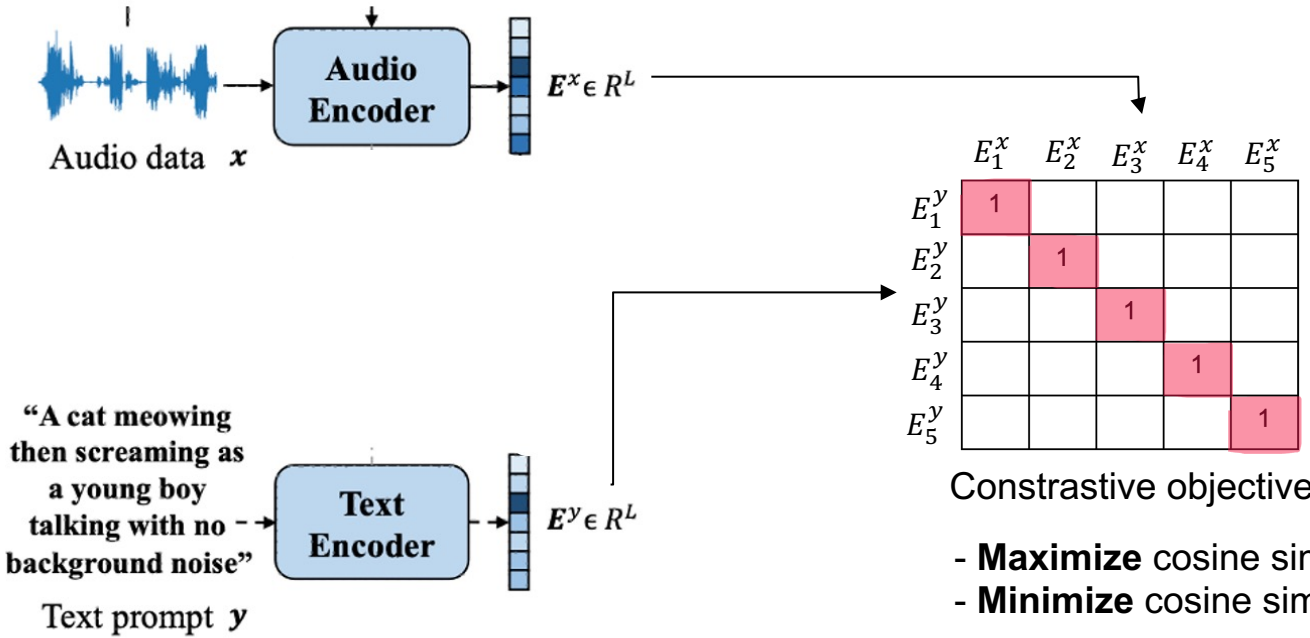


Image Source: AudioLDM

1. Contrastive Language-Audio Pretraining (CLAP)

By Wu et.al



- **Maximize** cosine similarity of **correct** pairs
- **Minimize** cosine similarity of **incorrect** pairs

Image Source: AudioLDM

Step 2. Training Mel-spectrogram VAE

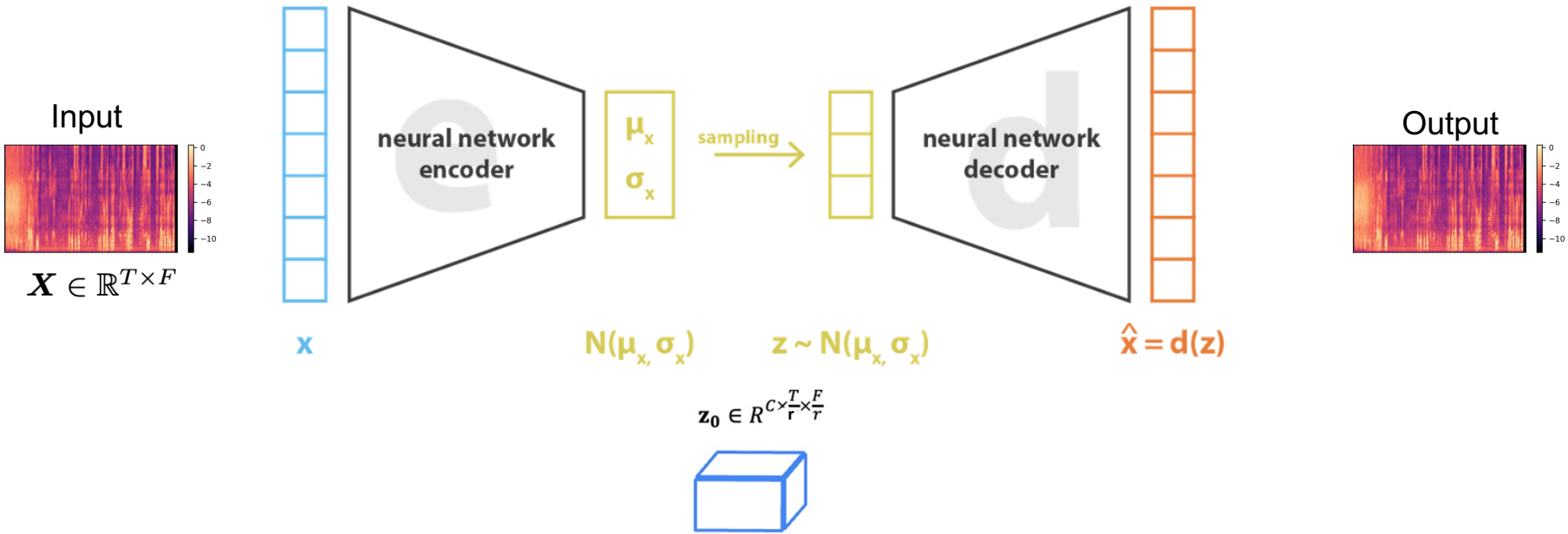


Image Source : https://lilianweng.github.io/posts/2018-08-12-vae/

Step 2. Training Mel-spectrogram VAE

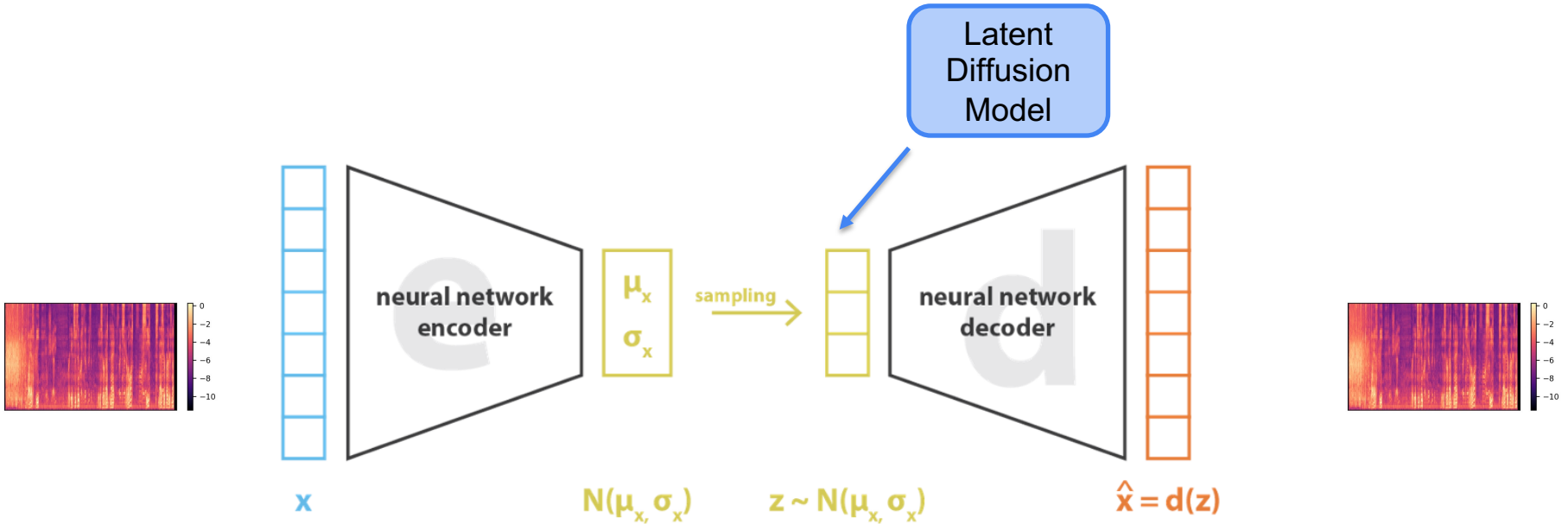
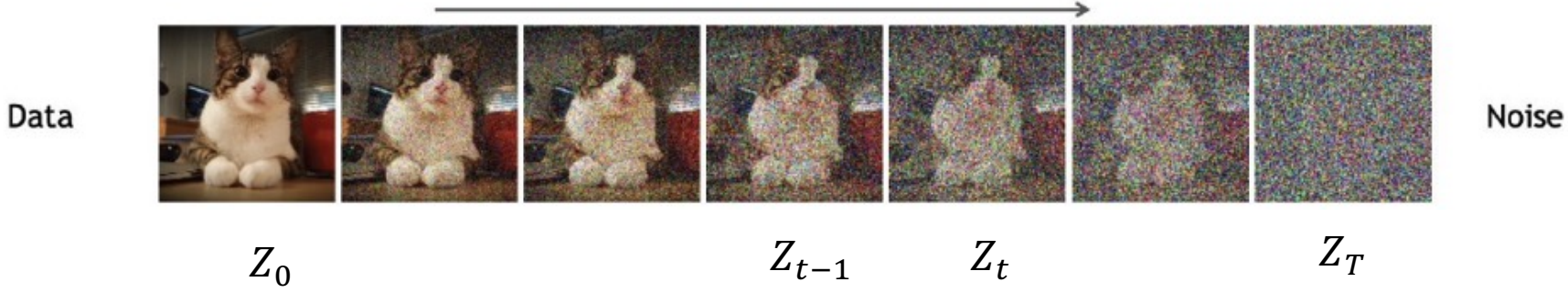


Image Source : https://lilianweng.github.io/posts/2018-08-12-vae/

Diffusion Model : Forward process

Single Forward step : $q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}\left(\mathbf{z_t}; \sqrt{1-\beta_t}\mathbf{z_{t-1}}, \beta_t \mathbf{I}\right)$



- β_t : predefined noise schedule

Image Source: https://uwaterloo.ca/data-analytics/sites/default/files/uploads/documents/lec17_from_stat940-allslidespdf_1-17.pdf

Forward process

- Diffusion models define a Markov chain of diffusion steps

$$q\left(\boldsymbol{z}_{1:T} \mid \boldsymbol{z}_0\right)=\prod_{t=1}^{T} q\left(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}\right)$$

- Using this notation

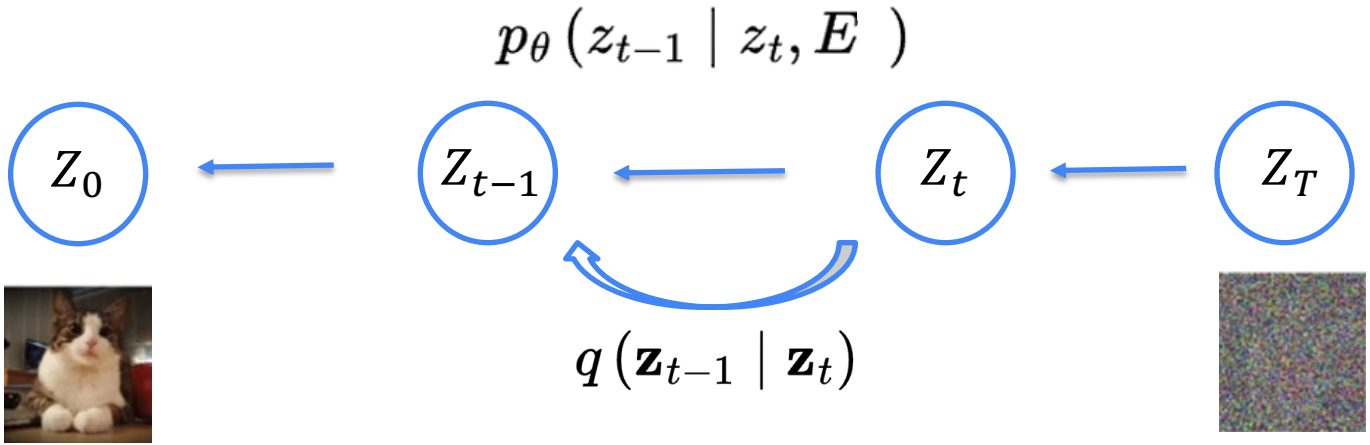
$$\alpha_t = 1 - \beta_t$$

$$\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$$

we can write: $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$

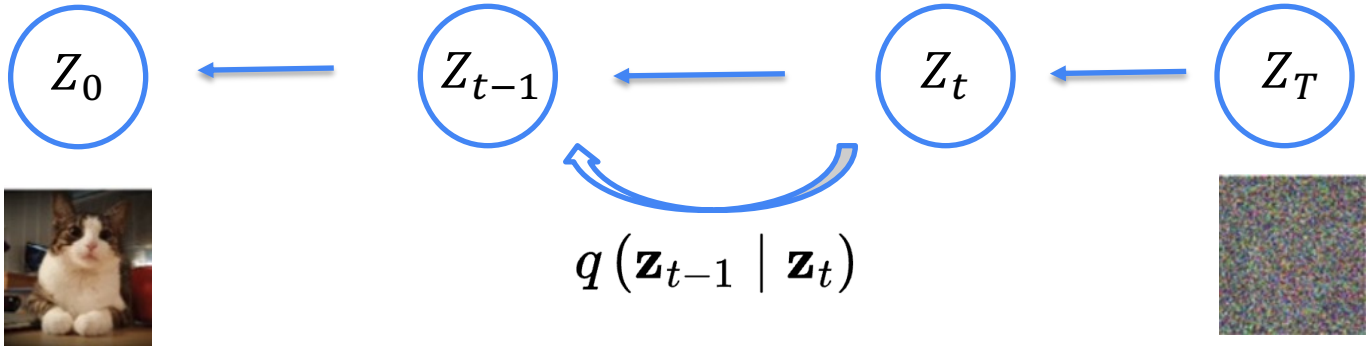
ϵ is from $\mathcal{N}(0, \mathbf{I})$

Reverse process



Reverse process

$$p_\theta\left(z_{t-1} \mid z_t, E\right)$$



$$p_\theta\left(z_{0:T} \mid E^y\right) := p\left(z_T\right) \prod_{t=1}^{T} p_\theta\left(z_{t-1} \mid z_t, E^y\right)$$

Reverse process

Following Denoising Diffusion Probabilistic Models (DDPM) by Ho et al.

$$p_\theta\left(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{E}^y\right)=\mathcal{N}\left(\boldsymbol{z}_{t-1} ; \boldsymbol{\mu}_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{E}^y\right), \boldsymbol{\sigma}_n^2 \boldsymbol{I}\right)$$

$$\boldsymbol{\mu}_\theta\left(z_t, t, \boldsymbol{E}^y\right)=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{z}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_t\right)$$

Reverse process

Following Denoising Diffusion Probabilistic Models (DDPM) by Ho et al.

$$p_\theta\left(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{E}^y\right) = \mathcal{N}\left(\boldsymbol{z}_{t-1}; \boldsymbol{\mu}_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{E}^y\right), \boldsymbol{\sigma}_n^2 \boldsymbol{I}\right)$$

$$\boldsymbol{\mu}_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{E}^y\right) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{z}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{E}^y\right)\right)$$

$$\boldsymbol{\sigma}_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

No need to learn

Reverse process

Following Denoising Diffusion Probabilistic Models (DDPM) by Ho et al.

$$p_\theta\left(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{E}^y\right)=\mathcal{N}\left(\boldsymbol{z}_{t-1} ; \boldsymbol{\mu}_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{E}^y\right), \boldsymbol{\sigma}_n^2 \boldsymbol{I}\right)$$

$$\boldsymbol{\mu}_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{E}^y\right)=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{z}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{E}^y\right)\right)$$

$$\boldsymbol{\sigma}_t^2=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t$$

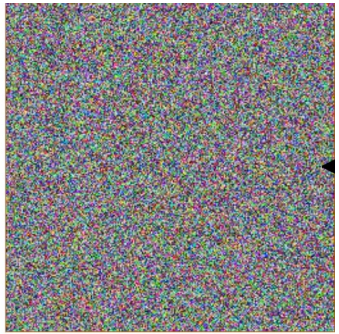
No need to learn

Loss function: $L_t(\theta)=\mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{\epsilon}, t}\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{E}^x\right)\right\|_2^2$

Noise Predictor U-Net

$\epsilon_\theta\left(z_t, t, \boldsymbol{E}\right)$

Predicted Noise



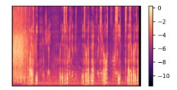
CLAP embedding

⊕

Step embedding for step t

Image Source: https://uwaterloo.ca/data-analytics/sites/default/files/uploads/documents/lec17_from_stat940-allslidespdf_1-17.pdf

Step 3. Training Diffusion model



Audio Sample

x

VAE Encoder

$\mathbf{z_0} \in R^{C \times \frac{T}{r} \times \frac{F}{r}}$

$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$

Inject noise ϵ

$z_t \in R^{C \times \frac{T}{r} \times \frac{F}{r}}$

$$L_t(\theta) = \mathbb{E}_{z_0, \epsilon, t} \| \epsilon - \epsilon_\theta (z_t, t, \boldsymbol{E}^x) \|_2^2$$

CLAP Audio Encoder

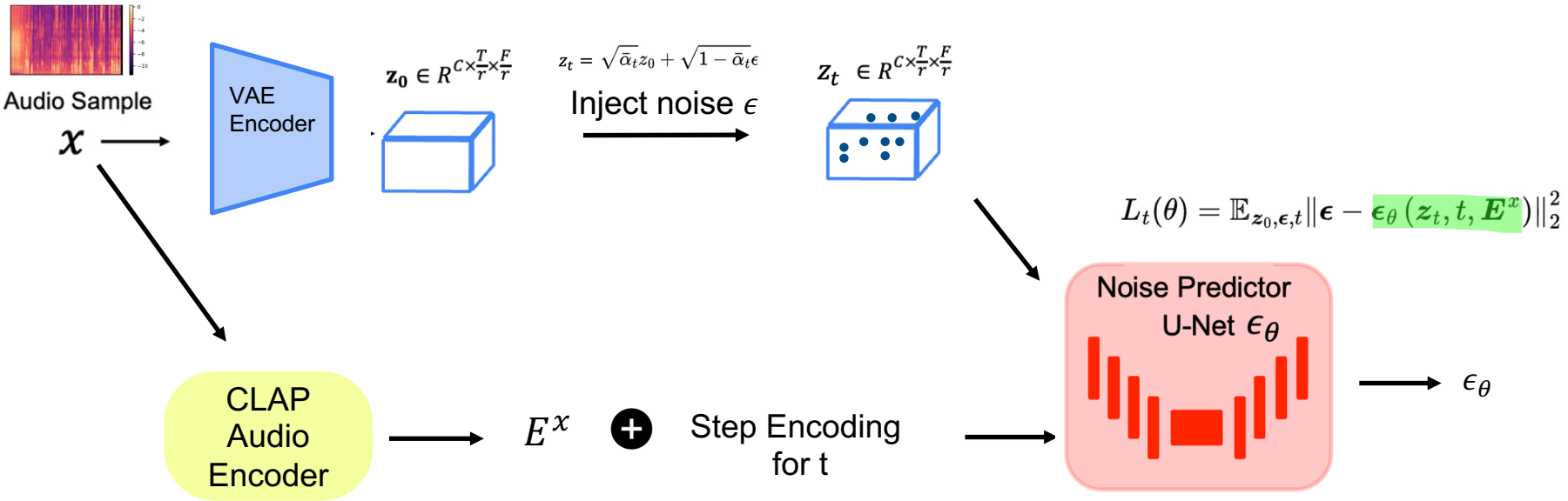
E^x

Step Encoding for t

Noise Predictor U-Net ϵ_θ

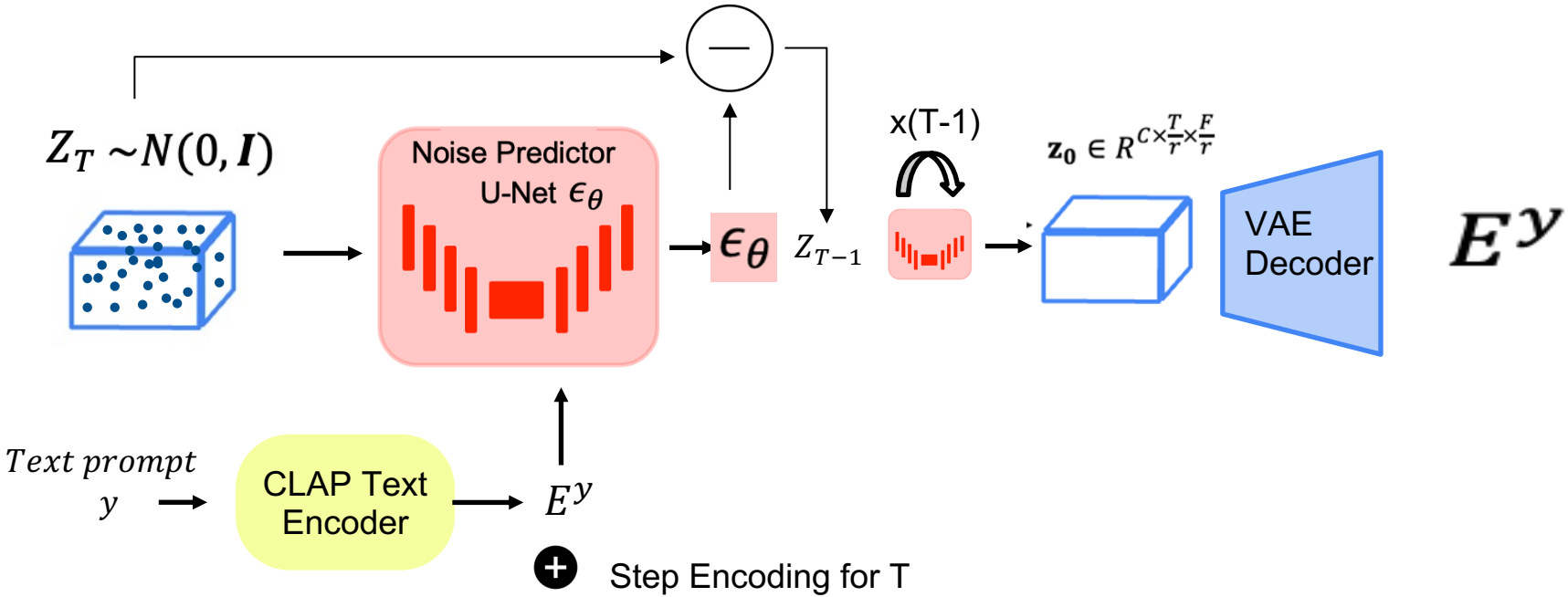
ϵ_θ

Step 3. Training Diffusion model



No Need for paired audio-text data

Generation


$Z_T \sim N(0, I)$
Noise Predictor
U-Net ϵ_θ
ϵ_θ
Z_{T-1}
x(T-1)
$\mathbf{z_0} \in R^{C \times \frac{T}{r} \times \frac{F}{r}}$
VAE
Decoder
E^y
$Text\ prompt$
y
CLAP Text
Encoder
E^y
Step Encoding for T

Text-guided audio Manipulation

- Style Transfer

From *trumpet* to *children singing*

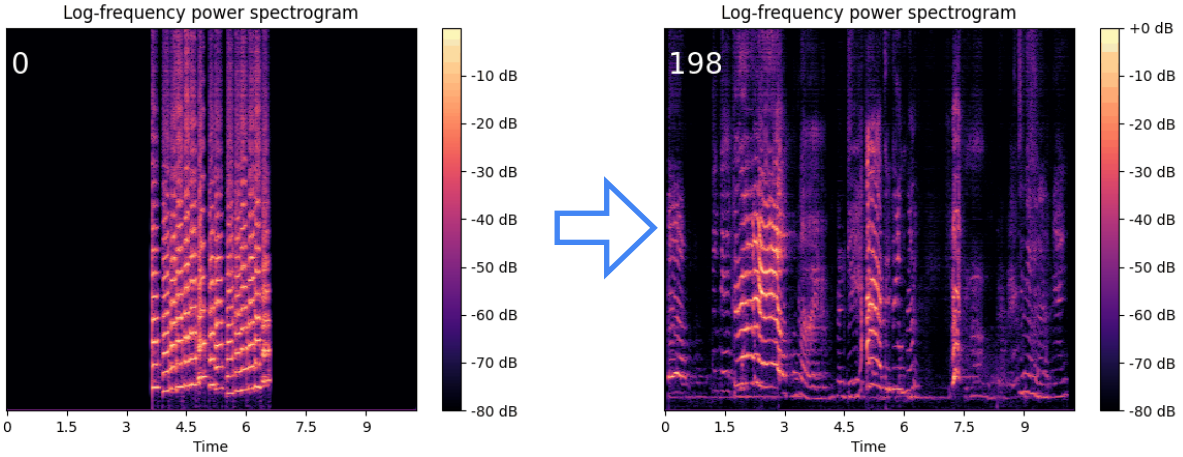
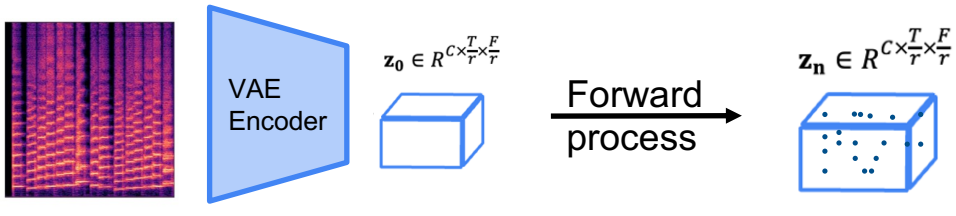


Image and Audio Source: AudioLDM

Style Transfer

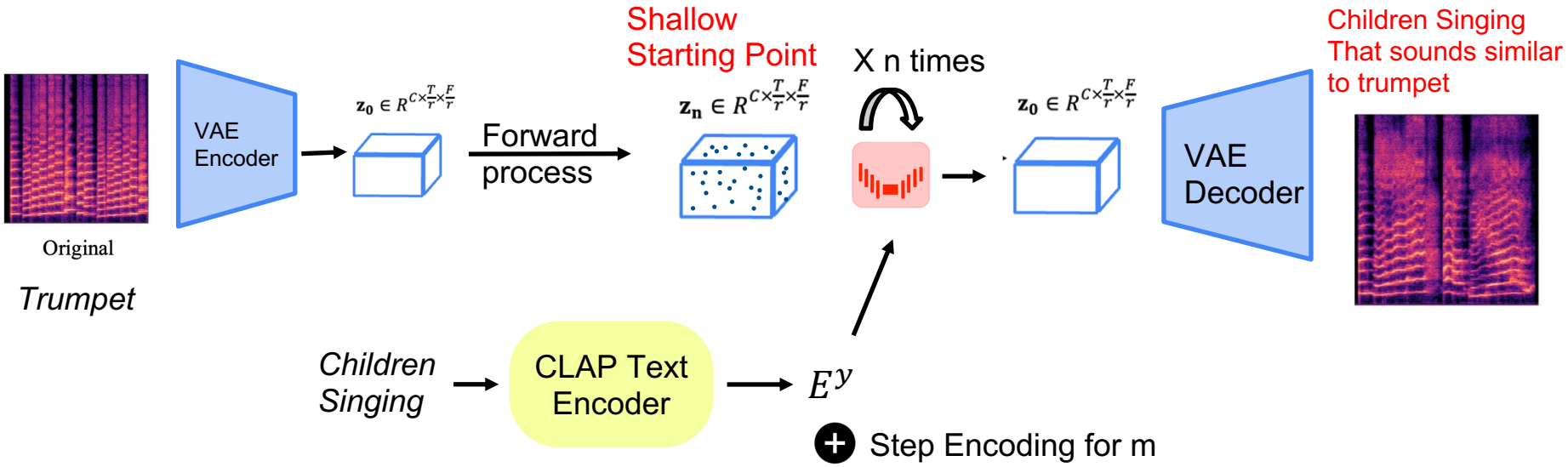


Original

Trumpet

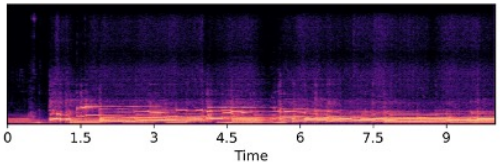
Only do n<T steps and
not the full forward process!

Style Transfer

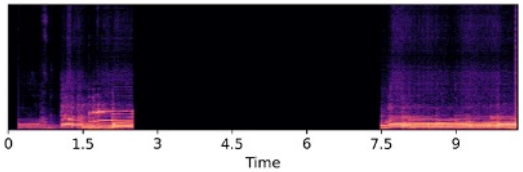


Inpainting

Ground-Truth



Unprocessed



Inpainting result

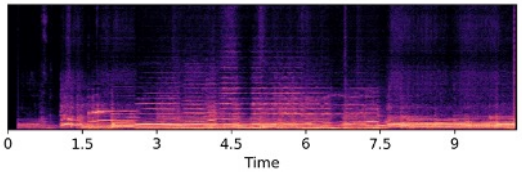
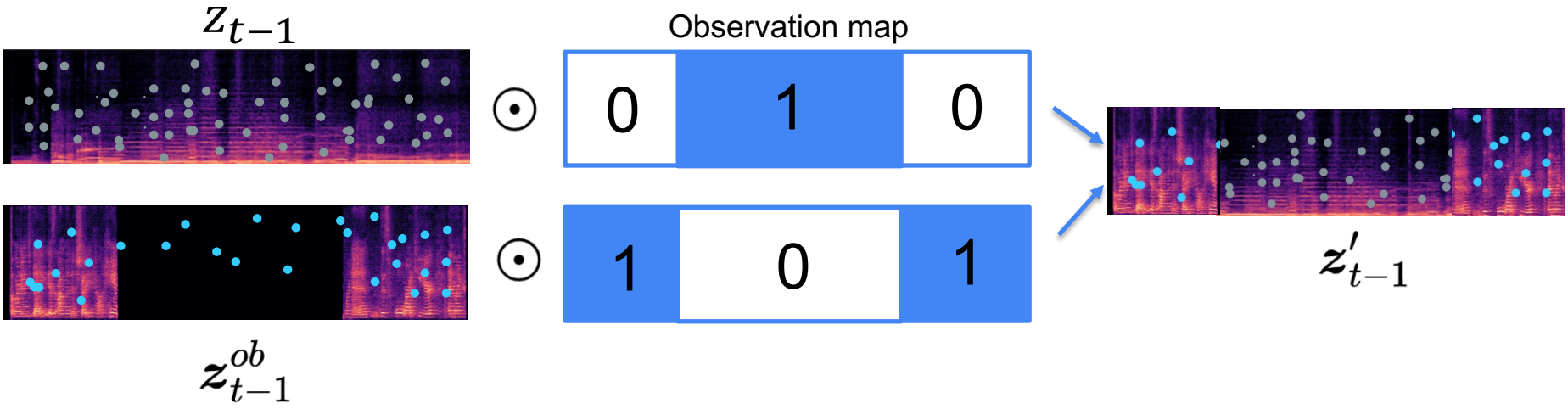


Image and Audio Source: AudioLDM

At each denoising step

- replace generated z_{t-1} with $\boldsymbol{z}'_{t-1} = (1 - \boldsymbol{m}) \odot \boldsymbol{z}_{t-1} + \boldsymbol{m} \odot \boldsymbol{z}^{ob}_{t-1}$



Results

Training Data

- AudioSet
- AudioCaps
- FreeSound
- BBC Sound Effect Library

=> ~9000 hours of audio samples with 16kHz sampling rate

Evaluation Metric

- Objective evaluation
  - FAD: Frechet Audio Distance
  - KL: Kullback-Leibler Divergence
  - IS: Inception Score

Evaluation Metric

- Objective evaluation
  - FAD: Frechet Audio Distance
  - KL: Kullback-Leibler Divergence
  - IS: Inception Score
- Subjective evaluation
  - overall quality (OVL)
  - relevance to the input text (REL)

Evaluation on AudioCaps Dataset

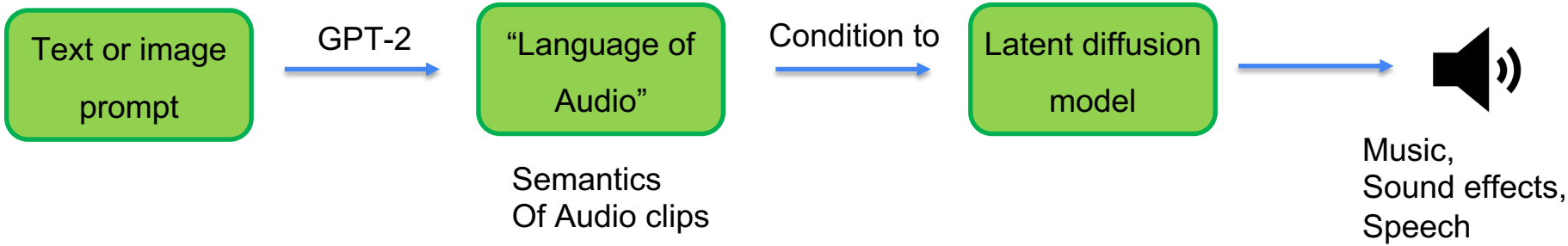
Model	Params	Duration (h)	IS ↑	KL ↓	FAD ↓	OVL ↑	REL ↑
Ground truth	-	-	-	-	-	$83.61_{\pm 1.1}$	$80.11_{\pm 1.2}$
DiffSound† (Yang et al., 2022)	400M	5420	4.01	2.52	7.75	$45.00_{\pm 2.6}$	$43.83_{\pm 2.3}$
AudioGen† (Kreuk et al., 2022)	285M	8067	-	2.09	3.13	-	-
AudioLDM-L-Full	739M	8886	**8.13**	**1.59**	**1.96**	$\mathbf{65.91}_{\pm 1.0}$	$\mathbf{65.97}_{\pm 1.6}$

Limitation

- Low sampling rate (16kHz) of data
- Limited generation quality
    - Incomprehensible speech generation

Future work : AudioLDM 2

- Higher sampling rate (48kHz)
- Improved Text-to-speech generation
- SOTA performance on Music and Sound effect generation
- Image Prompt



Thank you!