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

Anonymous Author(s)
Affiliation
Address
email

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. 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

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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

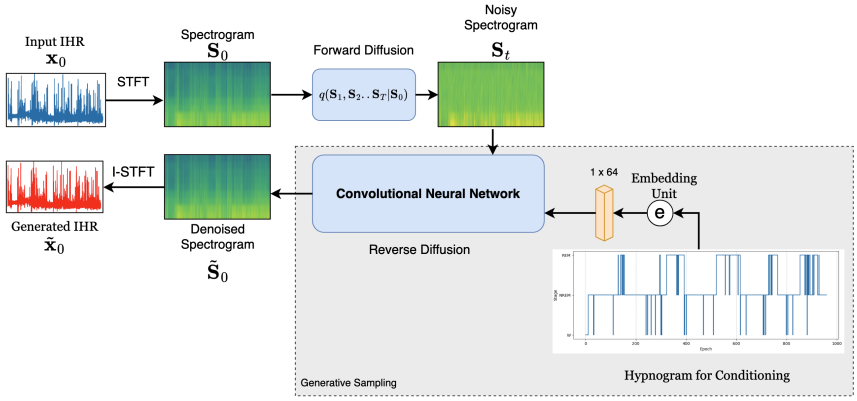


Figure 1: Overview of our method.

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 ($patience = 50$, $delta = 10^{-4}$).

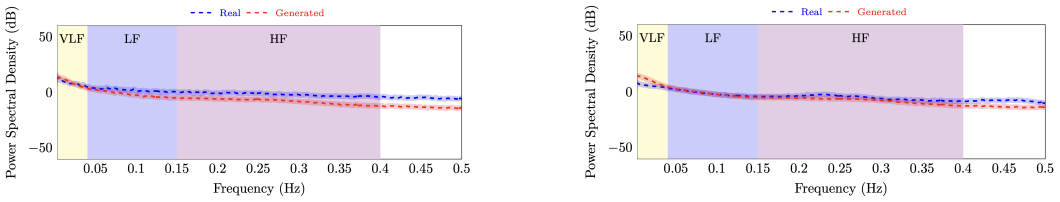
4 Results and Discussions

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



Wake NREM

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

4.1 Spectral Analysis of Generated Heart Rate

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	0.72	0.53	0.38	0.69	0.52	0.35
DREAMT	0.76	0.55	0.58	0.79	0.55	0.60	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). 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

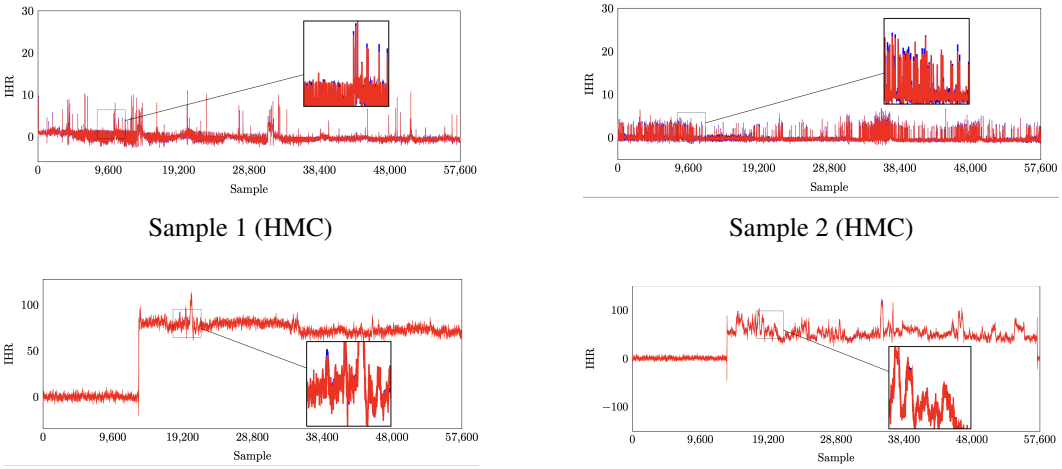
Table 3: Architecture of the convolutional neural network.

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



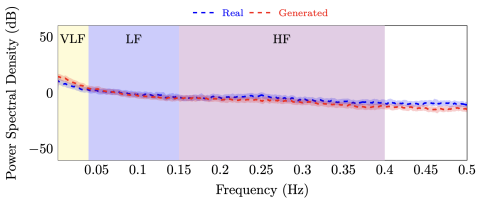
Sample 1 (HMC) Sample 2 (HMC)

Sample 3 (DREAMT) Sample 4 (DREAMT)

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

The visualizations in Figure 3 support the interpretation from Table 1, showing that although numerical errors increase, the reconstructed signals remain closely aligned with the originals, demonstrating preservation of physiologically relevant dynamics despite dataset differences.

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 match232 between original and generated signals, with a slightly elevated VLF component visible

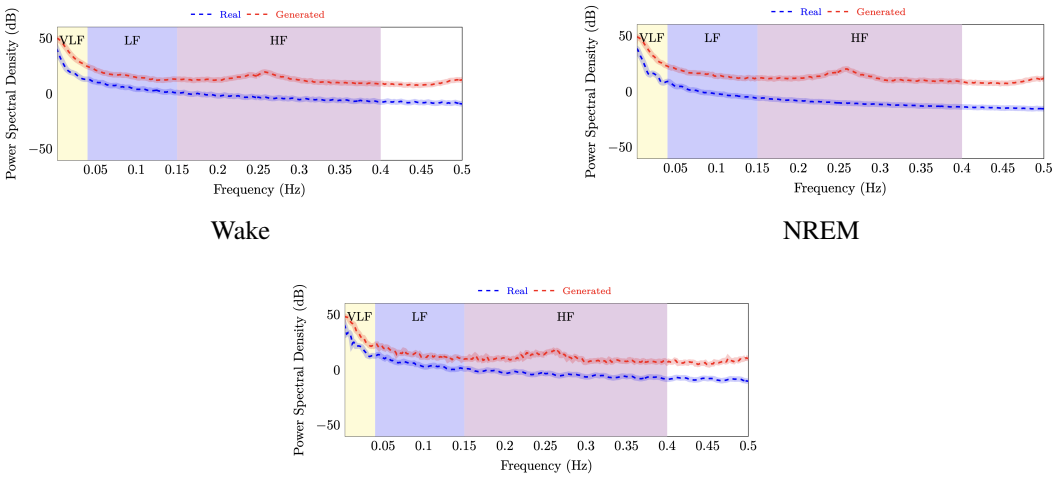


Figure 5: PSD Comparison for Wake, NREM and REM in DREAMT.

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.

NeurIPS Paper Checklist

limit.

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer:[Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use publicly available datasets. Code will be made available upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Right now we report only the mean value from the experiments because of time constraints. We will add error bar charts for camera ready version.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.