

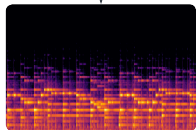
Training Waveforms
(Sleep Music)



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

Waveform
Processor
Component

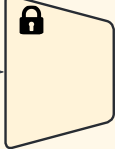
STFT



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

VAE Component

VAE encoder



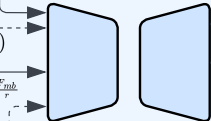
$$q(x_t|x_{t-1})$$

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

x1000

x1

U-Net



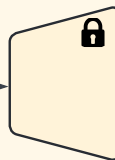
$$\mathbf{z}_N \sim \mathcal{N}(0, \mathbb{I})$$

Diffusion Component

Denoising

$$p_\theta(x_{t-1}|x_t)$$

VAE decoder



New Sleep Music



Vocoder
Component

Vocoder
or
Griffin-Lim

