SleepLong: Towards Generating Long-Sequence Sleep Heart Rate Signals with Conditional Diffusion

Manasa Mariam Mammen* **Priyanka Mary Mammen†** **Emil Joswin‡**

Abstract

Sleep-stage classification is a critical step in assessing sleep quality. Wearable sleep trackers offer a promising solution for long-term monitoring outside traditional clinical settings. Most wearable sleep trackers are heart-rate-based, but their effectiveness is limited by shortage of good-quality publicly available data. To address this, diffusion models offer a privacy-aware approach to generate data for augmentation and to train classification models. Existing generation methods typically focus on individual sleep stages in isolation, without modeling the dependencies and continuity across stages. This paper explores a spectrogram-based diffusion model to generate a long range sleep heart-rate sequence conditioned on sleep-stage labels (hypnogram), as opposed to generating the individual stages in isolation. We verify the effectiveness of the approach in sleep-stage classification tasks using two publicly available datasets, HMC and DREAMT.

1 Introduction

Sleep is a fundamental physiological process vital for physical and mental health [1, 2]. The current gold standard for assessing sleep quality is polysomnography (PSG), which is a comprehensive overnight study conducted in a clinical setting [3, 4]. While highly accurate, PSG is expensive, labor-intensive, and not easily scalable [5]. Wearable devices have emerged as a promising alternative, offering a non-invasive, low-cost, and at-home solution for long-term sleep monitoring [6, 7]. A good majority of the wearable sleep trackers use Instantaneous Heart Rate (IHR), which has been shown to reflect sleep-stage related changes in autonomic activity, making it a valuable proxy for sleep quality assessment [8, 9, 10]. The main challenge in the widespread adoption of wearable trackers is the development of robust and generalizable machine learning models which is hindered by limited data availability due to privacy concerns, and the difficulty of collecting large-scale labeled datasets.

Generative modeling provides a way to mitigate this limitation by synthesizing physiologically plausible data to supplement real-world datasets. Among various generative modeling approaches, diffusion probabilistic models have recently demonstrated state-of-the-art performance in generating high-fidelity data across domains like images, audio, and biosignals [11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. Their iterative denoising process offers stability in training, fine-grained control over conditioning, and the ability to capture complex temporal dependencies.

Currently, diffusion models in the sleep-stage data augmentation domain generate short signal segments conditioned on individual sleep stages [21, 22]. A key limitation of this stage-wise approach is that concatenating these isolated segments into a long sequence often produces unrealistic transitions between stages. As a result, the generated signals lack temporal consistency, reducing their usefulness for downstream tasks that rely on continuous, physiologically plausible dynamics.

*Technical University of Munich, manasamariam.mammen@tum.de

†University of Massachusetts Amherst, pmammen@cs.umass.edu

‡IEEE, emiljoswin@ieee.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time Series for Health.

Our work addresses this limitation by using a diffusion model to directly synthesize long-range IHR signals conditioned on a full sleep period hypnogram. This method allows the model to learn and reproduce the natural, long-range temporal patterns and stage-transition dynamics that occur during sleep, ensuring both physiological realism and continuity in the generated data.

The main contributions of this work are:

- A novel classifier-free guided diffusion model for generating realistic, long-term IHR signals conditioned on a sleep stage labels, uniquely preserving both stage continuity and physiological plausibility.
- Demonstrate the efficacy of the data augmentation approach via sleep-stage classification using publicly available datasets such as HMC and DREAMT. We show that the inclusion of this synthetic data significantly improves the performance of a downstream classifier, validating the utility and quality of our generated signals.

2 Label Conditioned Heart-rate Generation

Problem Statement: Sleep is composed of five distinct stages: Wake, rapid eye movement (REM), and three non-REM (NREM) stages (N1, N2, N3), each characterized by variations in physiological signals [23]. The progression of these stages is represented by a hypnogram. Given a hypnogram as input, our goal is to generate a realistic heart-rate signal that captures the dynamics of sleep physiology across an entire night.

2.1 Diffusion Model Pipeline

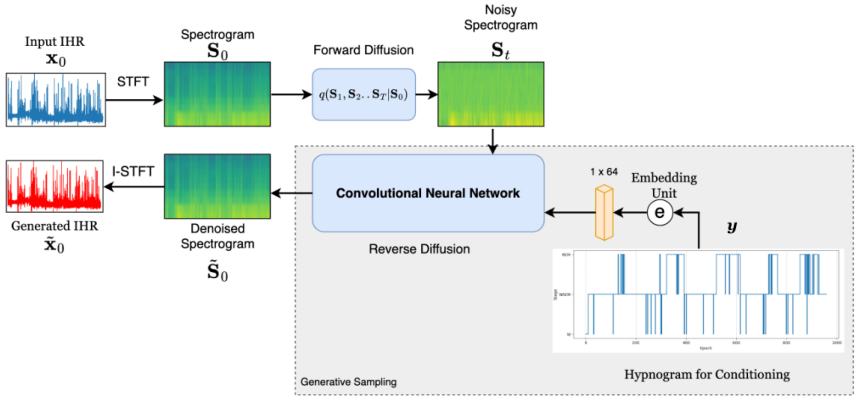


Figure 1: Overview of our method.

The model used in this work is a conditional denoising diffusion probabilistic model (DDPM) with classifier-free guidance (CFG), designed to generate heart rate signals conditioned on sequences of sleep stages. Each heart rate sequence in the dataset is represented as $\boldsymbol{x}_0^{(i)}$, paired with its corresponding sleep-stage annotations $\boldsymbol{y}^{(i)}$, where $i = 1, \ldots, N$. For notational simplicity, we will omit superscript indices and refer to a generic heart rate sequence as $\boldsymbol{x}_0$ and its sleep-stage sequence as $\boldsymbol{y}$. Each heart rate sequence $\boldsymbol{x}_0$ is first transformed into a two-dimensional time-frequency spectrogram $\boldsymbol{S}_0$ using Short-Time Fourier Transform (STFT). This representation captures both temporal and spectral characteristics of the signal, making stage-specific physiological patterns more separable. The diffusion process is then applied in the spectrogram domain, and after generation, the synthetic spectrograms are converted back into the time domain to reconstruct heart rate traces.

In the forward diffusion process, Gaussian noise is incrementally added to the spectrogram $\boldsymbol{S}_0$ over a fixed number of timesteps T according to a linear variance schedule. At each step t, the variance parameter is denoted as β_t, with $\alpha_t = 1 - \beta_t$ representing the fraction of the signal that is retained at step t and the cumulative product $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ measuring how much of the original signal remains

after t steps. A noisy spectrogram at step t can then be written in closed form as

$$\boldsymbol{S}_t = \sqrt{\bar{\alpha}_t}\,\boldsymbol{S}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is Gaussian noise. The number of timesteps T is a fixed hyperparameter that determines how gradually noise is injected and how many denoising steps are required for reconstruction. In this work, we set $T = 1000$, following the original DDPM framework, as this provides stable training and high-quality spectrogram generation [24].

The reverse process inverts the forward noising procedure by denoising $\boldsymbol{S}_t$ back to $\boldsymbol{S}_0$, conditioned on the sleep-stage sequence $\boldsymbol{y}$. The label sequence $\boldsymbol{y}$ is mapped into a learned embedding c using a linear label encoder, and the denoiser $\boldsymbol{\epsilon}_\theta(\boldsymbol{S}_t, t, c)$, implemented as a Convolutional Neural Network (CNN), predicts the noise component present in $\boldsymbol{S}_t$. To enable classifier-free guidance, conditioning is randomly dropped during training with probability p_{uncond}, replacing c with a null embedding $c_\emptyset$. The training objective minimizes the mean squared error between the true Gaussian noise $\boldsymbol{\epsilon}$ and the network prediction:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{S}_0, \boldsymbol{y}, t, \boldsymbol{\epsilon}}\Big[\, \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{S}_t, t, c)\|^2 \,\Big]. \tag{2}$$

During generation, the model starts from Gaussian noise $\boldsymbol{S}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and applies the reverse diffusion process. The classifier-free guidance combines the conditional and unconditional predictions to form a guided noise estimate [25]:

$$\tilde{\boldsymbol{\epsilon}}_\theta = (1+w)\,\boldsymbol{\epsilon}_\theta(\boldsymbol{S}_t, t, c) - w\,\boldsymbol{\epsilon}_\theta(\boldsymbol{S}_t, t, c_\emptyset), \tag{3}$$

where w is the guidance scale. This guided estimate $\tilde{\boldsymbol{\epsilon}}_\theta$ is then used to reconstruct a clean spectrogram $\tilde{\boldsymbol{S}}_0$, which is finally converted back into the time domain via the inverse STFT to yield a synthetic heart rate sequence $\tilde{\boldsymbol{x}}_0$. An overview of the proposed method is shown in Figure 1.

3 Experiments

3.1 Dataset and preprocessing

We utilized two publicly available datasets - HMC [26] and DREAMT [27] for our experiments. HMC consists of 151 whole-night polysomnographic sleep recordings. DREAMT is a collection of 100 whole-night sleep recordings of actigraphy data with technician-annotated labels from PSG data. Most of the DREAMT participants are diagnosed with sleep disorders. Instantaneous Heart Rate (IHR) from the ECG (HMC) and PPG (DREAMT) signals are extracted using Pan-Tompkins algorithm [28]. IHR is then resampled to 2Hz and padded to 57600 samples (8 hours) to ensure data uniformity. The sleep labels are annotated for every 30 second of the signal for both datasets.

3.2 Model and training details

All experiments were conducted on NVIDIA Tesla T4. The architecture of the denoising network is described in Appendix Table 3. We use $T = 1000$ diffusion steps with a linear variance schedule where β_t increases from 10^{-4} to 0.02. Classifier-free guidance is applied with a guidance scale of $w = 5.0$. Optimization is performed using Adam with a learning rate of 10^{-3} and a batch size of 2. Training runs for up to 1000 epochs with early stopping ($patience = 50$, $delta = 10^{-4}$).

4 Results and Discussions

Diffusion model training is validated using mean squared error (MSE) between predicted and true noise on the test dataset, reflecting how well the network learns the denoising objective. For further assessment, predicted noise is used to denoise the spectrograms, and the resulting signals are compared with the original IHR to verify preservation of temporal dynamics. To complement these metrics, the appendix provides qualitative illustrations of selected denoising results, including magnified regions for closer examination.

As shown in Table 1, predicted noise errors remain low for both datasets, confirming that the model learns the denoising objective effectively. In contrast, denoising errors are considerably higher for the DREAMT dataset than for HMC, reflecting the greater variability and scale differences in signals collected from individuals with sleep disorders.

Table 1: Mean squared error (MSE) for noise prediction and heart rate denoising.

Dataset	Noise Prediction MSE ↓	Denoised Heart Rate MSE ↓
HMC	2.10e-1	1.53e-2
DREAMT	6.05e-2	1.15e+1

4.1 Spectral Analysis of Generated Heart Rate

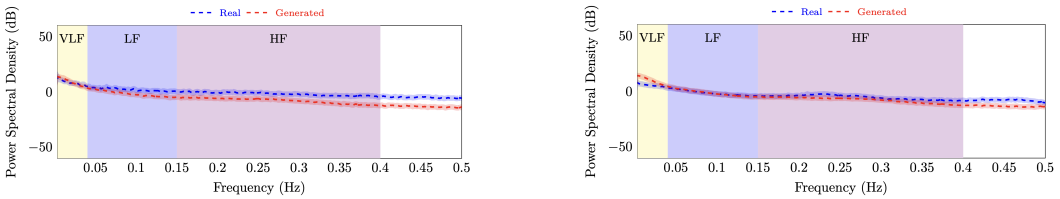


Figure 2: PSD Comparison for Wake and NREM in HMC dataset.

We analyze power spectral density (PSD) across sleep stages to assess realism, as it can show how much of the heart rate signal comes from slow or fast fluctuations. With a 2 Hz sampling rate (Nyquist = 1 Hz), we focus on three bands: very low frequency (VLF, 0–0.04 Hz), low frequency (LF, 0.04–0.15 Hz), and high frequency (HF, 0.15–0.4 Hz). As shown in Figure 2, the generated signals reproduce the main stage-specific patterns, with slightly higher VLF in NREM and moderate deviations in LF and HF bands in Wake. But overall, they still preserve the expected spectral organization, showing that the model captures key features of heart rate variability. Results for REM are provided in the appendix.

4.2 Downstream Task Evaluation

Table 2: Downstream task evaluation.

Dataset	Original			With Augmentation (Full Sequence)			With Augmentation (Sequence Stitching)		
	Acc	Macro F1	κ	Acc	Macro F1	κ	Acc	Macro F1	κ
HMC	0.68	0.48	0.36	<u>0.72</u>	<u>0.53</u>	<u>0.38</u>	0.69	0.52	0.35
DREAMT	0.76	<u>0.55</u>	0.58	<u>0.79</u>	<u>0.55</u>	<u>0.60</u>	0.78	0.54	0.59

We evaluate the utility of generated data via downstream sleep stage classification using a model similar to [29] and evaluate its performance using macro F1 score, accuracy and kappa score (κ). Table 2 compares three setups: i) real data only, ii) augmentation with our full sequence diffusion model (Model A), and iii) augmentation with the sequence stitching baseline (Model B). Results are averaged over 5 runs; standard deviations were consistently below 0.005 and are omitted for clarity. On the HMC dataset, adding 20% synthetic data from Model A improved accuracy (+5.9%), macro F1 (+10.4%), and κ (+5.6%), while Model B gave smaller or inconsistent gains. On the DREAMT dataset with 10% synthetic data, Model A improved accuracy (+3.9%) and κ (+3.4%), with no change in F1, whereas Model B gave weaker improvements. These results show that even modest synthetic augmentation enhances performance, and that the full sequence model is more effective than sequence stitching by leveraging cross stage temporal context.

5 Conclusions

We introduced a spectrogram based diffusion framework for generating heart rate signals conditioned on sleep stages. The model efficiently produces long synthetic recordings that preserve spectral patterns and is validated through downstream sleep stage classification, where synthetic data improved performance. By modeling cross stage temporal context, it outperforms a stage specific baseline. In future, we plan to improve our model further by adding a hypnogram generator as well so that we don't rely on fixed labels.

References

[1] G. S. Perry, S. P. Patil, and L. R. Presley-Cantrell, "Raising awareness of sleep as a healthy behavior," *Preventing chronic disease*, vol. 10, p. E133, 2013.

[2] M. R. Rosekind, K. B. Gregory, M. M. Mallis, S. L. Brandt, B. Seal, and D. Lerner, "The cost of poor sleep: workplace productivity loss and associated costs," *Journal of occupational and environmental medicine*, vol. 52, no. 1, pp. 91–98, 2010.

[3] J. V. Rundo and R. Downey III, "Polysomnography," *Handbook of clinical neurology*, vol. 160, pp. 381–392, 2019.

[4] R. F. Kaplan, Y. Wang, K. A. Loparo, M. R. Kelly, and R. R. Bootzin, "Performance evaluation of an automated single-channel sleep–wake detection algorithm," *Nature and science of sleep*, pp. 113–122, 2014.

[5] J. Bennett and W. Kinnear, "Sleep on the cheap: the role of overnight oximetry in the diagnosis of sleep apnoea hypopnoea syndrome," 1999.

[6] A. Inc., "Apple watch." `https://www.apple.com/apple-watch/`, 2022.

[7] F. LLC, "Fitbit devices." `https://www.fitbit.com/`, 2020.

[8] W. Baust and R. Engel, "The correlation of heart and respiratory frequency in natural sleep of man and their relation to dream content.," *Electroencephalography and clinical neurophysiology*, vol. 30, no. 3, pp. 262–263, 1971.

[9] J. L. Aldredge and A. J. Welch, "Variations of heart rate during sleep as a function of the sleep cycle," *Electroencephalography and Clinical Neurophysiology*, vol. 35, no. 2, pp. 193–198, 1973.

[10] M. Bonnet and D. Arand, "Heart rate variability: sleep stage, time of night, and arousal influences," *Electroencephalography and clinical neurophysiology*, vol. 102, no. 5, pp. 390–396, 1997.

[11] S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. J. Fleet, "Synthetic data from diffusion models improves imagenet classification," *arXiv preprint arXiv:2304.08466*, 2023.

[12] W. H. L. Pinaya, P.-D. Tudosiu, R. Gray, G. Rees, P. Nachev, S. Ourselin, and M. J. Cardoso, "Unsupervised brain anomaly detection and segmentation with transformers," *arXiv preprint arXiv:2102.11650*, 2021.

[13] W. H. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, "Brain imaging generation with latent diffusion models," in *MICCAI workshop on deep generative models*, pp. 117–126, Springer, 2022.

[14] W. H. Pinaya, M. S. Graham, R. Gray, P. F. Da Costa, P.-D. Tudosiu, P. Wright, Y. H. Mah, A. D. MacKinnon, J. T. Teo, R. Jager, *et al.*, "Fast unsupervised brain anomaly detection and segmentation with diffusion models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 705–714, Springer, 2022.

[15] V. Fernandez, W. H. L. Pinaya, P. Borges, P.-D. Tudosiu, M. S. Graham, T. Vercauteren, and M. J. Cardoso, "Can segmentation models be trained with fully synthetically generated data?," in *International workshop on simulation and synthesis in medical imaging*, pp. 79–90, Springer, 2022.

[16] V. Fernandez, P. Sanchez, W. H. L. Pinaya, G. Jacenków, S. A. Tsaftaris, and M. J. Cardoso, "Privacy distillation: reducing re-identification risk of diffusion models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 3–13, Springer, 2023.

[17] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[19] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[20] E. Adib, A. S. Fernandez, F. Afghah, and J. J. Prevost, "Synthetic ecg signal generation using probabilistic diffusion models," *IEEe Access*, vol. 11, pp. 75818–75828, 2023.

[21] B. Aristimunha, R. Y. de Camargo, S. Chevallier, A. G. Thomas, O. Lucena, J. Cardoso, W. H. L. Pinaya, and J. Dafflon, "Synthetic sleep eeg signal generation using latent diffusion models," in *DGM4H 2023-1st Workshop on Deep Generative Models for Health at NeurIPS 2023*, 2023.

[22] Y. Yin, "Sleep-cbddm: Sleep data augmentation method based on conditional bilateral denoising diffusion model," in *Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science*, pp. 745–749, 2023.

[23] R. B. Berry, R. Brooks, C. Gamaldo, S. M. Harding, R. M. Lloyd, S. F. Quan, M. T. Troester, and B. V. Vaughn, "Aasm scoring manual updates for 2017 (version 2.4)," 2017.

[24] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[25] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[26] D. Alvarez-Estevez and R. M. Rijsman, "Inter-database validation of a deep learning approach for automatic sleep scoring," *PloS one*, vol. 16, no. 8, p. e0256111, 2021.

[27] W. K. Wang, J. Yang, L. Hershkovich, H. Jeong, B. Chen, K. Singh, A. R. Roghanizad, M. M. H. Shandhi, A. R. Spector, and J. Dunn, "Addressing wearable sleep tracking inequity: a new dataset and novel methods for a population with sleep disorders," in *Proceedings of the fifth Conference on Health, Inference, and Learning (Proceedings of Machine Learning Research, Vol. 248), Tom Pollard, Edward Choi, Pankhuri Singhal, Michael Hughes, Elena Sizikova, Bobak Mortazavi, Irene Chen, Fei Wang, Tasmie Sarker, Matthew McDermott, and Marzyeh Ghassemi (Eds.). PMLR*, pp. 380–396, 2024.

[28] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE transactions on biomedical engineering*, no. 3, pp. 230–236, 2007.

[29] N. Sridhar, A. Shoeb, P. Stephens, A. Kharbouch, D. B. Shimol, J. Burkart, A. Ghoreyshi, and L. Myers, "Deep learning for automated sleep staging using instantaneous heart rate," *NPJ digital medicine*, vol. 3, no. 1, p. 106, 2020.

A Technical Appendices and Supplementary Material

A.1 Model Architecture

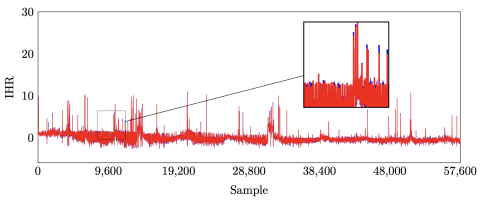
Table 3: Architecture of the convolutional neural network.

Layer	Input channels	Output channels	Kernel size / Padding
Label projection (Linear)	64	1	–
Concatenation with x_t and timestep map	–	3	–
Conv2D + ReLU	3	64	3×3 / 1
Conv2D + ReLU	64	64	3×3 / 1
Conv2D (output)	64	1	3×3 / 1

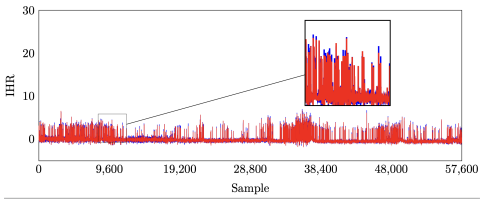
Table 3 outlines the architecture of the proposed denoising network. The model first projects the conditioning label into a embedding space, which is concatenated with the noisy spectrogram input S_t and a sinusoidal timestep encoding, forming a three-channel input. This is processed by a stack

of lightweight convolutional layers with ReLU activations to capture local temporal and spectral correlations. The final convolutional layer reduces the representation back to a single channel, yielding the denoised spectrogram estimate.

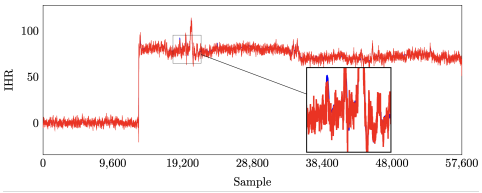
A.2 Supplementary Results: Qualitative Evaluation on Test Dataset



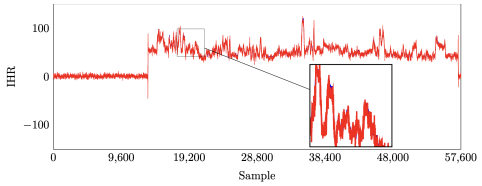
Sample 1 (HMC)



Sample 2 (HMC)



Sample 3 (DREAMT)

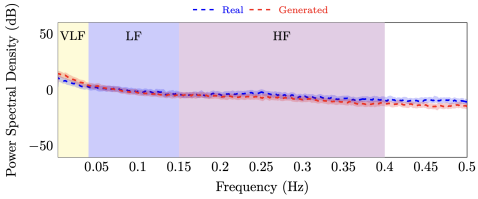


Sample 4 (DREAMT)

Figure 3: Comparison of original (blue) and denoised (red) IHR Signals.

To complement the quantitative evaluation in Table 1, Figure 3 shows representative examples of original and denoised IHR signals from the test dataset. The examples were randomly selected to represent typical cases. The visual comparison confirms that the denoised IHR signals follow the ground-truth waveforms closely, with only minor deviations. While reconstruction errors for the DREAMT dataset were slightly higher than for HMC, the overall temporal patterns and physiologically relevant fluctuations remain well preserved, indicating that the model generalizes effectively to more variable test conditions.

A.3 Supplementary Results: PSD Comparison (Cont..)



REM

Figure 4: PSD comparison for REM in HMC.

The PSD comparison of the REM stage in Figure 4 for the HMC dataset shows a very close match between original and generated signals, with a slightly elevated VLF component visible

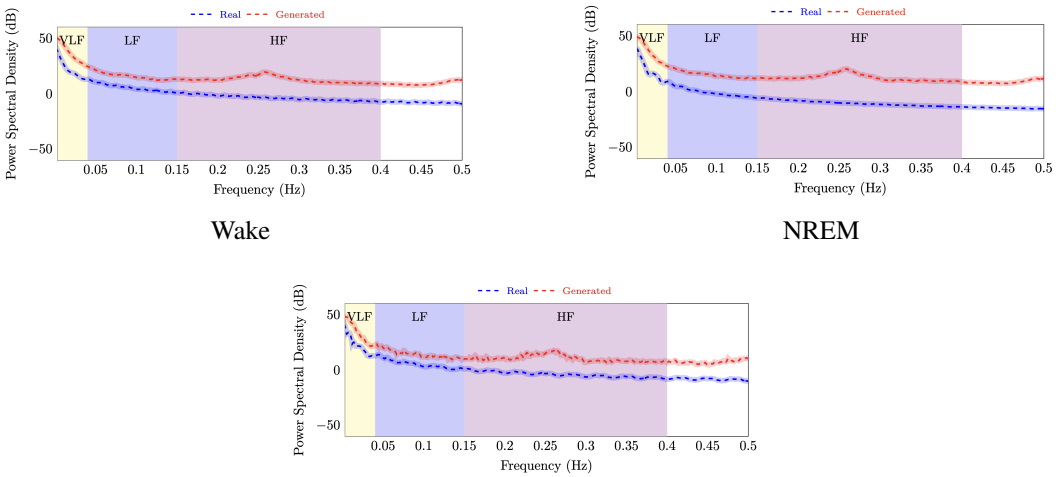


Figure 5: PSD Comparison for Wake, NREM and REM in DREAMT.

Across all sleep stages in Figure 5, the generated signals exhibit a spike in the high-frequency (HF) band, indicating the presence of residual noise rather than physiological patterns. Despite this deviation, downstream task performance remained unaffected, showing that the generated signals still capture the task-relevant information.