

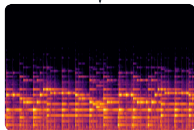
Training Waveforms
(Sleep Music)



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

Waveform
Processor
Component

STFT



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

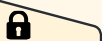
Training

Inference

VAE Component

Diffusion Component

VAE encoder



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

Denoising

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

U-Net

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

Diffusion

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

VAE decoder



Vocoder
Component

Neural Vocoder
or
Griffin-Lim

New Sleep Music

