Generating Music with **Diffusion Models**

Timo Hromádka | Dr Sam Nallaperuma-Herzberg | Dr Carl Henrik Ek

Table of Contents

0. Quick Motivation

Some Motivation

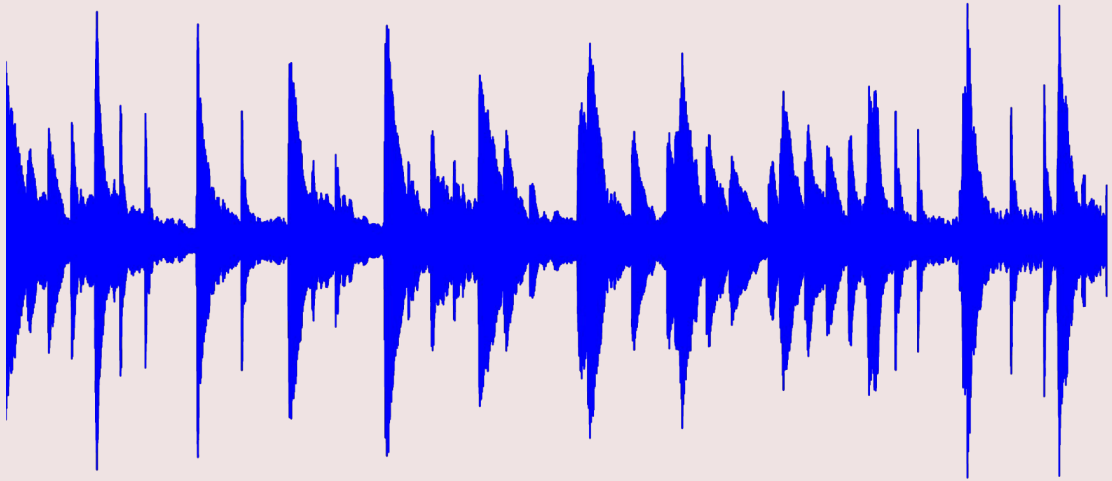
Music Production Tools

Insomnia Therapy

Reconstruction/Restoration of old audio

1. What is Music?

Representing Music: Waveform



Representing Music: Waveform

(Zoomed in)

Representing Music: Waveform

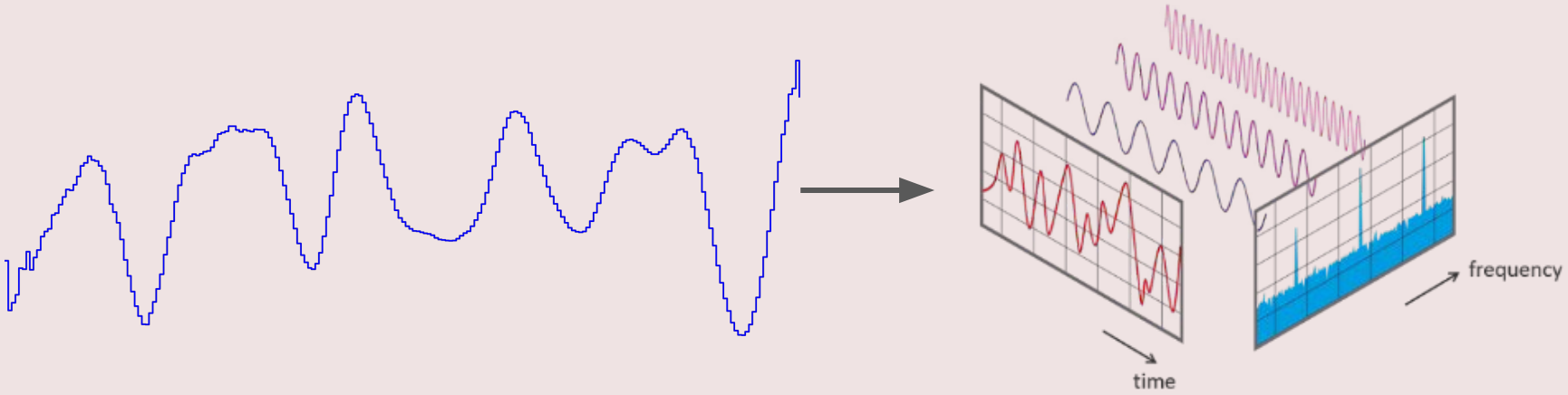
Stereo (2-channel)

```
tensor([[[ 0.0209,  0.0284,  0.0259,  ..., -0.0784, -0.1018, 0.1061],
         [-0.0153,  0.0352, -0.0456,  ...,  0.0897,  0.0674, -0.0542]]])
```

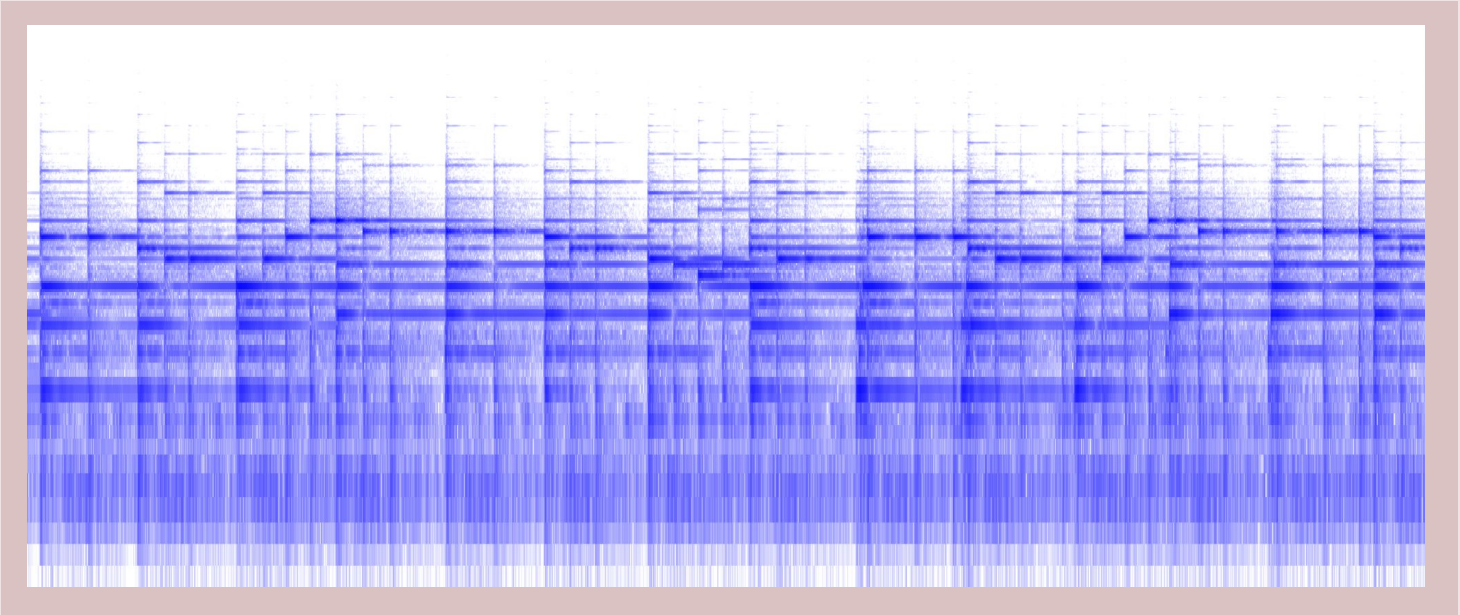
Mono (1-channel)

```
tensor([[ 0.0028,  0.0318, -0.0099,  ...,  0.0057, -0.0172,  0.0259]])
```

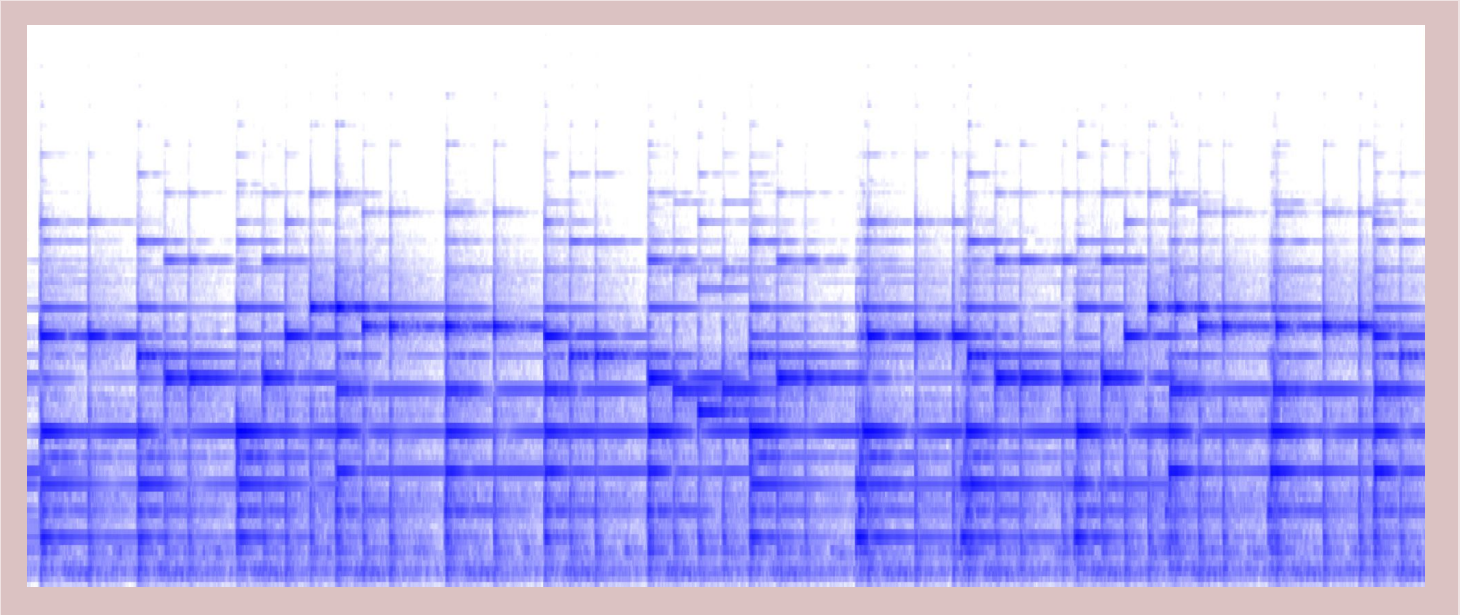
Representing Music: Spectrogram



Representing Music: Spectrogram



Representing Music: Mel-Spectrogram



2. What is Generative Modelling?

Generative Modelling

They are **probabilistic models of data**

- Tool for modelling probability distributions
- Describe the probabilistic process of generating an observation
- Allow for **new datapoint generation**

Approaches

Autoregressive models
- Transformer and RNN sequence models, NADE, etc.

Latent variable models
- Flow-based Models, Variational Autoencoders (VAEs), Diffusion Models

Implicit models
- Generative Adversarial Networks (GANs)

Approaches

~~**Autoregressive models**~~

- ~~Transformer and RNN sequence models, NADE, etc.~~

Latent variable models

- Flow-based Models, Variational Autoencoders (VAEs), **Diffusion Models**

~~**Implicit models**~~

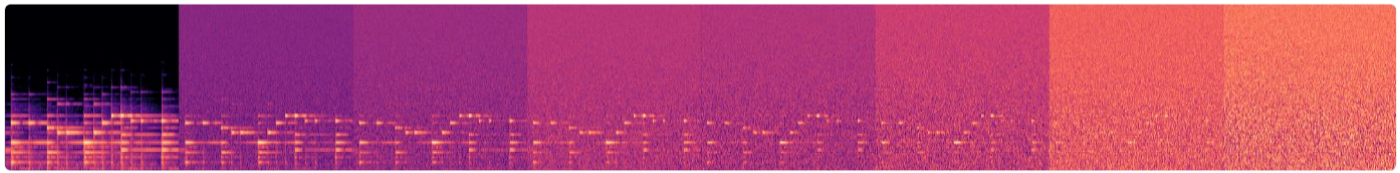
- ~~Generative Adversarial Networks (GANs)~~

Diffusion Models

forward process

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

t = 0 → t = T



$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

reverse process

θ are the parameters of our NN!

Diffusion Models

Loss Function - maximize log likelihood of generated sample belonging to original data distribution

$$p_\theta(x_0) := \int p_\theta(x_{0:T}) dx_{1:T}$$

$$L = -log(p_\theta(x_0))$$

Diffusion Models

Intractable due to integration over very dimensional data space over T timesteps

$$p_\theta(x_0) := \int p_\theta(x_{0:T}) dx_{1:T}$$

$$L = -log(p_\theta(x_0))$$

Diffusion Models

Approximation: Evidence Lower Bound (ELBO)

$$\mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] \leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] = \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t\geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right] =: L$$

$$\mathbb{E}_q\Bigg[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p(\mathbf{x}_T))}_{L_T} + \sum_{t>1} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\Bigg]$$

$$\begin{aligned}
L_{\text{vlb}} &:= L_0 + L_1 + \ldots + L_{T-1} + L_T \\
L_0 &:= -\log p_\theta(x_0|x_1) \\
L_{t-1} &:= D_{KL}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)) \\
L_T &:= D_{KL}(q(x_T|x_0) \parallel p(x_T))
\end{aligned}$$

$$L_{vlb} := L_{t-1} := D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))$$

Diffusion Models

$$L_{vlb} := L_{t-1} := D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))$$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}),$$

where $\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t$ and $\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

$$\downarrow$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma^2 I)$$

Diffusion Models

$$L_{vlb} := L_{t-1} := D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))$$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}),$$

$$\text{where} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

$$\downarrow$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma^2 I)$$

$$L_{t-1} = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2\right] + C$$

Diffusion Models

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2 \right] + C$$

$$\bar{\mu}(x_t, x_0) = \frac{1}{\sqrt{\bar{x}_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \right)$$

$$\mu_\theta(x_t, x_0) = \frac{1}{\sqrt{\bar{x}_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right)$$

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon}} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2 \right]$$

Diffusion Models

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2 \right] + C$$

$$\bar{\mu}(x_t, x_0) = \frac{1}{\sqrt{\bar{x}_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \right)$$

$$\mu_\theta(x_t, x_0) = \frac{1}{\sqrt{\bar{x}_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right)$$

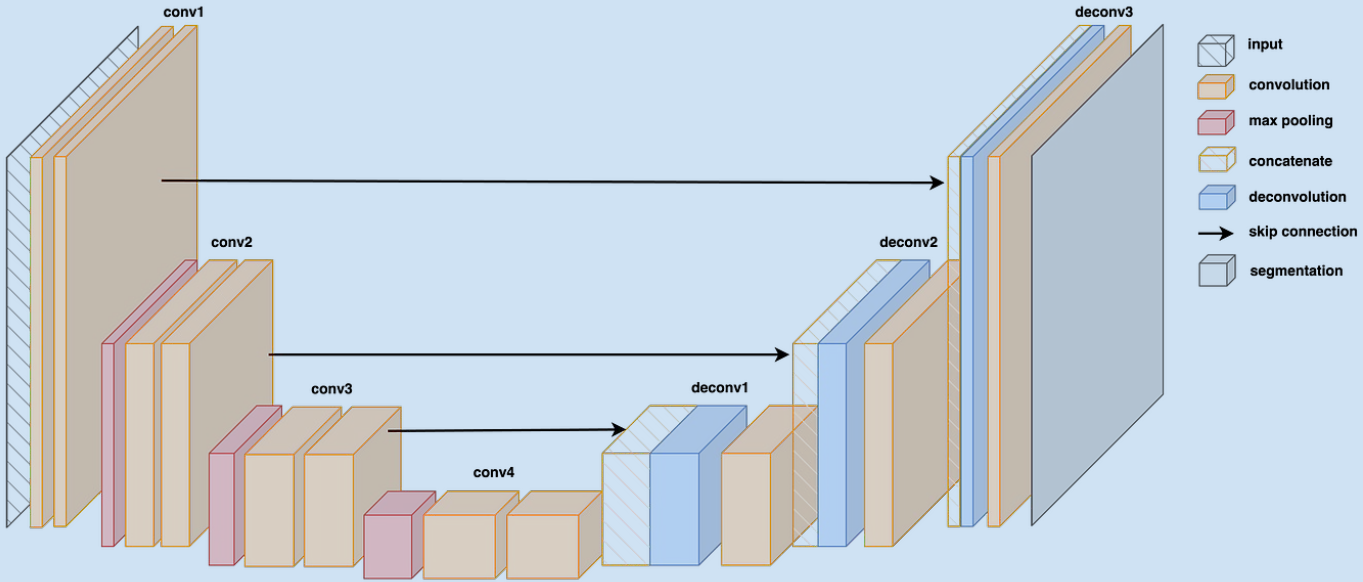
$$L_{\text{simple}}(\theta) := \mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon}} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2 \right]$$

Just a **MSE** between **noise added** in forward process and the **noise predicted** by model

$$L_{\text{simple}} = E_{t, x_0, \epsilon} \left[||\epsilon - \epsilon_\theta(x_t, t)||^2 \right]$$

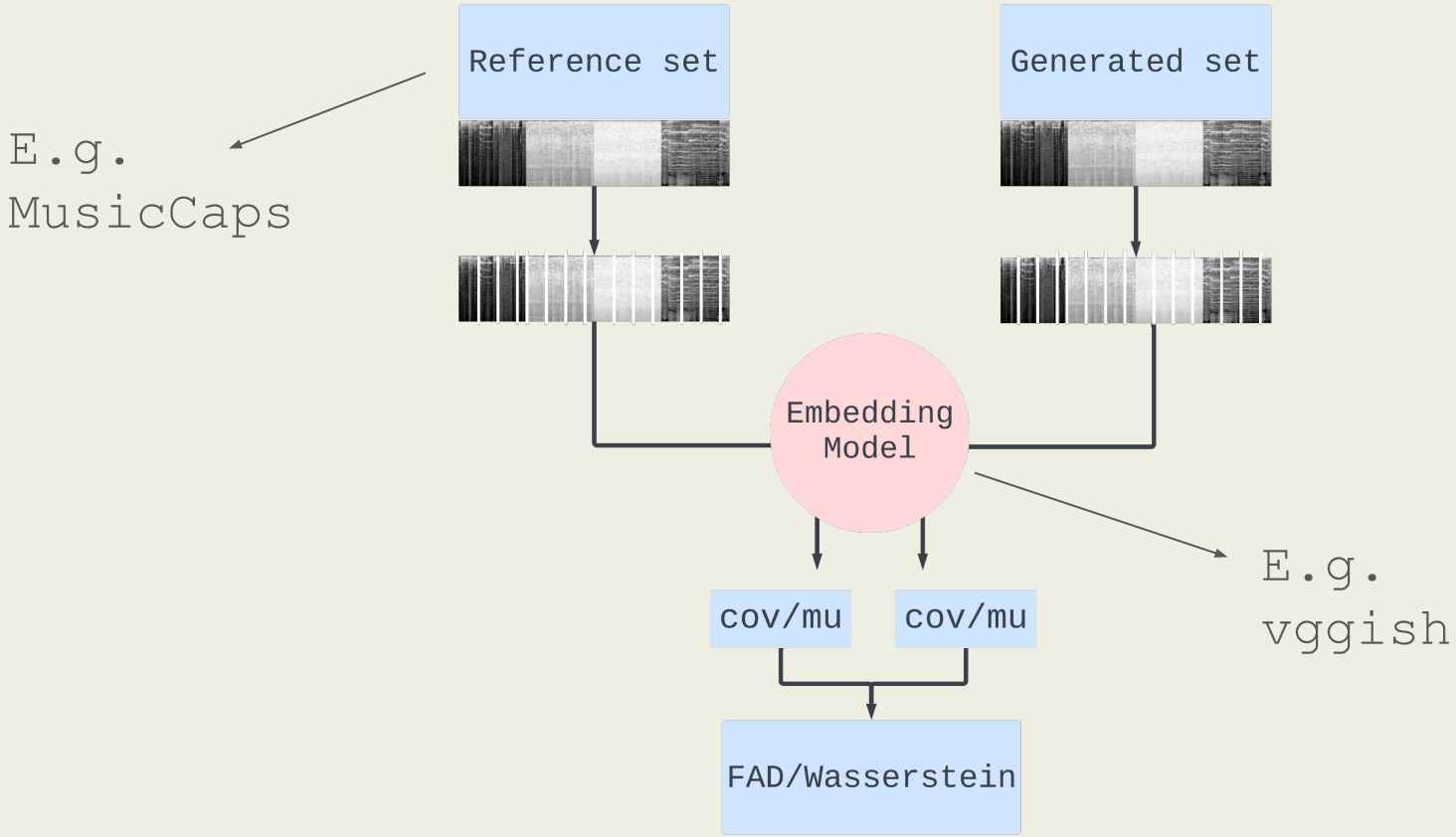
Diffusion Models

Diffusion models use a U-net for the NN itself.



3. Evaluation

Frechet Audio Distance



4. Current Approaches

Comparison

	Training Setup				Approach		Evaluation Setup			
Paper	**Training Dataset**	**Training Time**	**Audio Rate**	**Model Size**	**Method**	**Tasks**	**Evaluation Dataset**	**FAD Model**	**Num FAD Samples**	**Inference Speed**
Jen-1 [43]	Private (10s samples)	200k steps (x8 A100 GPUs)	48kHz	746M	1d Diffusion Model + Masked Autoencoder	text conditioning inpainting continuation	Musiccaps [2]	N/A	N/A	Unknown
Archisound [64]	N/A	1M steps (1 week)	48kHz	178M - 857M	Diffusion Autoencoder + Latent Diffusion	Text Conditioning	N/A	N/A	N/A	∼real time
MusicGen [6]	Private Shutterstock[1] Pond5[2]	500K steps	32kHz	300M - 3.5B	Autoregressive Transformer	Text Conditioning	Musiccaps [2]	VGGish	N/A	N/A
AudioLDM [44]	AudioCaps [33] AudioSet +2 more [3]	1.5M steps	16kHz	181M - 739M	Latent Diffusion	Text Conditioning	AudioCaps AudioSet	VGGish	N/A	minutes
Noise2Music [27]	Private	4.97M steps (all models)	16kHz	2M total (all models)	Cascading Diffusion Model	Text Conditioning	MusicCaps [2] AudioSet MagnaTagnTune	VGGish, Trill, MuLan	N/A	minutes
Audiogen [39]	AudioSet [18] AudioCaps [33] +7 more	200K steps (x128 A100 GPUs)	16kHz	285M - 1B	Autoregressive Diffusion Model	Text Conditioning	AudioSet	N/A	N/A	hours
Moûsai [65]	Private	1M steps	48kHz	∼1B	Latent Diffusion	Text Conditioning	Private	PANN [37]	N/A	∼real-time
Make-an-Audio [28]	AudioSet [18] AudioCaps [33] +12 more [76, 55, 12, 47, 15, 3] [4][5][6][7][8][9]	2M steps (x18 V100 GPUs)	16kHz	332M	Latent Diffusion	Text Conditioning	AudioCaption	FID, KL	N/A	N/A (but slower than real-time)

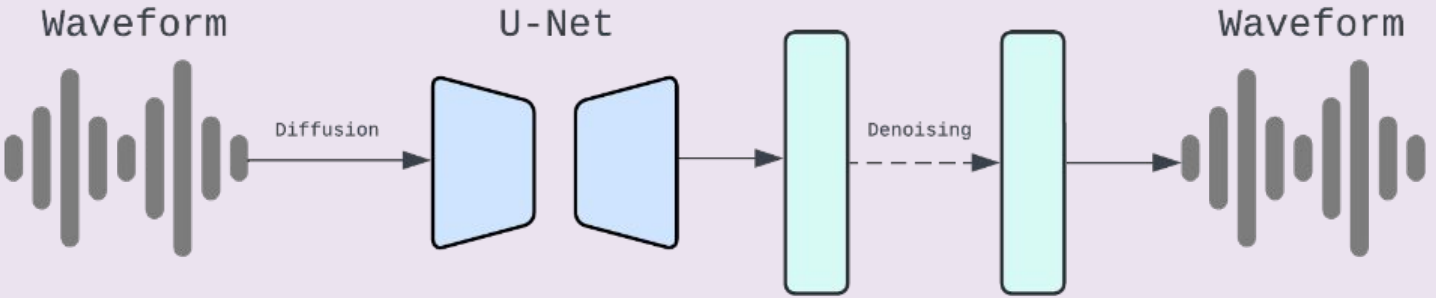
5. Aim of my Project

The focus of my Project

- **Real-world application** (sleep treatment, music production tools)
- **Real-time music generation** (on a budget)

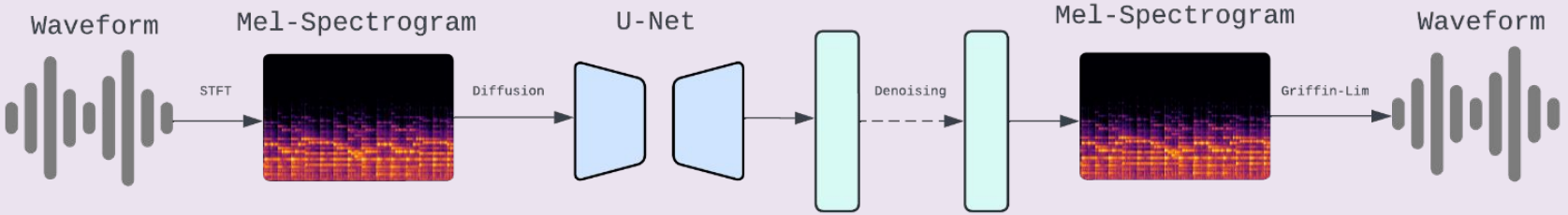
Setup - Waveform model

1.



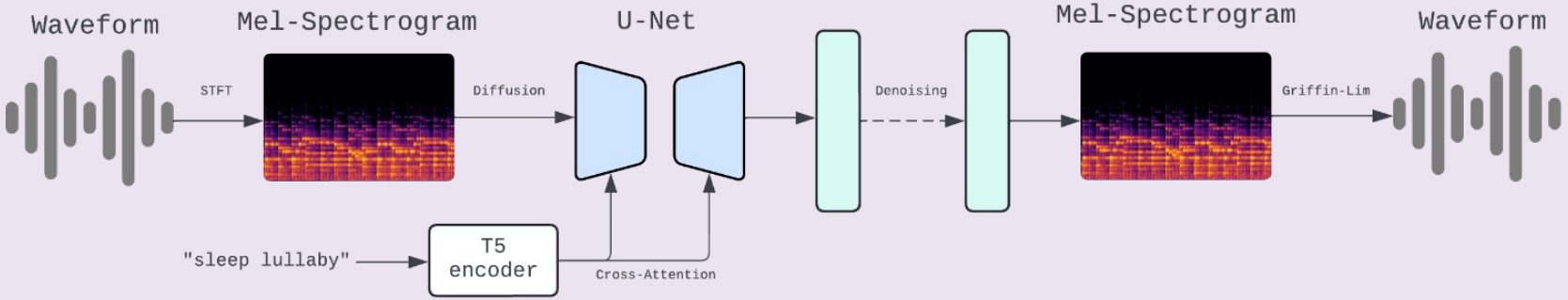
Setup - Mel-spec model

2.



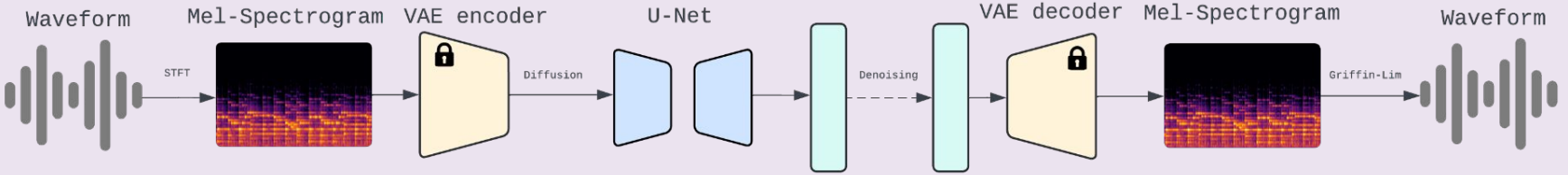
Setup - Conditional mel-spec model

3.



Setup - Latent Space model

4.

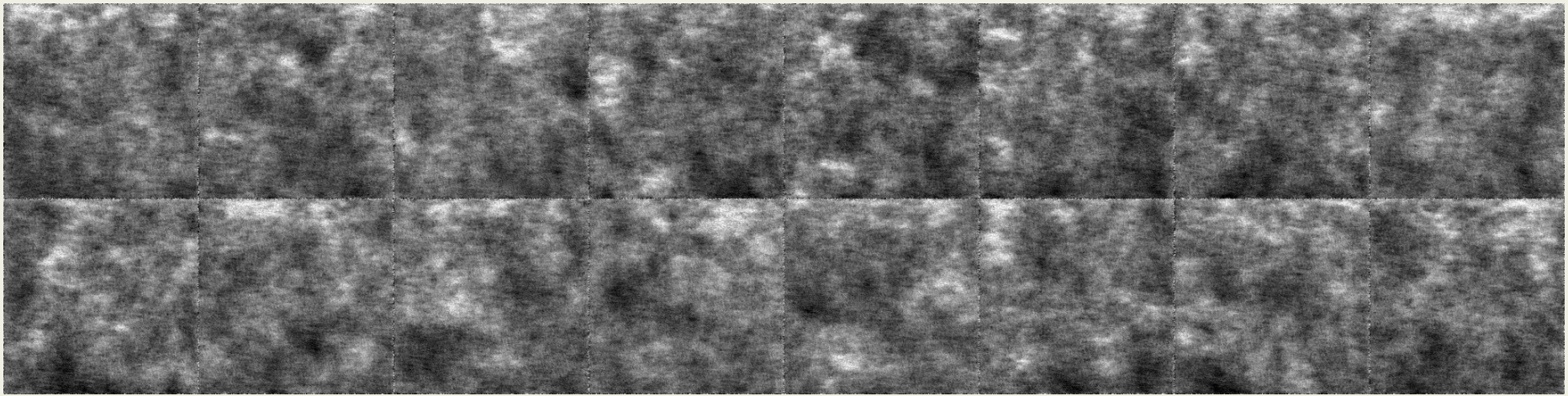


6. Experiment

Experiment - FAD Evaluation

Experiment

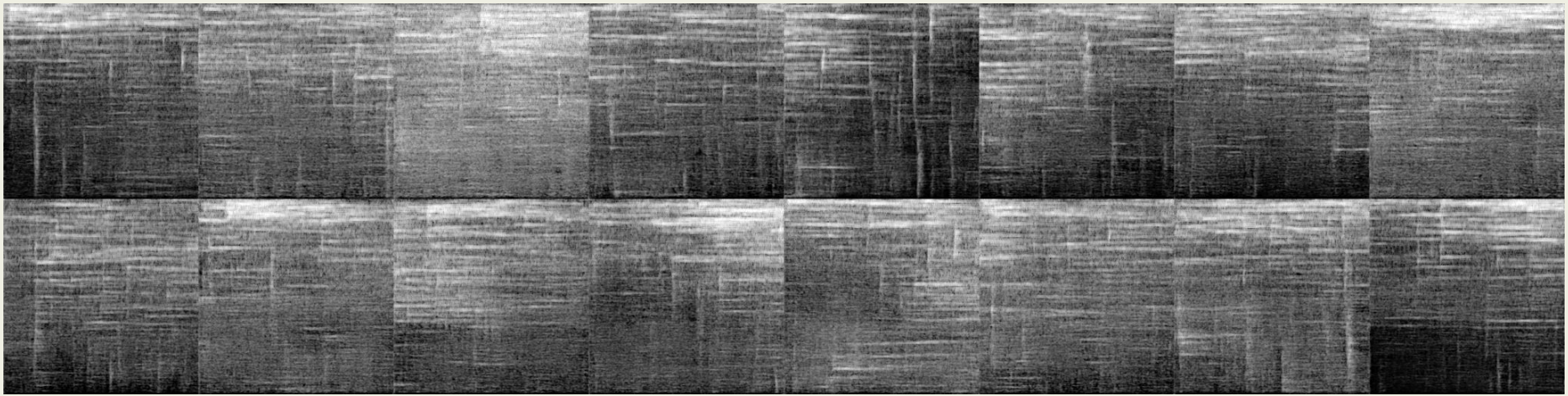
Training Set: Spotify Sleep Dataset



100 steps

Experiment

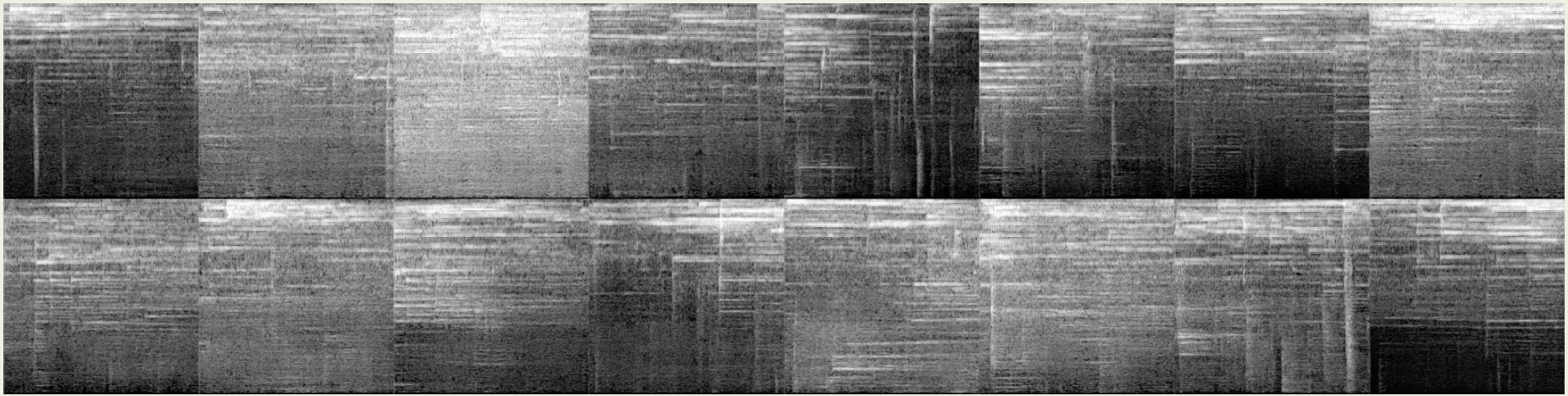
Training Set: Spotify Sleep Dataset



500 steps

Experiment

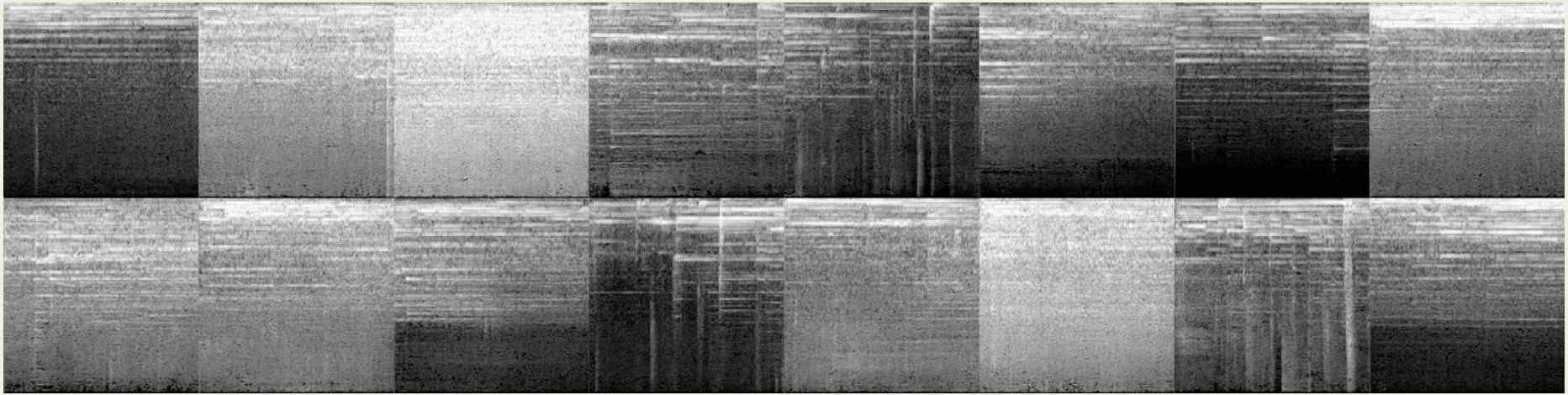
Training Set: Spotify Sleep Dataset



1000 steps

Experiment

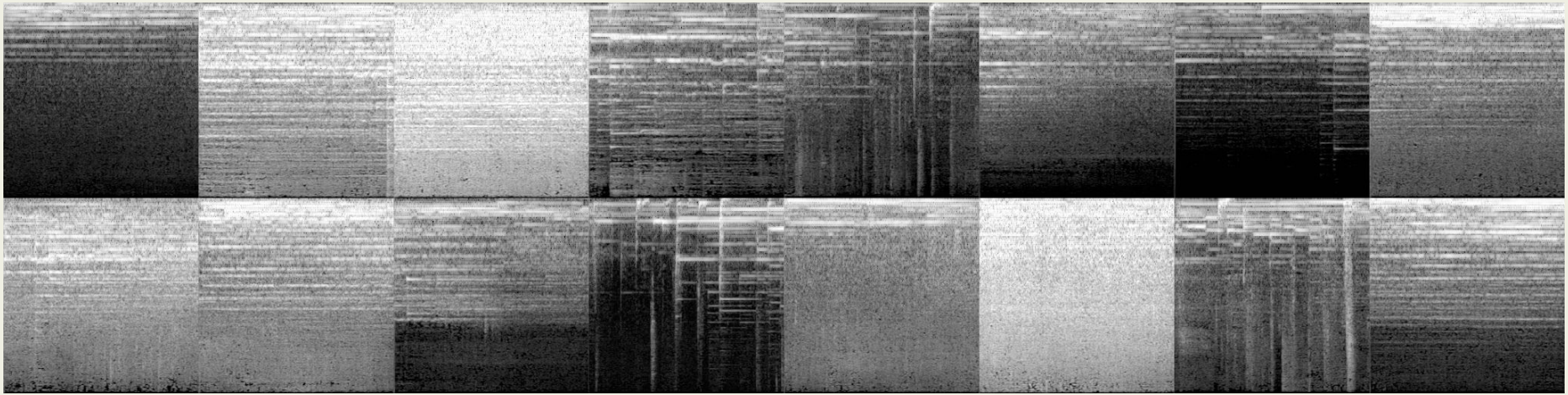
Training Set: Spotify Sleep Dataset



2500 steps

Experiment

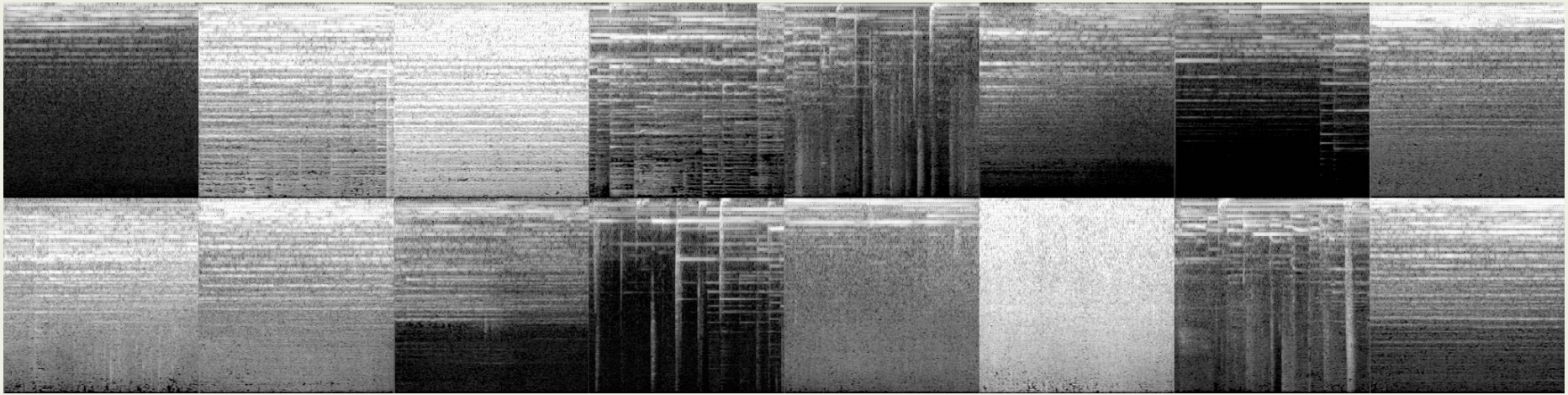
Training Set: Spotify Sleep Dataset



5000 steps

Experiment

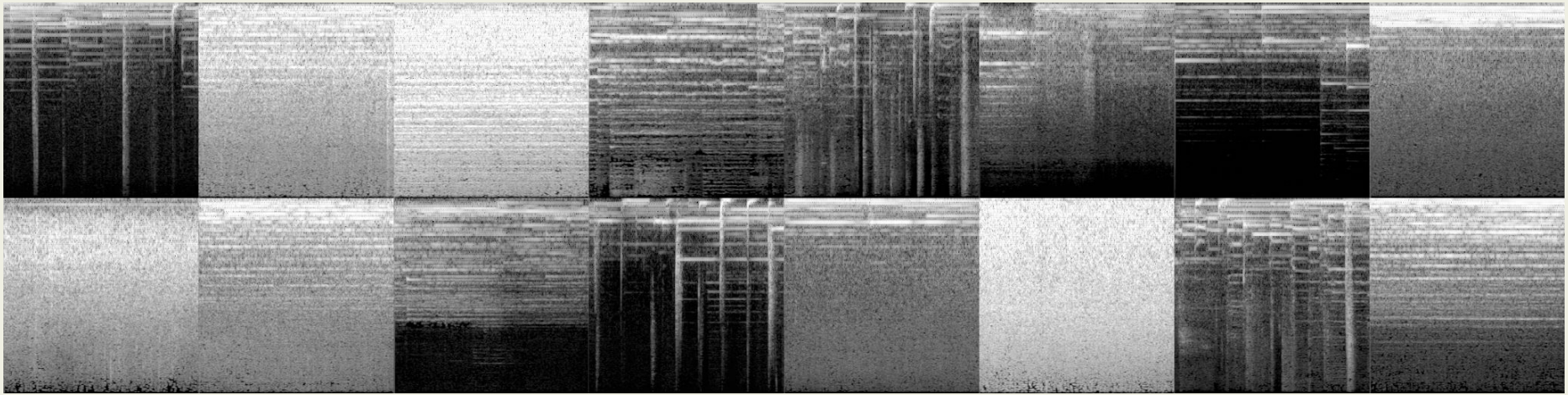
Training Set: Spotify Sleep Dataset



10000 steps

Experiment

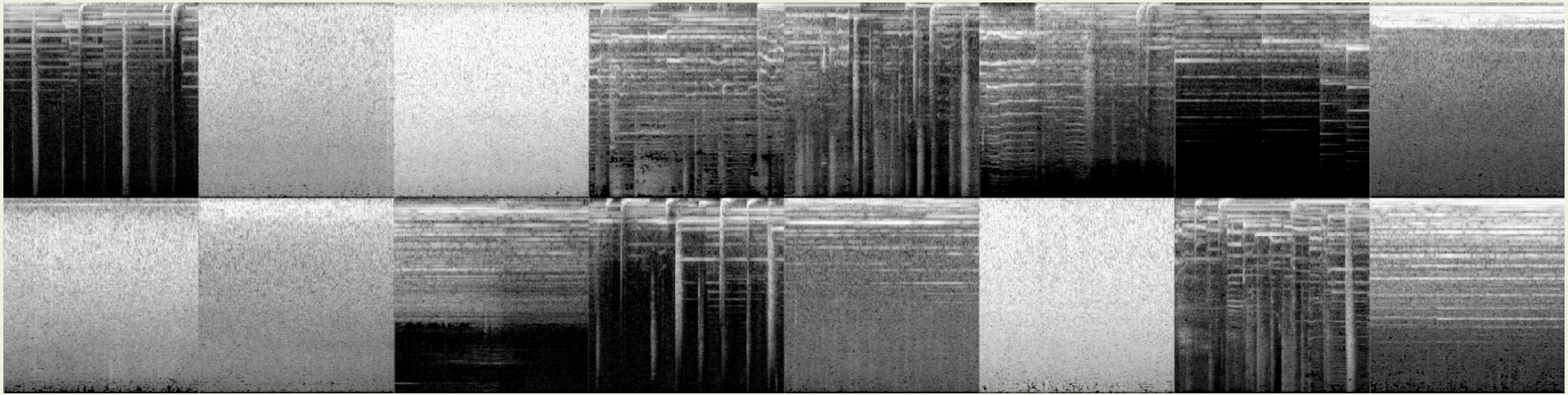
Training Set: Spotify Sleep Dataset



20000 steps

Experiment

Training Set: Spotify Sleep Dataset



40000 steps

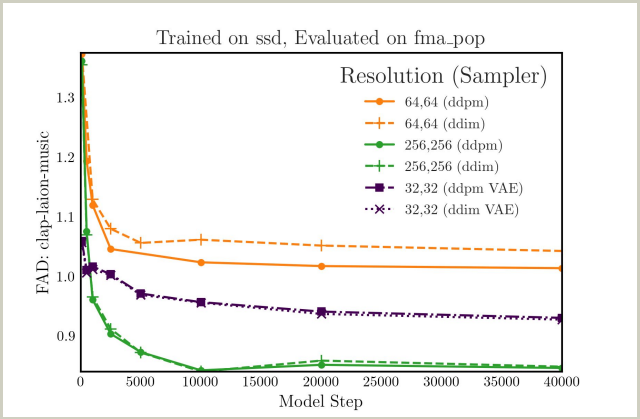
Experiment - FAD evaluation

FAD Embedding Model: clap-laion-music

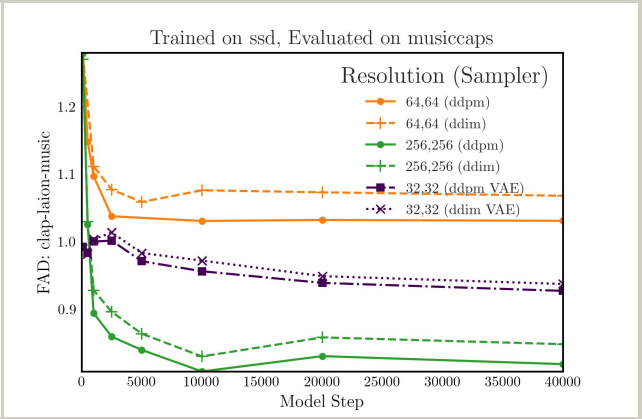
Training Set: Spotify Sleep Dataset

Evaluation Sets:

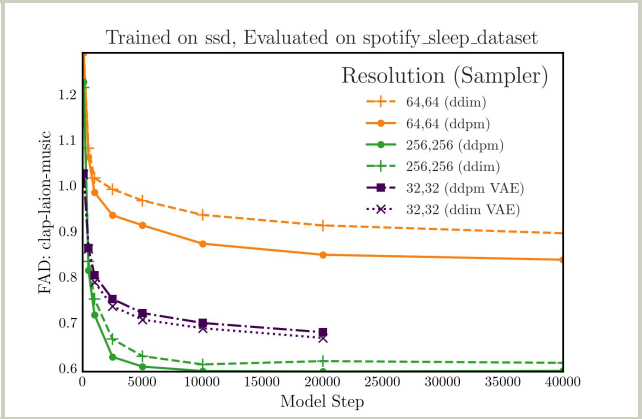
FMA Pop



MusicCaps



Spotify Sleep Dataset



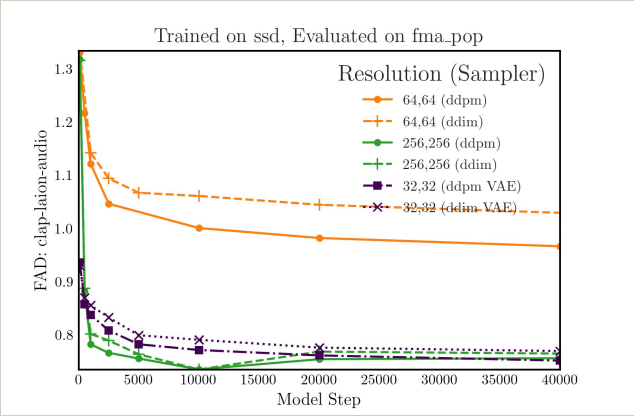
Experiment - FAD evaluation

FAD Embedding Model: clap-laion-audio

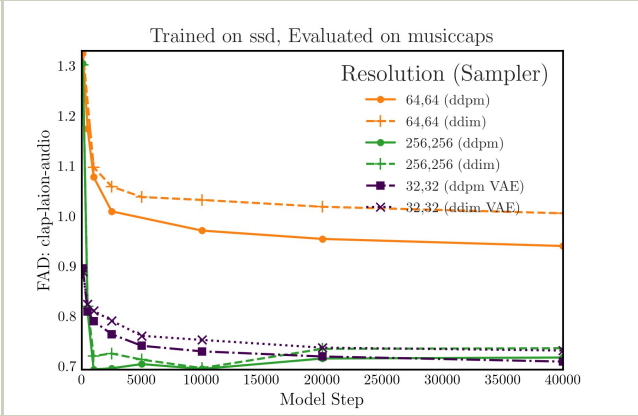
Training Set: Spotify Sleep Dataset

Evaluation Sets:

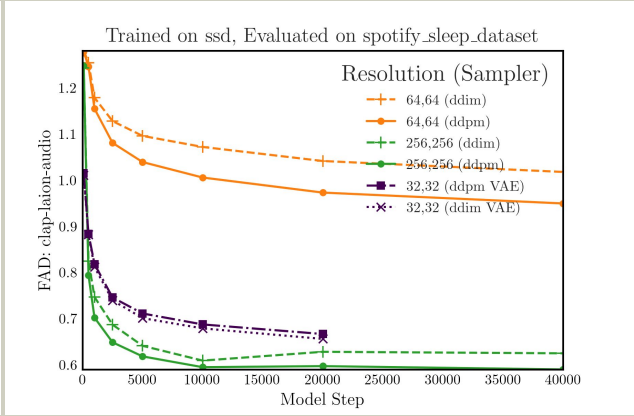
FMA Pop



MusicCaps



Spotify Sleep Dataset



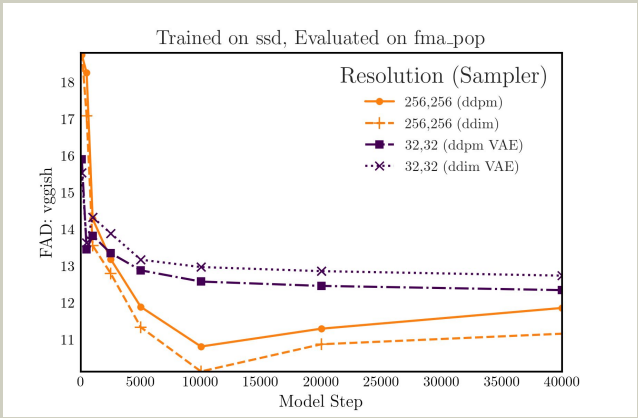
Experiment - FAD evaluation

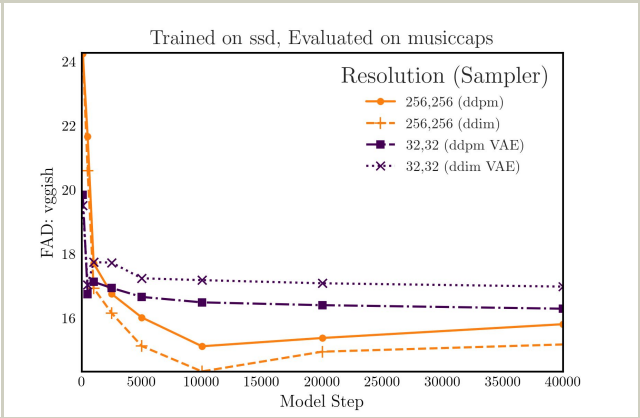
FAD Embedding Model: vggish

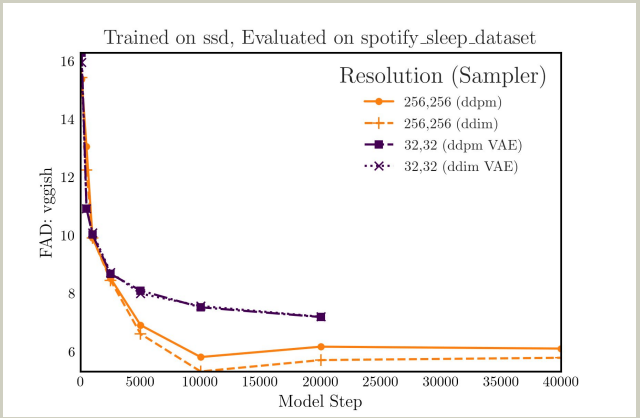
Training Set: Spotify Sleep Dataset

Evaluation Sets:

FMA Pop | MusicCaps | Spotify Sleep Dataset

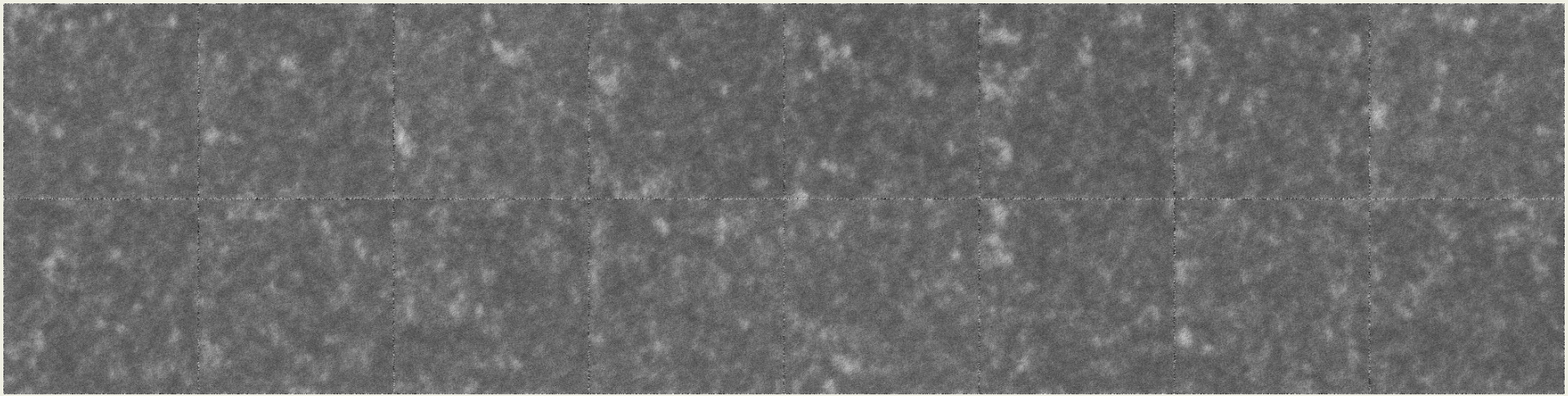






Experiment

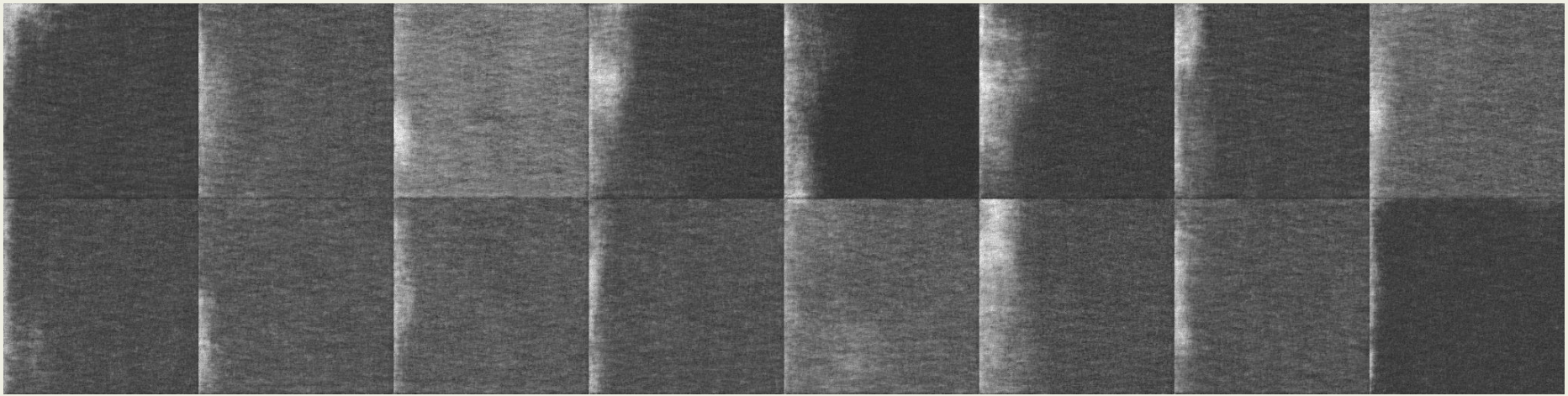
Training Set: Drum Samples



100 steps

Experiment

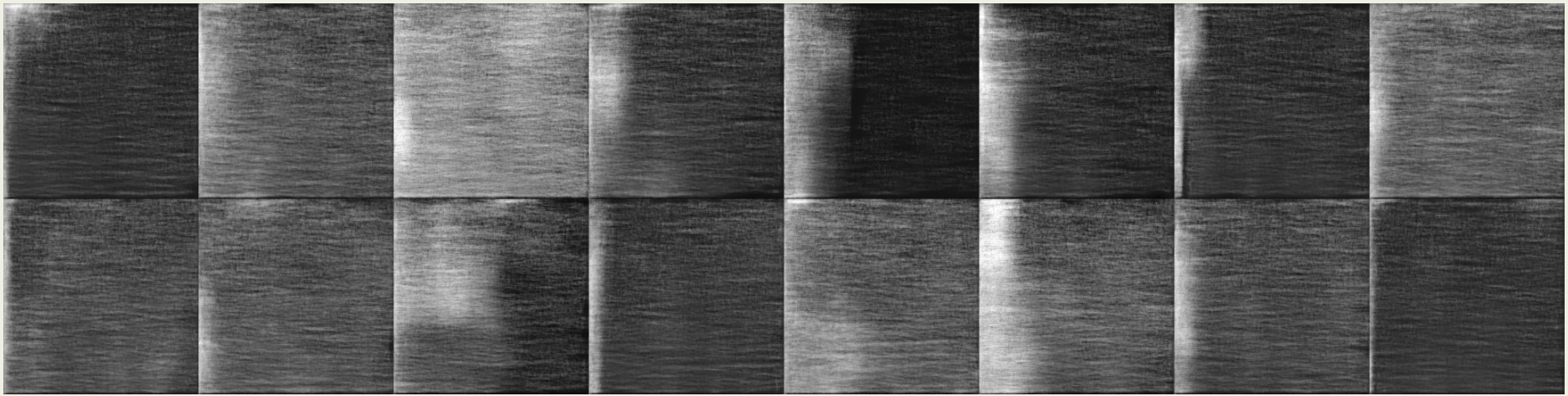
Training Set: Drum Samples



500 steps

Experiment

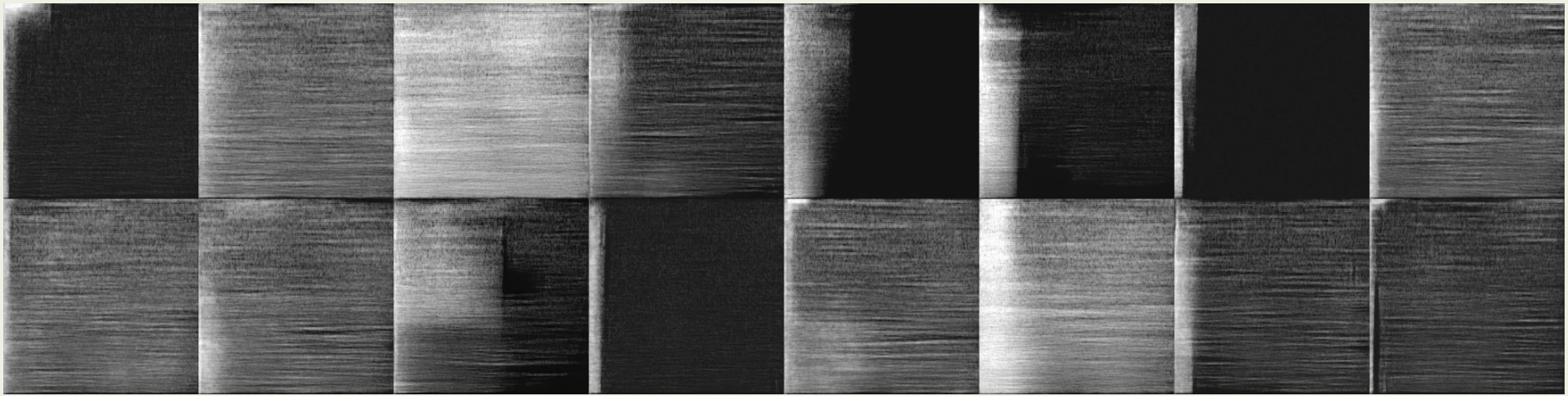
Training Set: Drum Samples



1000 steps

Experiment

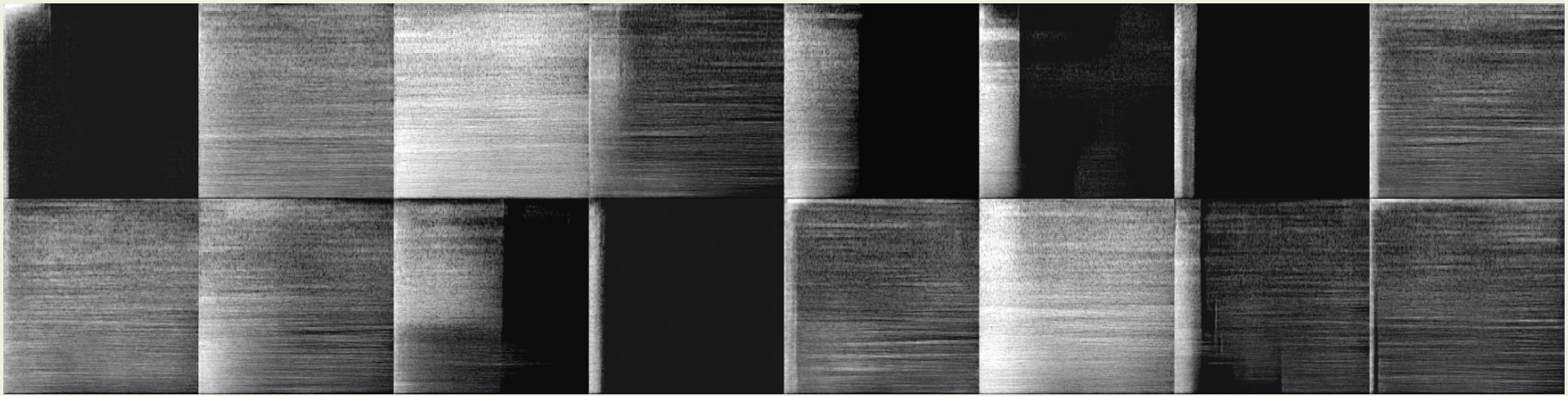
Training Set: Drum Samples



2500 steps

Experiment

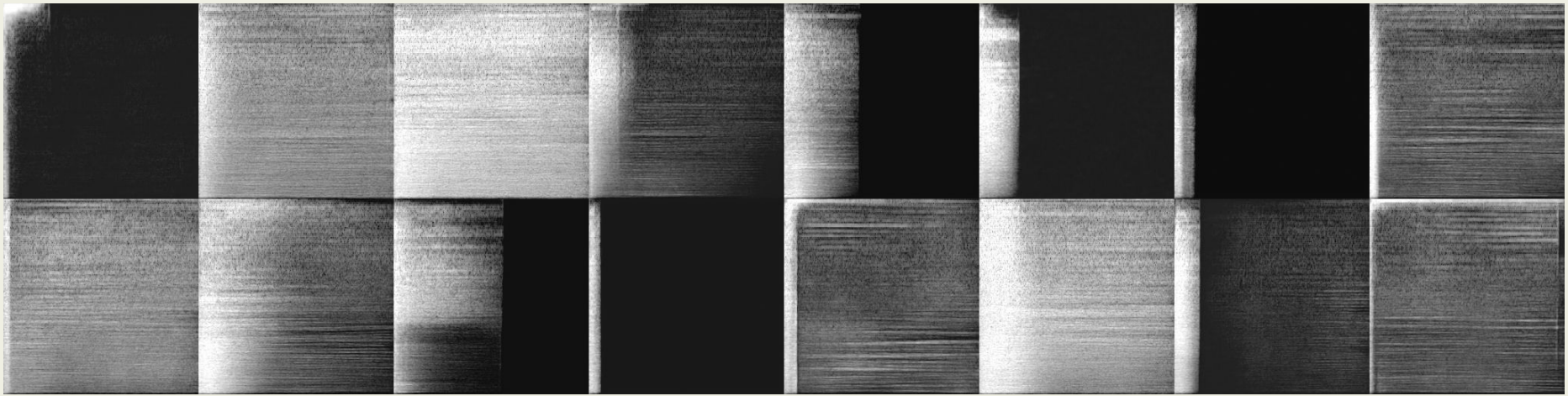
Training Set: Drum Samples



5000 steps

Experiment

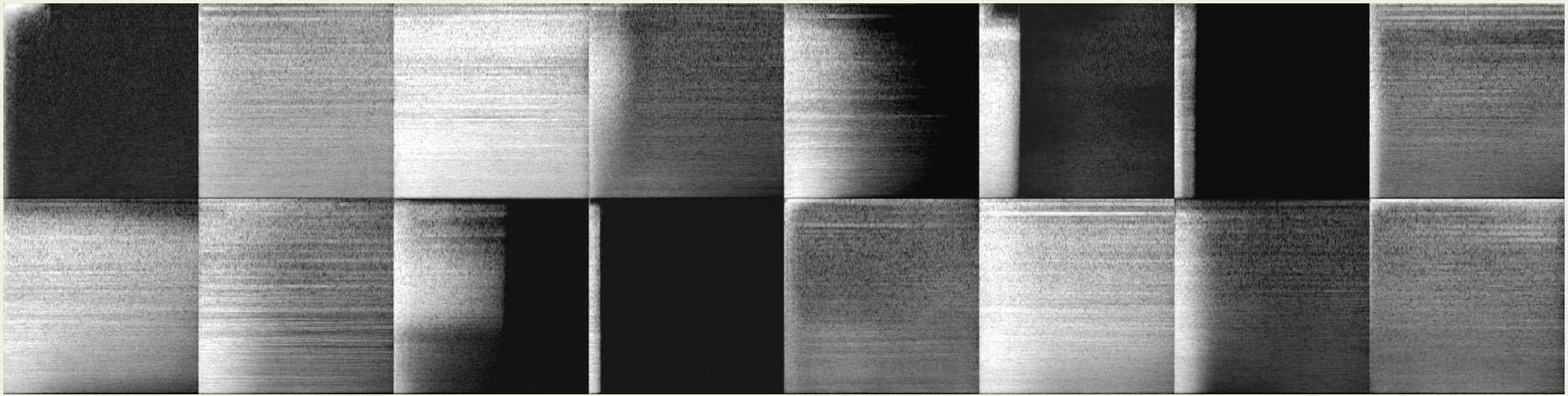
Training Set: Drum Samples



10000 steps

Experiment

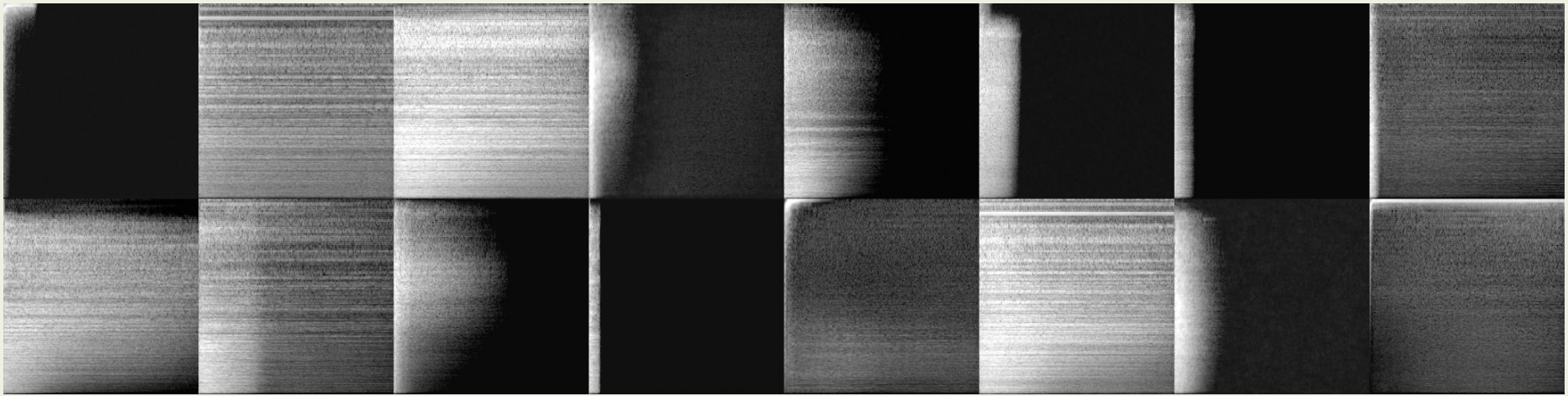
Training Set: Drum Samples



20000 steps

Experiment

Training Set: Drum Samples



40000 steps

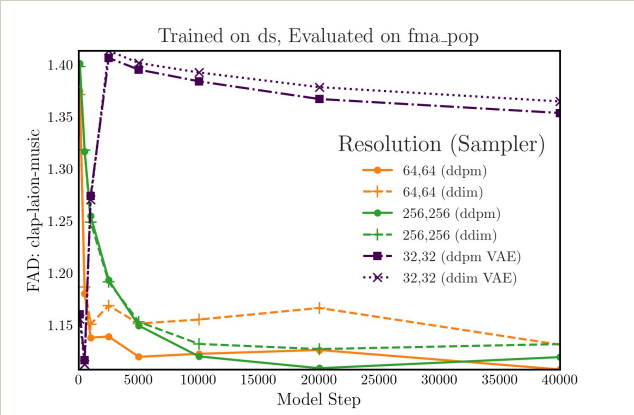
Experiment - FAD evaluation

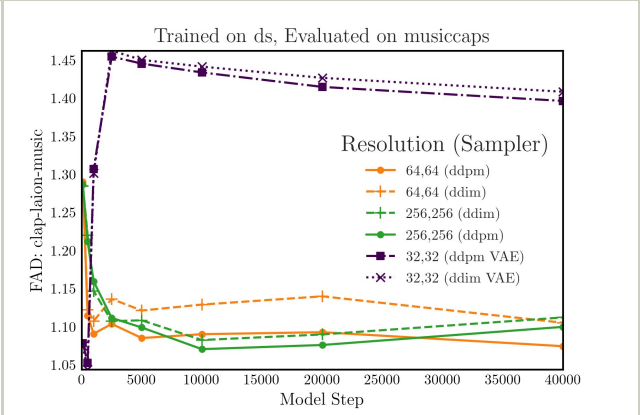
FAD Embedding Model: clap-laion-music

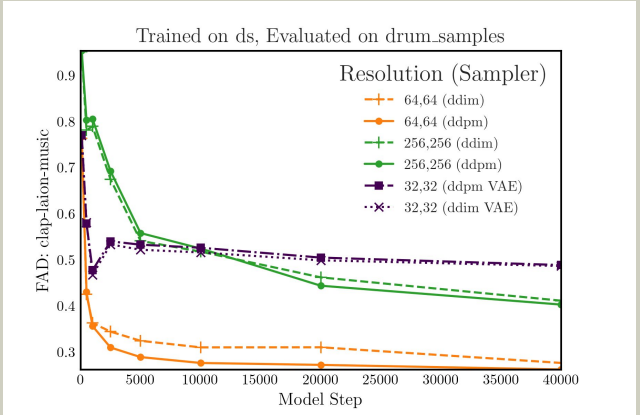
Training Set: Drum Samples

Evaluation Sets:

FMA Pop | MusicCaps | Drum Samples







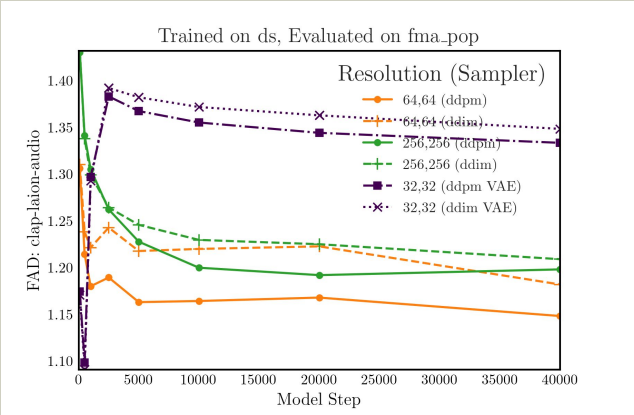
Experiment - FAD evaluation

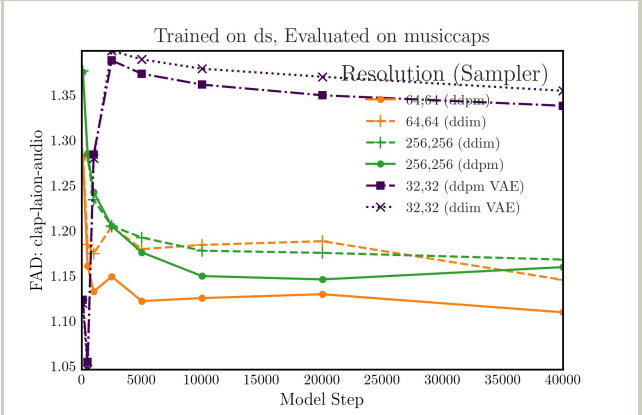
FAD Embedding Model: clap-laion-audio

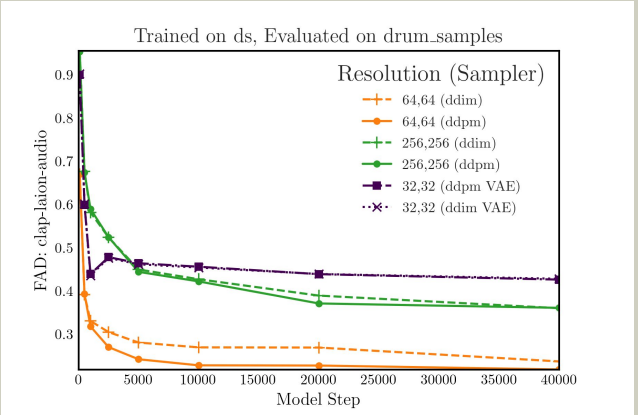
Training Set: Drum Samples

Evaluation Sets:

FMA Pop | MusicCaps | Drum Samples







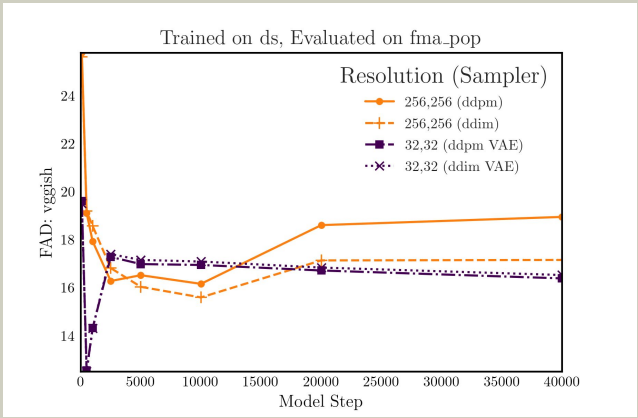
Experiment - FAD evaluation

FAD Embedding Model: vggish

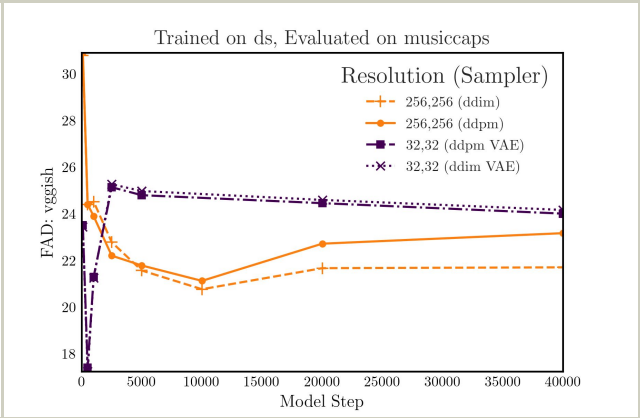
Training Set: Drum Samples

Evaluation Sets:

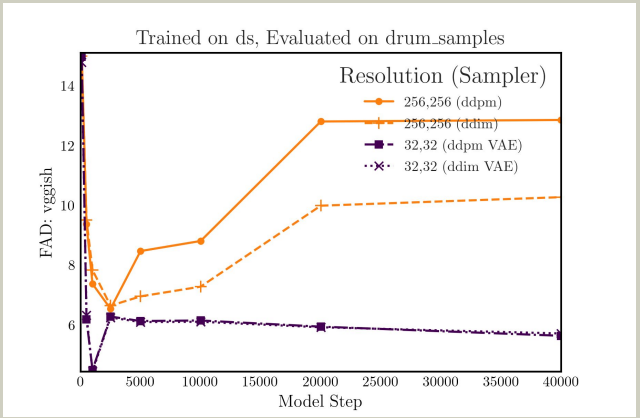
FMA Pop



MusicCaps



Drum Samples



Experiment - Speed Comparison

Experiment - Speed Comparison

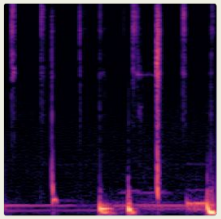
Data Type	Resolution	Params	Size (MB)	Sample Length (s)	Training Time 40k steps	Real-time Inference DDPM (s)	Real-time Inference DDIM (s)
Waveform	[1, 65536]	64M	259	2.96	14.41		
Mel-Spec	[64, 64]	113M	434	0.74	~1.16	1.162	**0.0743**
Mel-Spec	[256, 256]	113M	434	2.96	~12.82	3.784	**0.370**
Mel-Spec	[512, 32]	113M	434	5.93	3.88	**0.528**	**0.0474**
Mel-Spec	[512, 64]	113M	434	5.93	7.059	**0.951**	**0.092**
Mel-Spec	[512, 128]	113M	434	5.93	13.93	1.894	**0.187**
Mel-Spec	[1024, 128]	113M	434	11.86	15.01	1.876	**0.177**
VAE Latent Space	[256, 256] [32, 32]	113M	434	2.96	~2.99	**0.116**	**0.008**
VAE	[32, 32]	83M	319				
T5-Small (Encoder)	[1, 512]	35M	135				

Experiment - Noisy Samples

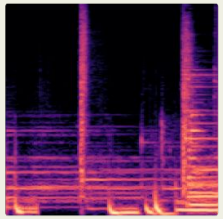
Experiment - Noisy Samples

Mean decibel (dB) value

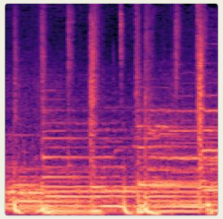
-72.18



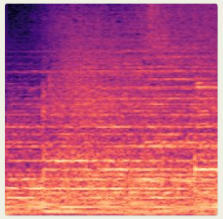
-61.74



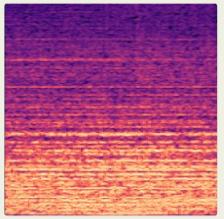
-45.33



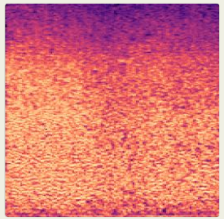
-34.66



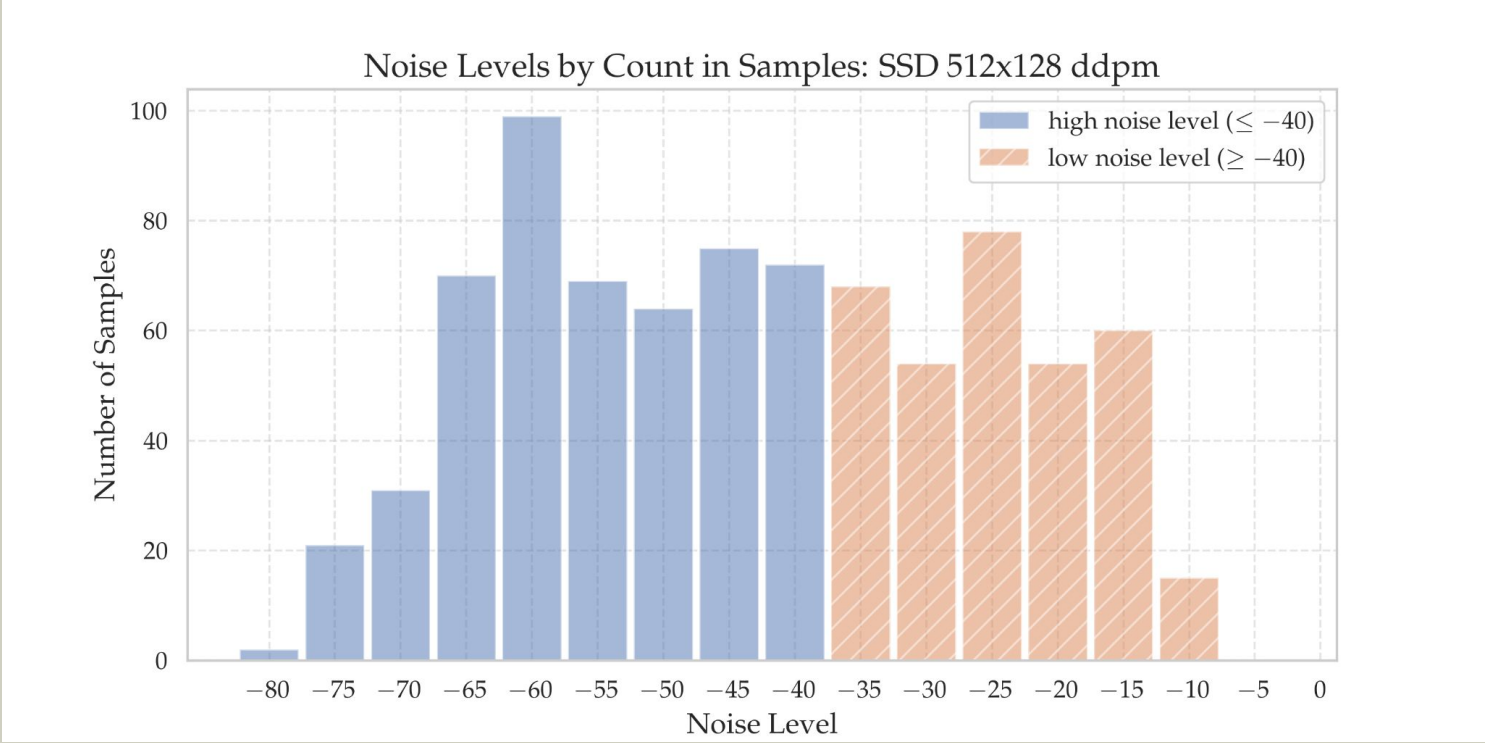
-21.17



-11.35



Experiment - Noisy Samples



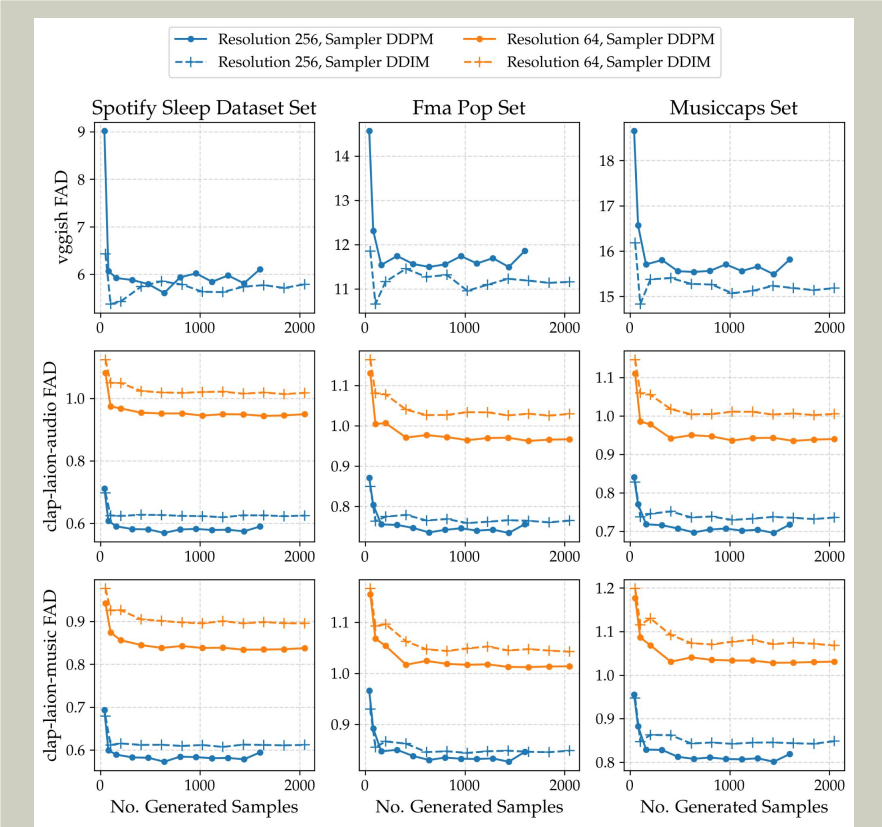
Experiment - Noisy Samples

Reference Dataset	Sampler	FAD Score ↓ (original/filtered)		
		VGGish	clap-laion-audio	clap-laion-music
Spotify Sleep Dataset	ddim	6.280/**5.373**	0.657/**0.579**	0.609/**0.571**
	ddpm	7.165/**5.026**	0.583/**0.509**	0.560/**0.529**
FMA Pop	ddim	10.706/**10.669**	0.774/**0.737**	**0.823**/0.831
	ddpm	12.464/**10.611**	0.724/**0.676**	**0.798**/0.806
Musiccaps	ddim	**15.228**/16.143	0.770/**0.731**	**0.830**/0.858
	ddpm	16.258/**15.922**	0.712/**0.664**	**0.782**/0.818

Table 5.2: Original FAD results (left) vs. the FAD results when noisy/loud samples are filtered out (right). Model used: **SSD 512x128**

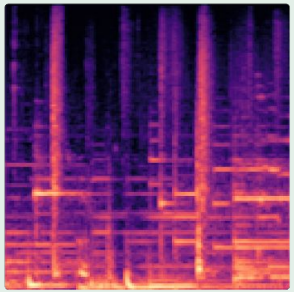
Experiment - Number of FAD Samples

Experiment - Number of FAD Samples

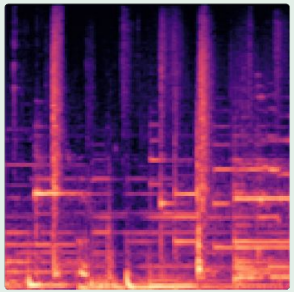


7. Creative Tools

Creative Tools - Looping

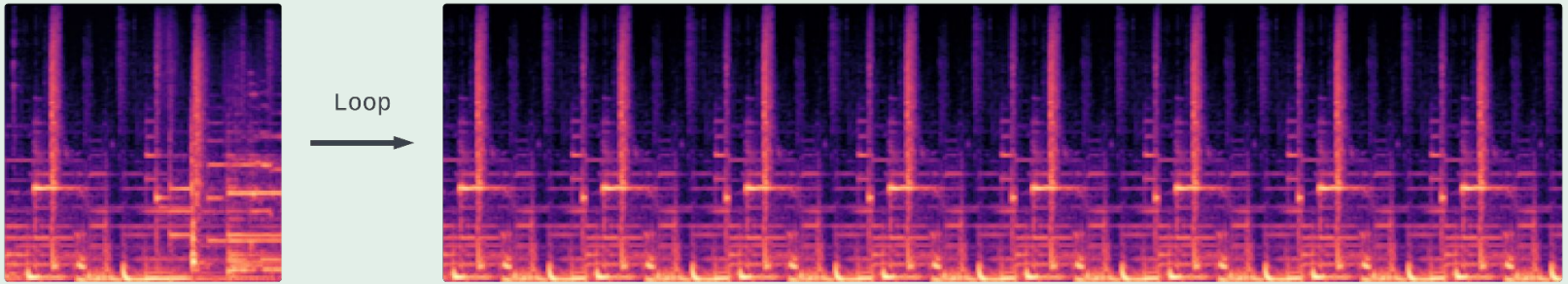


Creative Tools - Looping



Loop

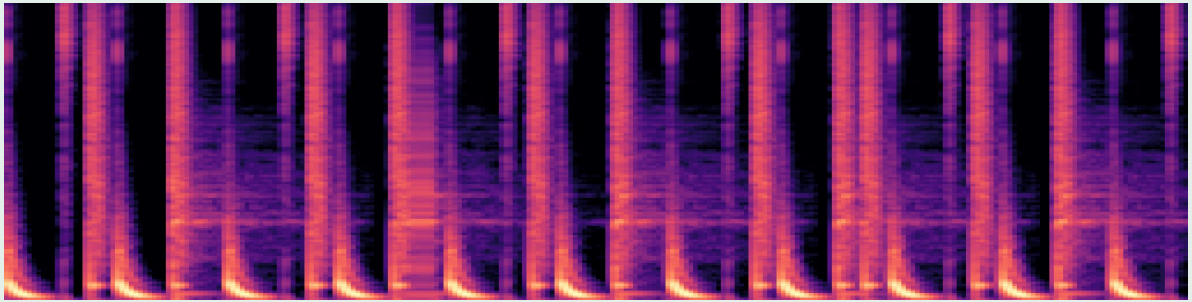
Creative Tools - Looping



Creative Tools - Music Variations

Reggaeton

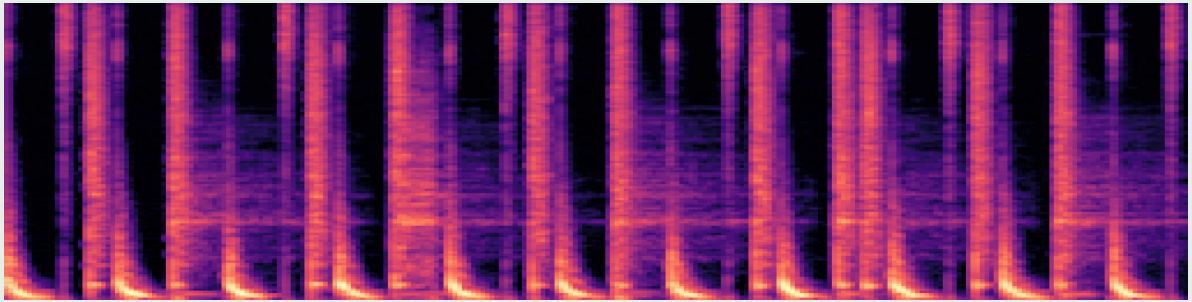
Original



Creative Tools - Music Variations

Reggaeton

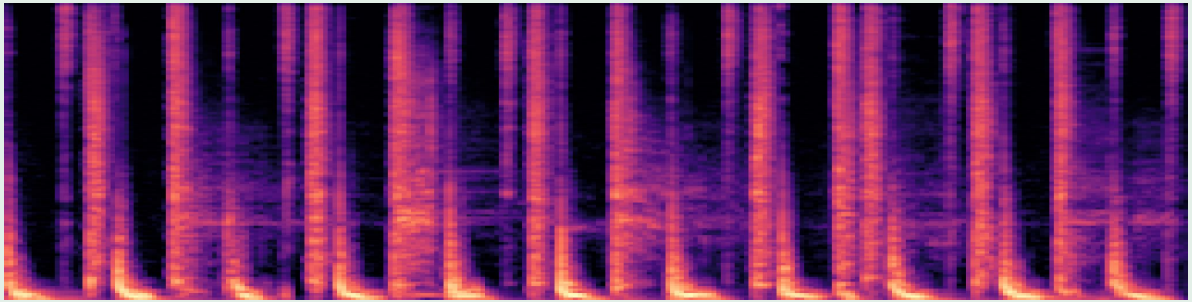
T=950



Creative Tools - Music Variations

Reggaeton

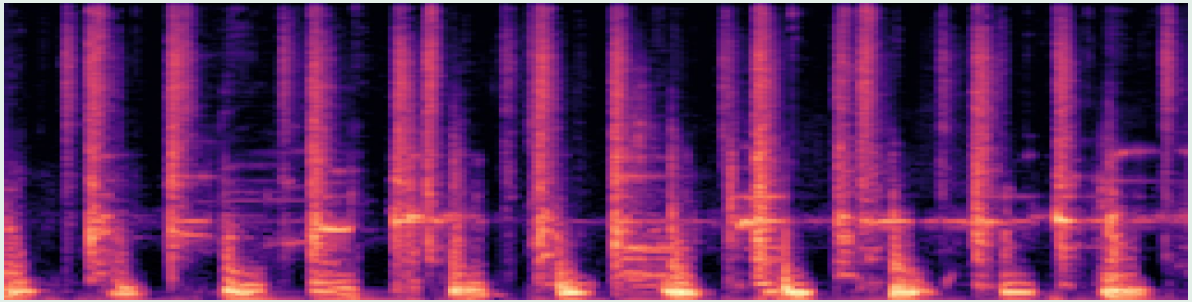
T=750



Creative Tools - Music Variations

Reggaeton

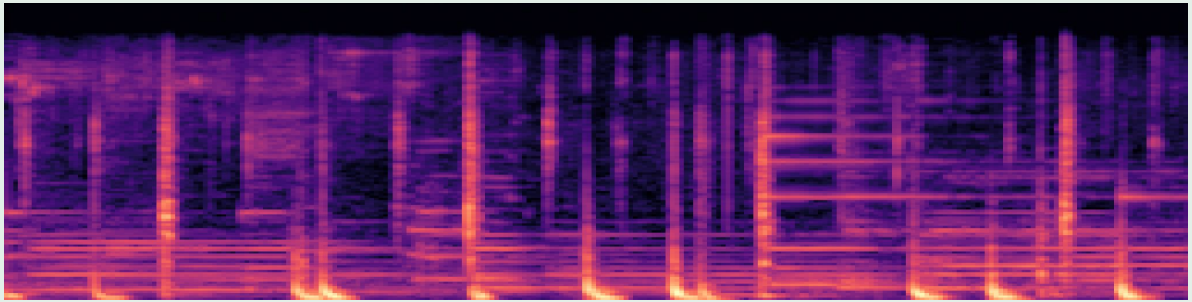
T=500



Creative Tools - Music Variations

Reggaeton

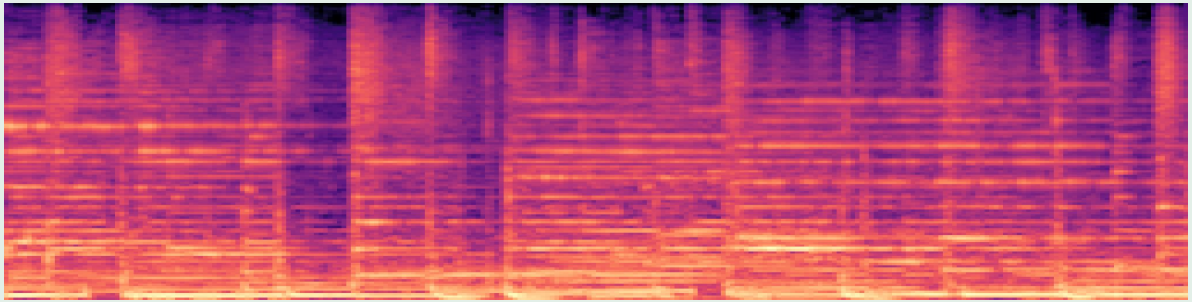
T=250



Creative Tools - Music Variations

Reggaeton

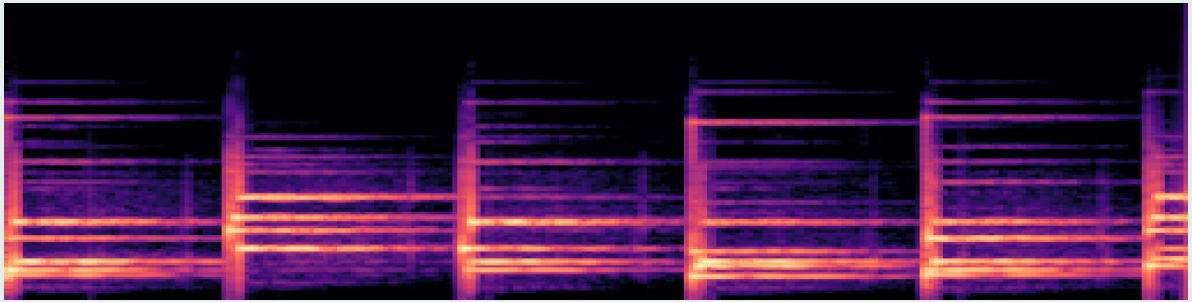
T=0



Creative Tools - Music Variations

R&B

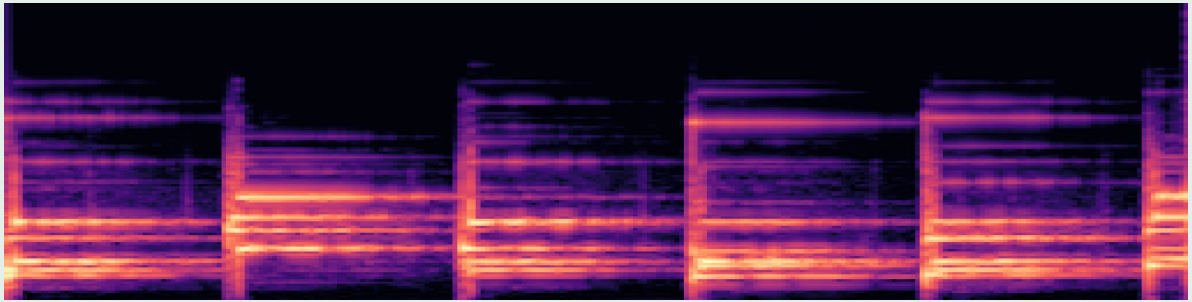
Original



Creative Tools - Music Variations

R&B

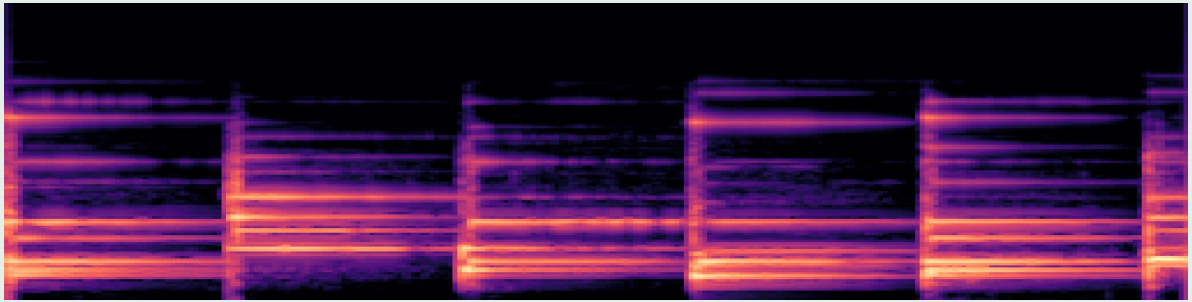
T=950



Creative Tools - Music Variations

R&B

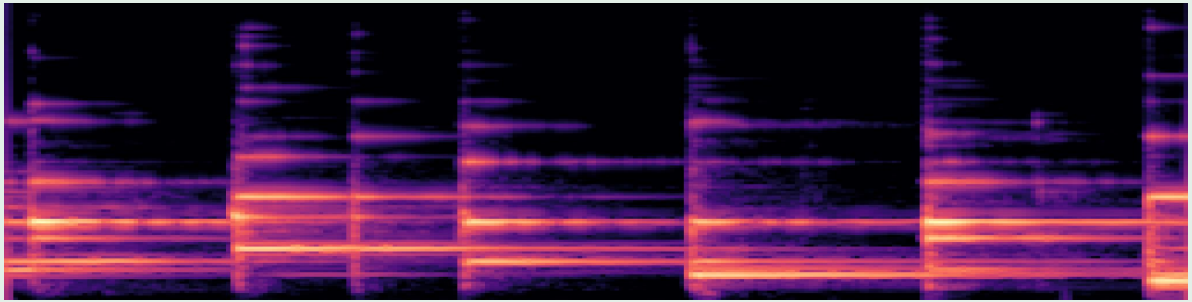
T=750



Creative Tools - Music Variations

R&B

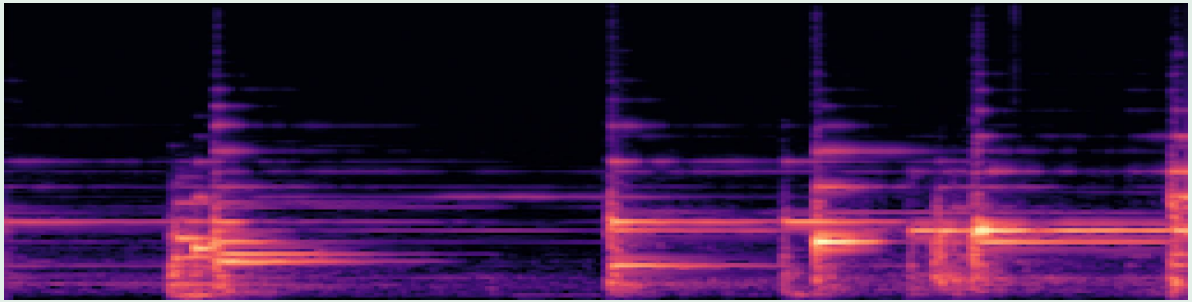
T=500



Creative Tools - Music Variations

R&B

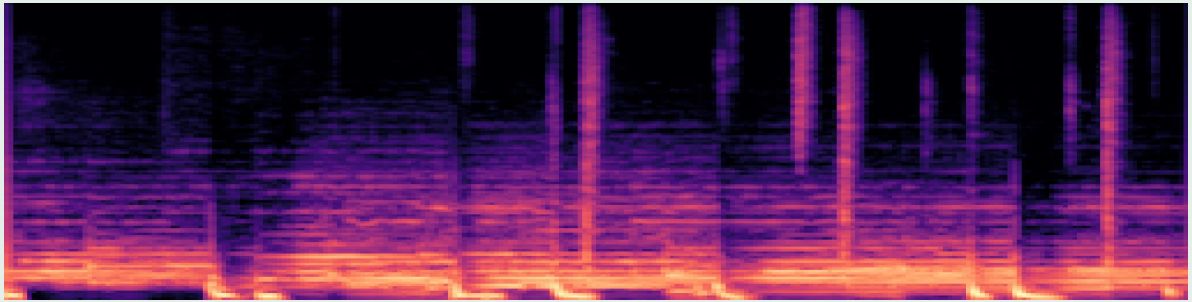
T=250



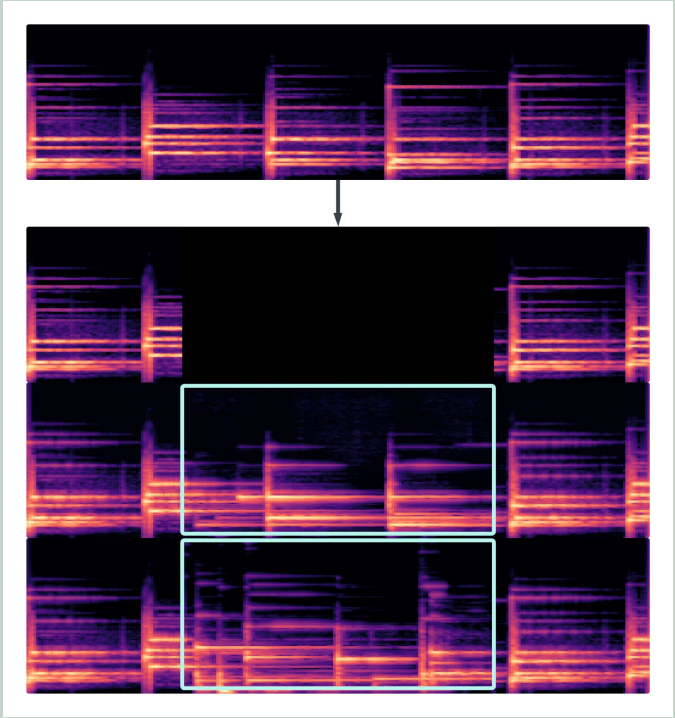
Creative Tools - Music Variations

R&B

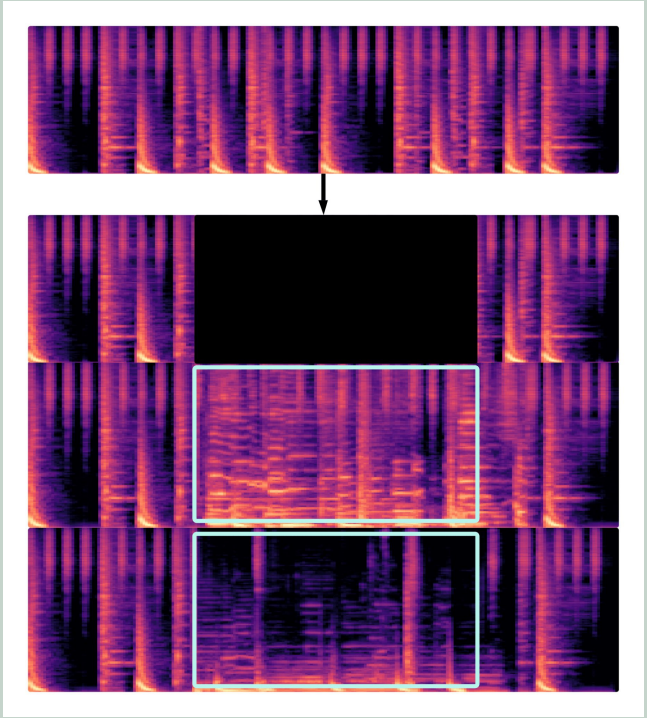
T=0



Creative Tools - Inpainting



R&B

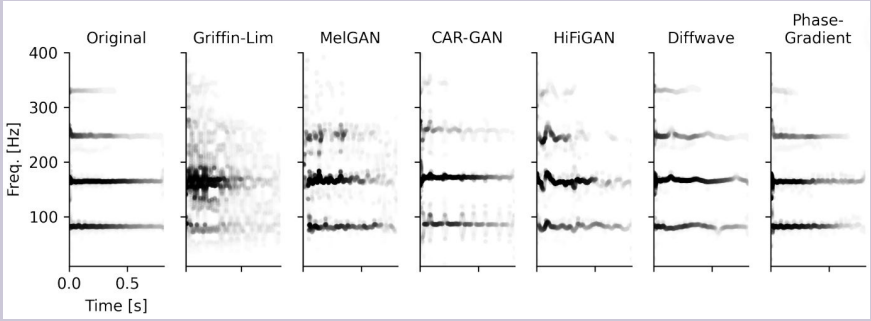


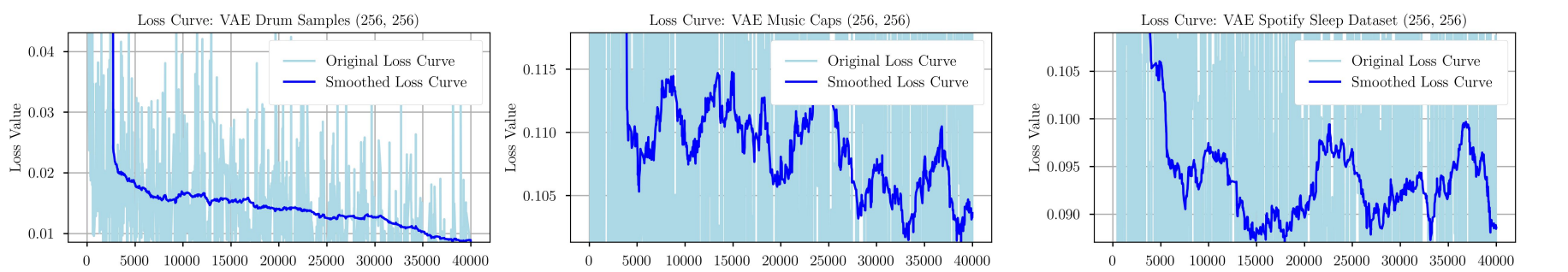
Hip Hop Beat

8. Limitations

Limitations

- More latent space experiments
- Neural vocoder
- Qualitative Analysis
- Loss Monitoring/Training





9. Discussion + Conclusion

Discussion + Conclusion

Lack of labelled training data

Ethics

Lack of evaluation consistency (datasets, embedding models, qualitative analysis, training/inference speed)

Diffusion models are **EXPENSIVE**

Free lo-fi samples up on BandCamp!

https://lofimphil.bandcamp.com/album/**mphil-lo-fi**

Thank you!

Questions? :)