A Novel Diffusion Model Based Approach for Sleep Therapeutic Music Generation

Anonymous submission

Editors: D. Herremans, K. Bhandari, A. Roy, S. Colton, M. Barthet

Abstract

Sleep disorders, particularly insomnia, and mental health conditions affect a significant fraction of adults worldwide, posing serious mental and physical health risk. Music therapy offers promising, low-cost, and non-invasive treatment, but current approaches rely heavily on expert-curated playlists, limiting scalability and personalization. We propose a low-cost generative system leveraging recent advances in diffusion models to synthesize music for therapy. We focus on insomnia and curate a dataset of waveform sleep music to generate audio tailored to sleep. To ensure real-world feasibility, we optimize our system for training and use on a single GPU, balancing quality and efficiency through extensive ablation studies. We show through subjective human evaluations that our generated music matches or outperforms existing baselines in both perceived quality and relevance to sleep therapy, while using only a fraction of the computational cost.

Keywords: Diffusion Models, Music Therapy, Insomnia, Mental Health, Sleep Music Generation

1. Introduction

It is estimated nearly a third of the world population suffers from insomnia (Bhaskar et al., 2016), with one in ten suffering from chronic insomnia (Ellis et al., 2023; Riemann et al., 2022). Mental health conditions such as depression and anxiety are also known to be comorbid with insomnia (Palagini et al., 2022; Blank et al., 2015), and together they contribute to many lifestyle diseases with severe impact on health such as lessened metabolic, immunologic, and cardiovascular health. Additionally, sleep disruption increases susceptibility to numerous chronic conditions (Ramar et al., 2021; Medic et al., 2017; van Cauter et al., 2007; Gebara et al., 2018). The progressive decline in sleep duration in the public is considered a public health epidemic (Consensus Conference Panel et al., 2015) and most common treatments for insomnia are drug based, which have often been associated with adverse side effects.

Music therapy is emerging as an effective, non-invasive, and low-cost solution in the treatment of sleep and mental health disorders (Huang et al., 2023a; Mohamad Zamani et al., 2022; Jespersen et al., 2015; Lee and Thyer, 2013), as there is strong evidence that music can improve sleep quality (Jespersen et al., 2015; Dickson and Schubert, 2020; Loewy, 2020; Wang et al., 2014) as well as the recovery approach in mental health care (Lee and Thyer, 2013). However, producing music for therapy on-demand is underexplored (Richter, 2019). Existing systems are static, lacking the adaptability to cater to individual needs. Moreover, these approaches rely on expert therapists to curate music selections, limiting their accessibility and scalability for broader public use.

To address the up-and-coming field of music therapy generation, this paper focuses on low-cost sleep music generation that can be both trained and deployed efficiently, making

it accessible for wider use, and operates on publicly available sleep music. To summarize, our main contributions are as follows: **(a)** we are the first to focus on generative sleep therapy using waveform audio, a more expressive and available audio data format, that allows for greater adaptability to personal preferences, surpassing the limitations of previous MIDI-based or algorithm-based approaches; **(b)** we are the first to apply the BigVGAN vocoder to music generation, demonstrating its superior performance despite being originally trained for speech synthesis; **(c)** we propose multiple model architectures that offer various trade-offs between generation quality and computational efficiency, with our fastest models leveraging a latent diffusion framework combined with the BigVGAN vocoder. We specifically focus on training models on a single consumer GPU, employing a minimal latent diffusion architecture and carefully chosen hyperparameters to maximize efficiency without compromising performance.; **(d)** we curate a specialized dataset of sleep music and **(e)** through comprehensive objective and subjective evaluations, including human listener studies, we show that our models outperform existing methods in terms of music relevance to sleep therapy. We achieve these improvements while maintaining smaller model size and reduced computational requirements compared to other state-of-the-art music models.

2. Background in Computer Music Generation

Previously, music intervention studies often relied on generic sleep music with no expert curation (Loewy, 2020; Wang et al., 2014; Richter, 2019). While some studies improve upon this by involving expert-selected music (Dickson and Schubert, 2020; Jespersen et al., 2015), it has been shown that music preference is individualistic and could have an impact on music therapy (Chang et al., 2012; Yamasato et al., 2020). Moreover, there still remains a critical gap: across nearly all studies there is a severe lack of analysis on the specific sleep music characteristics (Loewy, 2020), including melody, harmony, rhythm, etc. Implementing personalized music interventions on a large scale remains costly and logistically challenging.

Music generation methods can be broadly categorized into MIDI-based and waveform-based approaches. Transformers (Vaswani et al., 2023) have been widely employed to sequentially model MIDI notes, treating the task as a language modeling problem (Huang and Yang, 2020; Qu et al., 2024; Zhang, 2023; Jiang et al., 2020). However, MIDI only provides instructions for synthesizing sound; it does not directly produce audio and requires additional processing through instruments or software synthesizers to generate sound, thus lacking the expressiveness and fidelity required for generating full, synchronized soundtracks. As such, MIDI datasets need to often be created manually, in contrast to being able to use readily available sleep music in the audio (waveform) format.

On the other hand, learning from raw audio (waveform) provides both flexibility and ease of access to training data. Instead of needing to build music from MIDI instructions, music can be directly created from waveforms, allowing for more naturally sounding and expressive results. Additionally, this approach benefits from a large amount of sleep music available as waveforms, whereas MIDI versions of these songs are rarely accessible.

Specifically for sleep music, only two music generation approaches have been put forth so far. The first uses MIDI-based music generation (Yang et al., 2022), while the other utilizes an algorithmic approach using randomization and Markov Chains to construct music elements (Tulilaulu et al., 2012). Thus, no works up to date have focused on generating

expressive sleep music from waveform audio. As a result, existing methods lack the ability to produce music that is both highly expressive, natural-sounding, and adaptable to users' preferences, limiting their effectiveness. Nevertheless, we cover the most popular approaches in literature for general music generation.

Recently, diffusion models have shown exceptional performance in generative AI tasks involving waveform audio. Existing approaches in the literature can be categorized as follows: **(a)** Spectrogram-based Diffusion Models, which process spectrogram data using encoders such as VAEs and employ a neural vocoder to convert spectrograms back into audio (Chen et al., 2023; Liu et al., 2023; Huang et al., 2023b; Yang et al., 2023; Huang et al., 2023c), or operate directly on waveforms (Schneider et al., 2023; Schneider, 2023; Li et al., 2023); **(b)** Diffusion Transformers (DiT), which generate audio by operating in the latent space, and use a combined diffusion model and transformer approach (Agostinelli et al., 2023; Ning et al., 2025); **(c)** Auto-regressive Language Models (most commonly transformers), which generate audio based on trained codecs (Copet et al., 2023; Evans et al., 2024; Lan et al., 2024; Wu et al., 2024a,b); and **(d)** Alternative Generative Methods, such as Consistency Autoencoders (CAE) (Pasini et al., 2024).

However, in the field of music generation, few released models are able to accommodate use on a single consumer GPU, often requiring large amounts of compute to train and run for inference. Our research addresses this limitation by developing lightweight models that can be trained and deployed for inference on a single consumer GPU.

3. Proposed Approach

Our framework consists of four main components: **(a)** the Waveform Processor, **(b)** Variational Autoencoder (VAE), **(c)** Diffusion Model, **(d)** and Vocoder. Figure 1 depicts the overall pipeline of the proposed approach.

3.1. Framework Components

The **Waveform Processor (a)** starts the pipeline by taking raw audio waveforms $x \in \mathbb{R}^{T_s}$ from the training dataset, where T_s is the length of the audio signal in samples (e.g. an audio file with a sampling rate of 44.1kHz has 44,100 samples per second). To optimize computational efficiency, the audio is downmixed to mono and resampled to 22 kHz, unless stated otherwise. The processed waveform is then transformed into a mel-spectrogram $\mathbb{X}_{\text{m}} \in \mathbb{R}^{T_{px} \times F_{mb}}$, where T_{px} is the time resolution and F_{mb} is the number of mel-frequency bins. The mel-spectrograms are sliced to the appropriate length for the next stage.

Next, the **Variational Autoencoder (VAE) (b)** encodes mel-spectrograms into a compact latent space, enabling efficient training and inference. Specifically, $\mathbb{X}_{\text{m}} \in \mathbb{R}^{T_{px} \times F_{mb}}$ is encoded into $\mathbf{z} \in \mathbb{R}^{C \times \frac{T_{px}}{r} \times \frac{F_{mb}}{r}}$, where C is the latent channel size and r is the downsampling factor. The decoder reverses this process, reconstructing a mel-spectrogram from the latent representation. Once the VAE is trained, its parameters are frozen during the training and inference stages of the diffusion model.

Then, the **Diffusion Model (c)** generates novel samples $\mathbf{z}' \in \mathbb{R}^{C \times \frac{T_{px}}{r} \times \frac{F_{mb}}{r}}$ from Gaussian noise, either in the pixel space (if no VAE is used), or in latent space (if the VAE is used). Using a U-Net backbone, the model progressively denoises the input noise through a reverse

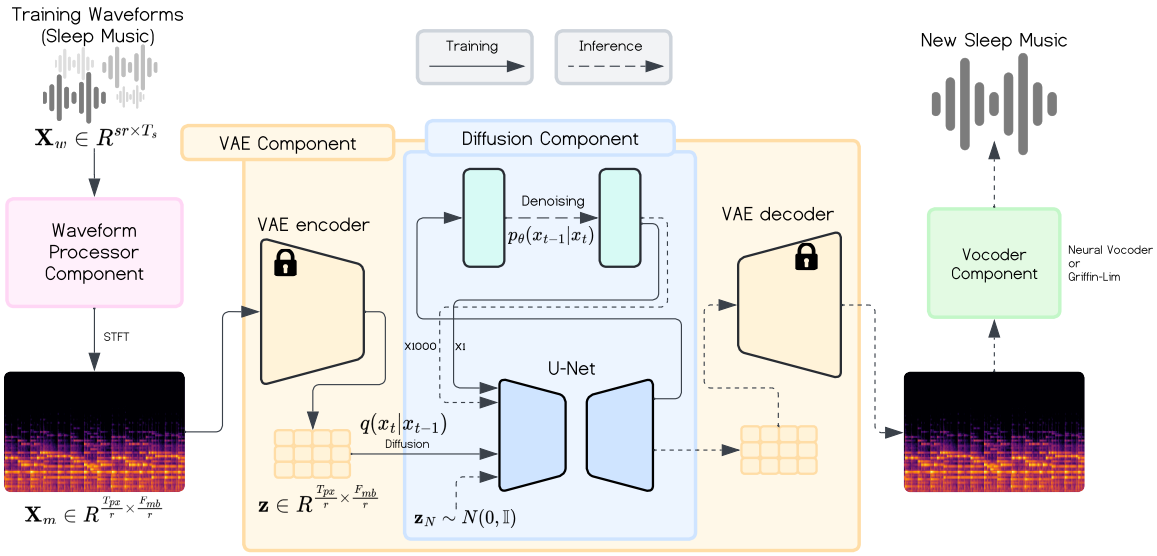

Figure 1: Diagram of our proposed framework, which consists of four components: Waveform Processor, VAE, Diffusion, and Vocoder. Raw audio (waveform) of length T_s from our sleep dataset is downsampled to sr = 22kHz and converted to mono. It is then transformed into mel-spectrograms (X_m) with a time (x) resolution of T_{px} and a frequency (y) resolution of F_{mb}. A pre-trained VAE encodes the mel-spectrogram into a compressed 2-dimensional latent space (z) with a compression ratio r. During training (bold arrows), the diffusion model processes the batch once to predict the added noise in training samples, computing the loss for backpropagation. During inference (dashed arrows), the process starts with Gaussian noise and performs 1000 steps of ancestral sampling to generate a novel sample, which is passed through the VAE decoder to reconstruct the mel-spectrogram. Finally, the Vocoder converts the mel-spectrogram back into audio.

diffusion process over N iterative steps. During training, the model learns to predict the noise added at each step of the forward diffusion process. In inference, it generates novel samples by iteratively reversing this process, starting from pure Gaussian noise.

Finally, the **Vocoder Component (d)** converts the generated mel-spectrograms $\mathbb{X}'_{\text{mel}} \in \mathbb{R}^{T_{px} \times F_{mb}}$ back into waveform audio $\mathbb{X}' \in \mathbb{R}^{T_s}$. Since phase information of mel-spectrograms is no longer present, this approach is not straightforward. We experiment with two approaches two make the conversion back to audio: the deterministic Griffin-Lim algorithm and the BigVGAN neural vocoder. The BigVGAN vocoder was pretrained on a diverse dataset, originally for speech synthesis, but we show it performs well for the music generation tasks.

Component	Hyperparameter	Value
Waveform Processor	Sampling Rate	22 kHz or 44 kHz
	Mel-spec Pixel Width	512 px or 2048 px
	Mel Bands	Variable
	Hop Length	Variable
	FFT Size	Variable
VAE	KL Divergence Reg. Term	1×10^{-6}
	Discriminator	After 50,001 steps
	Base Channel Width	32
	Channel Multipliers	$[1, 2, 4, 4]$
	Input/Output Channels	1
	Residual Blocks per Level	2
	Learning Rate	4.5×10^{-6}
	Batch Size	32
	EMA	inv. gamma $= 1.0$, power $= \frac{3}{4}$, max decay $= 0.9999$
	Compression Rate	Variable
Diffusion Model	Down Blocks	`[DownBlock, DownBlock, DownBlock, DownBlock, AttnDownBlock, DownBlock]`
	Up Blocks	`[UpBlock, AttnUpBlock, UpBlock, UpBlock, UpBlock, UpBlock]`
	Out Block Channels	$[128, 128, 256, 256, 512, 512]$
	Learning Rate	10^{-4} w/ 500 warmup steps
	Adam Optimizer	$\beta_1 = 0.95, \beta_2 = 0.999, \epsilon = 10^{-8}$, weight decay $= 10^{-6}$
	EMA	inv. gamma $= 1.0$, power $= \frac{3}{4}$, max decay $= 0.9999$
	Training Steps	1000 DDPM Steps
	Noising Parameters	$\beta_1 = 10^{-4}, \beta_T = 0.02$
	Noising Schedule	Cosine
	Inference Steps	1000 DDPM steps or 100 DDIM steps
	Batch Size	16 or 8
Vocoder	Neural Vocoder	bigvgan_v2_44khz_128band_256x
	Griffin-Lim Iterations	32

Table 1: Hyperparameters for each component. While most parameters are fixed throughout the paper following standard literature practices, others are the focus of the paper and vary during individual experiments and are marked as 'Variable'.

3.2. Training Dataset

To train our model, we selected a suitable dataset of sleep music. We used samples from the publicly available Spotify Sleep Playlist Dataset (Scarratt et al., 2023). After filtering, the dataset has 19,000 30-second samples, amounting to ∼158 hours of music. For brevity, this paper often refers to this dataset as `SSD` (**S**potify **S**leep **D**ataset).

3.3. Hyperparameters and Training Details

This section describes important hyperparameters and technical details of the relevant components used in our approach. An overview of the parameters is shown in Table 1.

4. Evaluation Setup

We perform objective and subjective evaluation for the generated music. We evaluate generated music using the Fréchet Audio Distance (FAD) metric and both objective and subjective methods when comparing to recent baseline models.

4.1. Objective Assessments

In our experiments, generated samples are evaluated using the Fréchet Audio Distance (FAD) metric (Kilgour et al., 2019). FAD measures the similarity between the distributions of two sets of audio tracks, one generated and one reference, by computing the Fréchet Distance (Heusel et al., 2018), also known as the Wasserstein-2 distance, between their respective embeddings. We leverage the library and incorporate improved parameters for FAD calculation from (Gui et al., 2024), addressing their recent findings that traditional FAD metrics may not reliably align with human judgment. Specifically, the widely used VGGish (Hershey et al., 2017) embedding model has been shown to inadequately reflect human perception. Instead, we leverage the CLAP music model backbone (Wu et al., 2024c), which offers a stronger alignment with human evaluations, and comes in two variations: `clap-laion-audio` and `clap-laion-music`. This is in line with other new works such as (Novack et al., 2024; Manor and Michaeli, 2024), which also adopt the CLAP model backbone to compute embeddings for FAD calculation. To remain consistent with recent literature, we nonetheless still report VGGish scores for comparison. For brevity, from now on we refer to FAD scores calculated using each of the three models as $\texttt{FAD}_{\texttt{cla}}$, $\texttt{FAD}_{\texttt{clm}}$, and $\texttt{FAD}_{\texttt{vgg}}$, respectively.

We use the FMA Pop dataset (Defferrard et al., 2017) as the reference set for FAD calculation to evaluate music quality, following recommendations for generative music tasks (Gui et al., 2024). The FMA Pop dataset provides a more reliable benchmark for assessing music quality compared to the commonly used MusicCaps dataset (Agostinelli et al., 2023). Next, to specifically evaluate how closely our generated samples align with sleep music characteristics, we compute FAD scores using our curated Spotify Sleep Dataset as a reference set. Throughout this paper, we report FAD scores for both reference sets of FMA Pop and the Spotify Sleep Dataset. To prepare our reference sets for FAD calculation, we download the FMA Pop dataset (Defferrard et al., 2017) from the official repository.

4.2. Subjective Assessments

We conduct a subjective human study following the same procedure as in existing studies (Li et al., 2023; Copet et al., 2023; Kreuk et al., 2023), to evaluate the generated music.

Human participants were asked to rate: **(a)** the overall perceived quality of the generated audio samples (**Qual**), and **(b)** the relevance of the generated audio samples to sleep music (**Rel**), both on a scale of 1 to 100. We leverage the Amazon Mechanical Turk platform to recruit participants and ensure they are paid at least the UK national minimum wage. Noisy annotations and outliers are dropped, such that responses from participants who did not listen to the full audio samples and/or annotators who rated the reference audio samples less than 85 are discarded. All audio samples were also normalized to -20.0 dBFS for fairness.

4.3. Baseline Models

Based on the findings from our literature review, we select AudioLDM (Liu et al., 2023) and MusicGen (Copet et al., 2023) as baseline models for comparison with our proposed model. We choose these models as they have publicly available APIs to generate samples and also offer configurations with small enough model sizes, which is in line with our goal of developing lightweight models for resource-constrained environments. We deploy these respective baseline models using the publicly available implementation on HuggingFace and generate samples by providing the prompt "relaxing sleep music perfect for sleep therapy" as well as other inputs based on the respective author's recommendations. Sleep music generated by our model is filtered to ensure the quality of the presented samples. We compute individual FAD scores (Gui et al., 2024) for these generated samples and select a subset with the top scores for further stages.

5. Evaluation Results

In this section, we first present a comparison between different model configurations followed by our objective and subjective evaluation results.

5.1. Model Configuration Comparison

Our ablation experiments showed that we can improve FAD scores by small architectural tweaks. To further enhance performance, we replace the Griffin-Lim algorithm for mel-spectrogram-to-audio conversion with a neural vocoder. We showcase all the main models and parameters side by side in Table 2.

Each long-sample model, corresponding to 2048×128-pixel mel-spectrograms, was trained for approximately 72 hours, while each short-duration model, using 512×128-pixel spectrograms, was trained for around 36 hours. All models were trained on a single A100 GPU. We selected these training durations to remain compute efficient, and to roughly have the same amount of training steps as other small models in the literature. Latent models performed worse than pure mel-spectrogram diffusion models, with noticeable improvements observed when employing the BigVGAN vocoder. The impact of the vocoder is more pronounced for shorter samples, suggesting that the generated mel-spectrograms are of higher quality. BigVGAN consistently outperforms the Griffin-Lim algorithm across all configurations, particularly when paired with a VAE during the diffusion process. This highlights BigVGAN's ability to mitigate reconstruction artifacts inherent in latent models. The strongest-performing setup combines a mel-spectrogram diffusion process with the BigVGAN vocoder. In terms of efficiency, VAE-based configurations are markedly faster, with generation times reduced from 49.75 to 3.44 seconds for longer samples and from 12.13 to 1.44 seconds for shorter samples. The BigVGAN vocoder enhances sample quality while introducing only minimal computational overhead, further solidifying its advantage over the Griffin-Lim algorithm. Additionally, VAE models achieve faster-than-real-time generation. Across *all* experiments, the DDIM sampler is capable of faster than real-time generation. however, this paper focuses on reporting results using the DDPM sampler. These findings demonstrate the feasibility of utilizing diffusion models for real-time music generation systems. We select the overall best performing model to carry forward into

the human evaluation section, which is a mel-spectrogram based diffusion model using a BigVGAN Vocoder producing samples of ∼12 seconds (*Mel-spec + BigvGAN*).

Configuration	**Sample Length, secs+kHz**	**Sampling Time DDPM/ DDIM (s)**	**Model Size (#Params/MB)**	**FMA Pop Reference Set**		**SSD Reference Set**	
				FAD_{cla}	**FAD_{clm}**	**FAD_{cla}**	**FAD_{clm}**
2048×128 Samples (Longer)							
Mel-spec	23.8 (22kHz)	49.75/4.975	113M/455MB	0.824	0.917	0.446	0.481
Mel-spec (hl512 nfft1024)	47.5 (22kHz)	49.75/4.975	113M/455MB	0.949	0.904	0.697	0.736
Mel-spec + 4× VAE	23.8 (22kHz)	3.44/0.344	113+1.3M/455+5MB	1.154	1.190	0.696	0.719
Mel-spec + BigVGAN	11.9 (44kHz)	50.93/6.155	113+112M/455+451MB	**0.607**	**0.630**	**0.357**	**0.460**
Mel-spec + 4× VAE + BigVGAN	11.9 (44kHz)	4.62/1.524	113+1.3+112M/455+5+451MB	-	-	-	-
512×128 Samples (Shorter)							
Mel-spec	5.9 (22kHz)	12.13/1.213	113M/455MB	0.988	1.012	0.692	0.628
Mel-spec (hl512 nfft1024)	11.9 (22kHz)	12.13/1.213	113M/455MB	0.807	0.879	0.484	**0.506**
Mel-spec + 4× VAE	5.9 (22kHz)	1.44/0.144	113+1.3M/455+5MB	1.119	1.129	0.894	0.760
Mel-spec + BigVGAN	3.0 (44kHz)	11.585/1.343	113+112M/455+451MB	**0.806**	**0.807**	0.613	0.671
Mel-spec + 4× VAE + BigVGAN	3.0 (44kHz)	1.645/0.349	113+1.3+112M/455+5+451MB	0.832	0.821	**0.478**	0.541

Table 2: FAD Score Evaluation on the FMA_Pop Dataset for Models With Different Configurations and Sample Sizes. Smaller Sample Sizes (512×128) and Longer Sample Sizes (2048×128) are indicated in the Table. The Mel-spectrogram + 4× VAE + BigVGAN configuration is omitted as training was unable to converge within the allocated training time.

5.2. Comparison with the Baselines

Model (100 samples)	# Params	Qual ↑	Rel ↑	FMA Pop Reference Set			SSD Reference Set		
				FAD_{cla} ↓	FAD_{clm} ↓	FAD_{vgg} ↓	FAD_{cla} ↓	FAD_{clm} ↓	FAD_{vgg} ↓
Sleep Dataset (human composed)	-	94.72 ± 0.81	92.31 ± 2.23	0.704*	0.827*	11.656*	0.027	0.022	0.336
AudioLDM-S	185M	66.14 ± 5.20	65.07 ± 5.59	**0.642**	0.834	**5.483**	0.864	0.825	9.069
MusicGen-S	300M	83.08 ± 3.34	82.41 ± 3.74	0.851	**0.825**	10.272	0.616	0.693	4.212
Proposed	**115M**	**83.70** ± 3.23	**85.74** ± 3.03	0.823	0.849	11.609	**0.251**	**0.416**	**2.782**

Table 3: Subjective (Qual and Rel) and Objective (FAD) comparison between the baseline models (AudioLDM, MusicGen) and our proposed model. FAD scores are computed using FMA Pop and the Spotify Sleep Dataset as reference as indicated. Lower FAD scores indicate greater similarity to the reference set and are typically preferred. Higher Qual and Rel, from the subjective evaluation surveys, indicate better perceived audio quality and relevance to sleep music and are therefore better.

Both objective and subjective results are reported in Table 3. We first compare our proposed model with the selected baseline models using objective metrics. When using FMA Pop as the reference set, our model achieves FAD scores comparable to MusicGen-S while using only about 1/3 of the number of parameters. On the other hand, when using the Spotify Sleep Dataset as reference, the proposed model considerably out performs the baseline models demonstrating better objective alignment with sleep music characteristics. Next, we present the mean opinion score and 95% confidence intervals from the human evaluation study (subjective results). Again, the proposed model performs similar to MusicGen-S in both audio quality and perceived relevance to sleep music and outperforms AudioLDM-S across the same metrics.

We also observe high FAD values on the Spotify Sleep Dataset (100 samples) when using FMA Pop as the reference set. The FMA Pop set consists of studio recordings of pop songs while the Spotify Sleep Dataset consists of sleep music. Sleep music refers to audio that is typically instrumental and calming. It often features slow tempos, and may incorporate nature sounds such as rain, ocean waves, wind, etc. This difference in type of music between the Sleep Dataset and the FMA Pop set could explain the divergence seen, underscoring the importance of genre-specific reference sets and evaluation metrics when assessing generative models for specialized tasks such as sleep music generation.

Our model has considerably less parameters and required far less training time when compared to the baseline models. We train our proposed model for 2 days on one A100 GPU with a batch size of 8, circa 200k steps, and similarly for our VAE (200k steps on a single A100 GPU). MusicGen-S trains for 1.5 million steps on 32 GPUs, and AudioLDM's VAE alone is trained on 1.5 million steps on a single GPU. Not only do we match performance but also surpass in terms of computational requirements.

On the whole, the proposed model outperforms AudioLDM-S and achieves performance that is comparable to, if not better than, MusicGen-S.

6. Conclusion and Future Work

In this work, we developed lightweight generative models tailored specifically for sleep music. Objective and subjective evaluations show that our models produce high-quality audio, often outperforming other approaches in the literature. We also demonstrate the successful application of the BigVGAN vocoder for music generation, achieving high fidelity. Our experiments examined key design choices for model architectures, balancing efficiency (training and sampling speed) with output quality. The results indicate that our lightweight models generate sleep music with strong resemblance to real examples, supported by low Fréchet Audio Distance (FAD) scores and similarities across acoustic and audio features. Key findings include the following: **(a)** using a curated sleep music dataset enables our models to achieve superior quality and sleep-music relevance as rated by human listeners with significantly fewer parameters compared to existing methods; **(b)** pretrained BigVGAN vocoders, originally designed for speech, are capable of high-quality music generation; **(c)** alternative mel-spectrogram configurations (e.g., non-standard `hop_length`, `n_fft`, and `mel_bands`) outperform conventional literature settings; and **(d)** confirming that VAE compression exhibits diminishing returns, with excessive compression degrading audio quality, as indicated by higher FAD scores. Based on these findings, we select a middle ground of 4× compression for lightweight diffusion model training, balancing efficiency with audio quality. Future research focuses on investigating which musical features best support specific sleep phases and integrating user data to enable adaptive, real-time music generation that corresponds to a user's sleep state. Continuous generation techniques such as successive conditioning or outpainting offer promising directions.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating Music From Text, January 2023. URL http://arxiv.org/abs/2301.11325. arXiv:2301.11325 [cs, eess].

Swapna Bhaskar, D. Hemavathy, and Shankar Prasad. Prevalence of chronic insomnia in adult patients and its correlation with medical comorbidities. *Journal of Family Medicine and Primary Care*, 5(4):780–784, 2016. ISSN 2249-4863. doi: 10.4103/2249-4863.201153.

Madeleine Blank, Jihui Zhang, Femke Lamers, Adrienne D Taylor, Ian B Hickie, and Kathleen R Merikangas. Health correlates of insomnia symptoms and comorbid mental disorders in a nationally representative sample of us adolescents. *Sleep*, 38(2):197–204, 2015.

En-Ting Chang, Hui-Ling Lai, Pin-Wen Chen, Yuan-Mei Hsieh, and Li-Hua Lee. The effects of music on the sleep quality of adults with chronic insomnia using evidence from polysomnographic and self-reported analysis: a randomized control trial. *International*

Journal of Nursing Studies, 49(8):921–930, August 2012. ISSN 1873-491X. doi: 10.1016/j.ijnurstu.2012.02.019.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies, August 2023. URL http://arxiv.org/abs/2308.01546. arXiv:2308.01546.

Consensus Conference Panel, Nathaniel F. Watson, M. Safwan Badr, Gregory Belenky, Donald L. Bliwise, Orfeu M. Buxton, Daniel Buysse, David F. Dinges, James Gangwisch, Michael A. Grandner, Clete Kushida, Raman K. Malhotra, Jennifer L. Martin, Sanjay R. Patel, Stuart F. Quan, and Esra Tasali. Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society on the Recommended Amount of Sleep for a Healthy Adult: Methodology and Discussion. *Sleep*, 38(8):1161–1183, August 2015. ISSN 1550-9109. doi: 10.5665/sleep.4886.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and Controllable Music Generation, June 2023. URL https://arxiv.org/abs/2306.05284v3.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A Dataset For Music Analysis, September 2017. URL http://arxiv.org/abs/1612.01840. arXiv:1612.01840 [cs].

Gaelen Thomas Dickson and Emery Schubert. Music on Prescription to Aid Sleep Quality: A Literature Review. *Frontiers in Psychology*, 11, July 2020. ISSN 1664-1078. doi: 10.3389/fpsyg.2020.01695. Publisher: Frontiers.

Jason Ellis, Luigi Ferini-Strambi, Diego García-Borreguero, Anna Heidbreder, David O'Regan, Liborio Parrino, Hugh Selsick, and Thomas Penzel. Chronic Insomnia Disorder across Europe: Expert Opinion on Challenges and Opportunities to Improve Care. *Healthcare*, 11(5):716, February 2023. ISSN 2227-9032. doi: 10.3390/healthcare11050716.

Zach Evans, Julian D. Parker, C. J. Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Long-form music generation with latent diffusion, April 2024. URL http://arxiv.org/abs/2404.10301. arXiv:2404.10301 [cs, eess].

Marie Anne Gebara, Nalyn Siripong, Elizabeth A. DiNapoli, Rachel D. Maree, Anne Germain, Charles F. Reynolds, John W. Kasckow, Patricia M. Weiss, and Jordan F. Karp. Effect of insomnia treatments on depression: A systematic review and meta-analysis. *Depression and Anxiety*, 35(8):717–731, 2018. ISSN 1520-6394. doi: 10.1002/da.22776.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting Frechet Audio Distance for Generative Music Evaluation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1331–1335, April 2024. doi: 10.1109/ICASSP48485.2024.10446663. ISSN: 2379-190X.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm

Slaney, Ron J. Weiss, and Kevin Wilson. CNN Architectures for Large-Scale Audio Classification, January 2017. URL http://arxiv.org/abs/1609.09430. arXiv:1609.09430 [cs, stat].

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, January 2018. URL http://arxiv.org/abs/1706.08500. arXiv:1706.08500 [cs, stat].

Jing Huang, Inga M. Antonsdottir, Richard Wang, Mengchi Li, and Junxin Li. Insomnia and Its Non-Pharmacological Management in Older Adults. *Current Geriatrics Reports*, September 2023a. ISSN 2196-7865. doi: 10.1007/s13670-023-00397-1.

Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, Jesse Engel, Quoc V. Le, William Chan, Zhifeng Chen, and Wei Han. Noise2Music: Text-conditioned Music Generation with Diffusion Models, March 2023b. URL http://arxiv.org/abs/2302.03917. arXiv:2302.03917 [cs, eess].

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models, January 2023c. URL http://arxiv.org/abs/2301.12661. arXiv:2301.12661 [cs, eess].

Yu-Siang Huang and Yi-Hsuan Yang. Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions, August 2020. URL http://arxiv.org/abs/2002.00212. arXiv:2002.00212.

Kira V. Jespersen, Julian Koenig, Poul Jennum, and Peter Vuust. Music for insomnia in adults. *The Cochrane Database of Systematic Reviews*, 2015(8):CD010459, August 2015. ISSN 1469-493X. doi: 10.1002/14651858.CD010459.pub2.

Junyan Jiang, Gus G. Xia, Dave B. Carlton, Chris N. Anderson, and Ryan H. Miyakawa. Transformer VAE: A Hierarchical Model for Structure-Aware and Interpretable Music Representation Learning. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 516–520, May 2020. doi: 10.1109/ICASSP40776.2020.9054554. ISSN: 2379-190X.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Interspeech 2019*, pages 2350–2354. ISCA, September 2019. doi: 10.21437/Interspeech.2019-2219.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually Guided Audio Generation, March 2023. URL http://arxiv.org/abs/2209.15352. arXiv:2209.15352 [cs, eess].

Gael Le Lan, Bowen Shi, Zhaoheng Ni, Sidd Srinivasan, Anurag Kumar, Brian Ellis, David Kant, Varun Nagaraja, Ernie Chang, Wei-Ning Hsu, Yangyang Shi, and Vikas Chandra. High Fidelity Text-Guided Music Editing via Single-Stage Flow Matching, October 2024. URL http://arxiv.org/abs/2407.03648. arXiv:2407.03648.

Jungup Lee and Bruce A Thyer. Does music therapy improve mental health in adults? a review. *Journal of Human Behavior in the Social Environment*, 23(5):591–603, 2013.

Peike Li, Boyu Chen, Yao Yao, Yikai Wang, Allen Wang, and Alex Wang. JEN-1: Text-Guided Universal Music Generation with Omnidirectional Diffusion Models, August 2023. URL http://arxiv.org/abs/2308.04729. arXiv:2308.04729 [cs, eess].

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models, September 2023. URL http://arxiv.org/abs/2301.12503. arXiv:2301.12503 [cs, eess].

Joanne Loewy. Music Therapy as a Potential Intervention for Sleep Improvement. *Nature and Science of Sleep*, 12:1–9, January 2020. ISSN null. doi: 10.2147/NSS.S194938.

Hila Manor and Tomer Michaeli. Zero-Shot Unsupervised and Text-Based Audio Editing Using DDPM Inversion, May 2024. URL http://arxiv.org/abs/2402.10009. arXiv:2402.10009 [cs].

Goran Medic, Wille , Micheline, , and Michiel EH Hemels. Short- and long-term health consequences of sleep disruption. *Nature and Science of Sleep*, 9:151–161, May 2017. ISSN null. doi: 10.2147/NSS.S134864.

Nur Azmina Mohamad Zamani, Nasiroh Omar, and Nur Damira Huda Azmi. Insomnia Audio Therapy Mobile Application with Music Recommender System. *Mathematical Sciences and Informatics Journal*, 3(1):29–38, May 2022. ISSN 27350703. doi: 10.24191/mij.v3i1.18264. URL https://myjms.mohe.gov.my/index.php/mij/article/view/18264.

Ziqian Ning, Huakang Chen, Yuepeng Jiang, Chunbo Hao, Guobin Ma, Shuai Wang, Jixun Yao, and Lei Xie. DiffRhythm: Blazingly Fast and Embarrassingly Simple End-to-End Full-Length Song Generation with Latent Diffusion, March 2025. URL http://arxiv.org/abs/2503.01183. arXiv:2503.01183 [eess].

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. DITTO: Diffusion Inference-Time T-Optimization for Music Generation, June 2024. URL http://arxiv.org/abs/2401.12179. arXiv:2401.12179 [cs].

Laura Palagini, Elisabeth Hertenstein, Dieter Riemann, and Christoph Nissen. Sleep, insomnia and mental health. *Journal of sleep research*, 31(4):e13628, 2022.

Marco Pasini, Stefan Lattner, and George Fazekas. Music2Latent: Consistency Autoencoders for Latent Audio Compression, August 2024. URL https://arxiv.org/abs/2408.06500v1.

Xingwei Qu, Yuelin Bai, Yinghao Ma, Ziya Zhou, Ka Man Lo, Jiaheng Liu, Ruibin Yuan, Lejun Min, Xueling Liu, Tianyu Zhang, Xinrun Du, Shuyue Guo, Yiming Liang, Yizhi Li, Shangda Wu, Junting Zhou, Tianyu Zheng, Ziyang Ma, Fengze Han, Wei Xue, Gus Xia, Emmanouil Benetos, Xiang Yue, Chenghua Lin, Xu Tan, Stephen W. Huang, Jie Fu, and Ge Zhang. MuPT: A Generative Symbolic Music Pretrained Transformer, November 2024. URL http://arxiv.org/abs/2404.06393. arXiv:2404.06393.

Kannan Ramar, Raman K. Malhotra, Kelly A. Carden, Jennifer L. Martin, Fariha Abbasi-Feinberg, R. Nisha Aurora, Vishesh K. Kapur, Eric J. Olson, Carol L. Rosen, James A. Rowley, Anita V. Shelgikar, and Lynn Marie Trotti. Sleep is essential to health: an American Academy of Sleep Medicine position statement. *Journal of Clinical Sleep Medicine*, 17(10):2115–2119, October 2021. ISSN 1550-9389, 1550-9397. doi: 10.5664/jcsm.9476.

Miriam Richter. Towards the development of music as an intervention for Insomnia treatment: A research synthesis. 2019. doi: 10.24382/1211. URL https://pearl-prod.plymouth.ac.uk/handle/10026.1/14697. Publisher: University of Plymouth.

Dieter Riemann, Fee Benz, Raphael J. Dressle, Colin A. Espie, Anna F. Johann, Tessa F. Blanken, Jeanne Leerssen, Rick Wassing, Alasdair L. Henry, Simon D. Kyle, Kai Spiegelhalder, and Eus J. W. Van Someren. Insomnia disorder: State of the science and challenges for the future. *Journal of Sleep Research*, 31(4):e13604, 2022. ISSN 1365-2869. doi: 10.1111/jsr.13604.

Rebecca Jane Scarratt, Ole Adrian Heggli, Peter Vuust, and Kira Vibe Jespersen. The audio features of sleep music: Universal and subgroup characteristics. *PLOS ONE*, 18(1): e0278813, January 2023. ISSN 1932-6203. doi: 10.1371/journal.pone.0278813. Publisher: Public Library of Science.

Flavio Schneider. ArchiSound: Audio Generation with Diffusion, January 2023. URL http://arxiv.org/abs/2301.13267. arXiv:2301.13267 [cs, eess].

Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Mo\ˆusai: Text-to-Music Generation with Long-Context Latent Diffusion, October 2023. URL http://arxiv.org/abs/2301.11757. arXiv:2301.11757 [cs, eess].

Aurora Tulilaulu, Joonas Paalasmaa, Mikko Waris, and Hannu Toivonen. Sleep Musicalization: Automatic Music Composition from Sleep Measurements. In Jaakko Hollmén, Frank Klawonn, and Allan Tucker, editors, *Advances in Intelligent Data Analysis XI*, pages 392–403, Berlin, Heidelberg, 2012. Springer. ISBN 978-3-642-34156-4. doi: 10.1007/978-3-642-34156-4_36.

Eve van Cauter, Ulf Holmbäck, Kristen Knutson, Rachel Leproult, Annette Miller, Arlet Nedeltcheva, Silvana Pannain, Plamen Penev, Esra Tasali, and Karine Spiegel. Impact of Sleep and Sleep Loss on Neuroendocrine and Metabolic Function. *Hormone Research*, 67 (Suppl. 1):2–9, February 2007. ISSN 0301-0163. doi: 10.1159/000097543.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Chun-Fang Wang, Ying-Li Sun, and Hong-Xin Zang. Music therapy improves sleep quality in acute and chronic sleep disorders: A meta-analysis of 10 randomized studies. *International Journal of Nursing Studies*, 51(1):51–62, January 2014. ISSN 0020-7489. doi: 10.1016/j.ijnurstu.2013.03.008.

Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kai-wei Chang, Ho-Lam Chung, Alexander H. Liu, and Hung-yi Lee. Towards audio language modeling – an overview, February 2024a. URL http://arxiv.org/abs/2402.13236. arXiv:2402.13236 version: 1.

Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kaiwei Chang, Jiawei Du, Ke-Han Lu, Alexander H. Liu, Ho-Lam Chung, Yuan-Kuei Wu, Dongchao Yang, Songxiang Liu, Yi-Chiao Wu, Xu Tan, James Glass, Shinji Watanabe, and Hung-yi Lee. Codec-SUPERB @ SLT 2024: A lightweight benchmark for neural audio codec models, September 2024b. URL http://arxiv.org/abs/2409.14085. arXiv:2409.14085 version: 1.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation, March 2024c. URL http://arxiv.org/abs/2211.06687. arXiv:2211.06687 [cs, eess].

Ami Yamasato, Mayu Kondo, Shunya Hoshino, Jun Kikuchi, Mayumi Ikeuchi, Kiyoyuki Yamazaki, Shigeki Okino, and Kenji Yamamoto. How Prescribed Music and Preferred Music Influence Sleep Quality in University Students. *The Tokai Journal of Experimental and Clinical Medicine*, 45(4):207–213, December 2020. ISSN 2185-2243.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete Diffusion Model for Text-to-sound Generation, April 2023. URL http://arxiv.org/abs/2207.09983. arXiv:2207.09983 [cs, eess].

Jing Yang, Chulhong Min, Akhil Mathur, and Fahim Kawsar. SleepGAN: Towards Personalized Sleep Therapy Music. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 966–970, May 2022. doi: 10.1109/ICASSP43922.2022.9747033. ISSN: 2379-190X.

Ning Zhang. Learning Adversarial Transformer for Symbolic Music Generation. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4):1754–1763, April 2023. ISSN 2162-2388. doi: 10.1109/TNNLS.2020.2990746. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.