

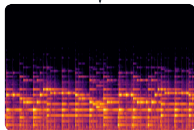
Training Waveforms
(Sleep Music)



$$\mathbf{X}_w \in \mathbb{R}^{s_r \times T_s}$$

Waveform
Processor
Component

STFT



$$\mathbf{X}_m \in \mathbb{R}^{\frac{T_{px}}{r} \times \frac{F_{mb}}{r}}$$

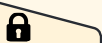
Training

Inference

VAE Component

Diffusion Component

VAE encoder



$$\mathbf{z} \in \mathbb{R}^{\frac{T_{px}}{r} \times \frac{F_{mb}}{r}}$$

$$q(\mathbf{x}_t | \mathbf{x}_{t-1})$$

Diffusion

$$\mathbf{z}_N \sim \mathcal{N}(0, \mathbb{I})$$

Denoising

$$p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)$$

U-Net

VAE decoder



Vocoder
Component

Neural Vocoder
or
Griffin-Lim

New Sleep Music

