# High-Fidelity Synthetic ECG Generation via Mel-Spectrogram Informed Diffusion Training

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

The development of machine learning for cardiac care is severely hampered by privacy restrictions on sharing real patient electrocardiogram (ECG) data. While generative AI offers a promising solution, synthesized ECGs produced by existing models often lack the morphological fidelity required for clinical utility due to their reliance on simplistic and general training objectives such as MSE loss. In this work, we address this critical gap by introducing **MIST-ECG (Mel spectrogram Informed Synthetic Training)**, a novel training paradigm that supervises the conditional diffusion-based Structured State Space Model (SSSD-ECG) with time–frequency domain objective to enforce structural realism. We train and rigorously evaluate our framework on the PTB-XL dataset, assessing the synthesized ECG signals on trustworthiness, fidelity, privacy preservation, and downstream task utility. MIST-ECG achieves substantial gains: it improves morphological coherence, preserves strong privacy guarantees with all metrics evaluated exceeding the baseline by 4%-8%, and notably reduces the interlead correlation error by an average of 74%. In critical low-data regimes, a classifier trained on datasets supplemented with our synthetic ECGs achieves performance comparable to a classifier trained solely on real data. This work demonstrates that ECG synthesizers, trained with the proposed time–frequency structural regularization scheme, can serve as high-fidelity, privacy-preserving surrogates when real data are scarce.

## 1 Introduction

Cardiovascular disease remains the leading cause of death worldwide, creating a staggering health and economic burden [9]. The electrocardiogram (ECG) is the cornerstone of cardiac diagnostics, and applying machine learning to these signals promises earlier and more accurate diagnoses [7]. However, this promise is constrained by a fundamental data access bottleneck. ECGs are not merely medical records; they are sensitive biometric data that reveal extensive personal health information [11]. Consequently, privacy regulations limit the sharing of large, diverse datasets needed to train robust and generalizable AI models. High-fidelity synthetic data generation has emerged as the most promising solution, offering a pathway to democratize research and accelerate innovation [2, 1].

As the field moves from feasibility to clinical application, it faces a critical challenge: morphological fidelity. The clinical utility of an ECG depends on the precise shape, duration, and interplay of its components (P-waves, QRS complexes, T-waves) and spatio-temporal coherence across all 12 leads. Signals that are statistically similar but morphologically flawed are not ready for clinical use. They risk biased algorithms, failed validation studies, and a loss of trust in AI-driven diagnostics. Current generative models [1] perpetuate this gap, relying on generic time-series losses such as Mean Squared Error. These metrics are structurally agnostic, treating each time point independently, and fail to

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

impose a physiologically grounded prior. As a result, signals that appear plausible may lack the subtle morphological integrity and inter-lead correlations essential for safe clinical use.

This paper argues that closing the morphological fidelity gap is essential for deploying generative ECG models responsibly. We introduce **MIST-ECG (Mel-Spectrogram Informed Synthetic Training for ECGs)**, a training paradigm designed to address this problem. By supervising the generative process in the time-frequency domain with the mel-spectrogram difference between real and synthetic 12-lead ECGs, MIST-ECG imposes a clinically relevant structural prior on generated waveforms. Our goal is not incremental improvements in point-wise error but a fundamental enhancement of physiological plausibility. We demonstrate the efficacy of our approach through a comprehensive evaluation, showing a 4%-8% improvements of all fidelity and morphological Metrics, an average 74% reduction in inter-lead correlation error and, critically, that downstream 71 disease label ECG classifier model trained on synthetic supplemented data can match and sometimes outperform solely real training data in diagnostic performance under low-data regimes. This establishes a scalable, privacy-conscious foundation for advancing cardiac research with synthetic ECGs.

## 2 Related Work

**Advances in Synthetic ECG Generation.** The synthesis of ECG signals has evolved rapidly from early approaches with Generative Adversarial Networks (GANs) [2, 14] and Variational Autoencoders (VAEs) [10, 5] to the current state-of-the-art: diffusion models [4]. Architectures like SSSD-ECG [1] have demonstrated superior sample quality and training stability. However, the success of these models has been primarily driven by architectural innovations, while the training objective has remained relatively simple, typically relying on point-wise losses that do not explicitly enforce morphological realism. **Ensuring Fidelity in Medical Time Series Generation.** A common challenge across generative modeling for private medical data, from electroencephalography (EEG) signals [13] to electronic health records (EHR) [3], is ensuring the structural integrity of the generated time series. The predominant training paradigm relies on time-domain losses like Mean Squared Error, which are often insufficient to enforce the morphological coherence essential for clinical realism. Concurrent work, such as ECG-DPM [6], has also recently explored using spectrogram-based diffusion models, based on UNet backbone and is not conditional. Building upon this principle, our work introduces **MIST-ECG**, a framework that not only systematically applies a mel-spectrogram-informed training paradigm but also provides the first rigorous evaluation of its impact on physiological coherence and its ability to serve as a surrogate for real data in data-scarce settings. This bridges a methodological gap by imposing a stronger, clinically relevant structural prior on the generated waveforms, addressing the morphological fidelity and inter-lead correlation gap left by previous methods.

## 3 Methods

Our methodology introduces a novel paradigm, **MIST-ECG**, leverages the 12-lead Mel Spectrogram representation of ECGs to supervise the generative model training. We selected SSSD-ECG [1] as our foundational architecture due to its proven success in generating high-fidelity 12-lead ECGs. The model leverages a score-based diffusion process to transform random noise into structured signals iteratively. Its core strength lies in its use of Structured State-Space Model (SSSM) layers, which are highly effective at capturing the long-range temporal dependencies crucial for modeling the physiological structure of an entire heartbeat and rhythm.

**SSSD-ECG: Diffusion-based Conditional ECG Generation with Structured State Space Models** In its original implementation, SSSD-ECG conditioned on a 71 length onehot vector representing diagnostic labels, which is projected into a continuous representation via a learnable weight matrix. Despite its strong performance, this framework has two primary limitations. First, its reliance on a mean squared error (MSE) loss treats each time step independently, failing to impose a global structural prior on the waveform's morphology. Second, its conditioning is limited to a monolithic vector of disease labels, which prevents the generation of personalized ECGs. Although both are important areas for improvement, our work focuses on resolving the morphological fidelity gap.

**MIST-ECG: Mel-Spectrogram Informed Synthetic Training for ECGs.** To address the limitation of standard diffusion training for physiological time-series generation—its reliance on point-wise losses like MSE that ignore spatiotemporal structure—we propose a paradigm that adds spatiotempo-

ral supervision to better capture ECG waveform morphology and duration. Our approach is inspired by Mel Spectrogram used as a higher fidelity of continuous representation other than vector quantization in audio synthesis [8] but is specifically adapted to the unique physiological characteristics of the ECG. The process begins by computing a multi-resolution Short-Time Fourier Transform (STFT) of the ECG signals, using window sizes $wl \in \{256, 512\}$ to capture both high-frequency, transient events (like the sharp QRS complex) and low-frequency, evolving waveforms (like the T-wave). We then warp the frequency axis of these spectrograms onto the perceptually-motivated Mel scale, using 64 and 128 Mel bands for each resolution. The Mel scale's non-linear compression of frequencies is uniquely suited for ECG analysis, as it naturally places greater emphasis on the diagnostically-rich low-frequency bands where information about ST segments and T-wave morphology resides. This imposes a strong inductive bias, forcing the model to prioritize the most clinically relevant spectral components. The MIST loss is defined as the summed $L_1$ distance between the multi-resolution mel-spectrograms of the generated ($\hat{y}$) and ground-truth ($y$) signals:

$$\mathcal{L}_{\text{MIST}}(\hat{y}, y) = \sum_{wl} \sum_{bands} \left\| M_{wl,bands}(\hat{y}) - M_{wl,bands}(y) \right\|_1 \tag{1}$$

The final training objective is a weighted sum: $\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{MSE}} + \beta \mathcal{L}_{\text{MIST}}$. We empirically selected $\beta = 0.02$, which stabilizing training and leading to consistent loss reduction. The mechanism of MIST-ECG and the visualization of 12-lead Mel-spectrogram of real data and MIST-ECG can be found in Appendix B 5.

# 4 Experiments and Results

To rigorously validate our proposed methods, we designed a multi-faceted evaluation framework on the public PTB-XL dataset [12]. Our investigation is structured as a direct comparative analysis between the baseline SSSD-ECG model and our proposed **MIST-ECG** framework. This allows us to systematically quantify the impact of our mel-spectrogram informed training paradigm on three critical dimensions: trustworthiness, privacy preservation, utility, and robustness.

**Signal Fidelity and Privacy Preservation.** We first quantify the quality and trustworthiness of the synthesized signals. Table 1 shows that MIST-ECG outperforms SSSD-ECG across all fidelity and morphology metrics while also strengthening privacy. Specifically, MIST-ECG reduces RMSE by 4.68% and MSE by 4.39%; increases SNR by +0.58,dB; lowers Fourier distance by 4.68% and Hausdorff distance by 8.51%; and raises SSIM by +0.0309 (+5.15%). Privacy also improves: Membership Inference Risk (MIR) drops by 18.18%, and Nearest Neighbor Adversarial Accuracy (NNAA) decreases by 0.0056, crossing below zero (0.0047 $\rightarrow$ -0.0009), indicating attack performance worse than chance. Collectively, these gains highlight higher morphological realism and fidelity with enhanced privacy guarantees.

Table 1: Comprehensive comparison of signal fidelity, morphology, and privacy. The MIST-ECG framework excels in morphological realism and offers the strongest privacy guarantees.

| | **Fidelity & Morphology Metrics** | | | | | | **Privacy Metrics** | |
|---|---|---|---|---|---|---|---|---|
| **Training Objective** | **RMSE** $\downarrow$ | **MSE** $\downarrow$ | **SNR (dB)** $\uparrow$ | **Fourier** $\downarrow$ | **Hausdorff** $\downarrow$ | **SSIM** $\uparrow$ | **MIR** $\downarrow$ | **NNAA** $\downarrow$ |
| SSSD-ECG | 0.2114 | 0.0524 | -3.086 | 0.2115 | 1.187 | 0.6004 | 0.0099 | 0.0047 |
| MIST-ECG | **0.2015** | **0.0501** | **-2.508** | **0.2016** | **1.086** | **0.6313** | **0.0081** | **-0.0009** |

**Interlead Correlation Matrix.**

To further assess physiological coherence, we analyzed the interlead correlations, a critical property of realistic ECGs. The results are shown in Table 2. The **MIST-ECG** framework demonstrates a 70% and 78% reduction in the average absolute interlead correlation error and max absolute interlead correlation error compared to the baseline. This confirms its superior ability to capture the complex spatio-temporal dependencies between leads, a key of clinical realism.

Table 2: Average and maximum absolute inter-lead correlation error relative to real data.

| **Model** | **Avg. Corr. Error** | **Max Corr. Error** |
|---|---|---|
| SSSD-ECG (Baseline) | 0.140 | 0.491 |
| MIST-ECG (Ours) | **0.042** | **0.108** |

**Validating Synthetic Data as a Surrogate for Real Data.** The ultimate test of synthetic data is its ability to serve as a surrogate for real data when it is most needed in low-data regimes. To evaluate this, we started with training 71 disease ECG classifier model on a full corpus of synthetic data (8 folds) and measured performance as we incrementally added folds of real data for training. The results are summarized in Table 3. In the most extreme case (0 folds of real data), a classifier trained exclusively on synthetic data from our **MIST-ECG** framework achieves an AUROC of **0.640**, significantly outperforming the baseline SSSD-ECG (0.541). This proves that our training paradigm is critical for generating diagnostically useful signals from scratch. Even more remarkably, in low-data regimes (1–3 folds), hybrid datasets with **MIST-ECG** data consistently *match or outperform* the real-data-only baseline. This provides strong evidence that our high-fidelity synthetic data can serve as a viable surrogate for real data in critical, data-limited scenarios.

Table 3: Downstream AUROC when augmenting a full synthetic dataset with real data folds. In low-data regimes (0-3 folds), synthetic data from the MIST-ECG framework matches or exceeds the real-data baseline. **Bold** indicates the best performance in each column. An asterisk (*) denotes performance statistically significantly lower than the best in that column ($p < 0.05$).

| | **Number of Real Data Folds Added** | | | | | |
|---|---|---|---|---|---|---|
| **Data Type** | **0** | **1** | **2** | **3** | **8** | **Avg Rank** |
| *Baseline* | | | | | | |
| Real Data Only | — | 0.901 ± 0.009 | 0.912 ± 0.003 | 0.916 ± 0.003 | 0.927 ± 0.005 | 2.62 |
| *Synthetic Models* | | | | | | |
| Synthetic (SSSD-ECG(Baseline)) | 0.541 ± 0.074* | 0.901 ± 0.007 | **0.914 ± 0.002** | 0.917 ± 0.004 | 0.927 ± 0.005 | 2.89 |
| Synthetic (MIST-ECG(Ours)) | **0.640 ± 0.094** | **0.902 ± 0.004** | 0.911 ± 0.002* | **0.919 ± 0.004** | **0.928 ± 0.002** | **2.78** |

**Robustness in Data-Rich Augmentation Scenarios.** Having established its value in data-scarce settings, we next evaluated the robustness of the MIST-ECG framework in a data-rich environment. This experiment evaluates the marginal utility of synthetic data by starting with a complete real dataset (8 folds) and incrementally adding folds of synthetic data. The results, shown in Table 4, confirm that while performance gains naturally plateau when real data is abundant, the **MIST-ECG** framework is the clear winner. It consistently achieves the highest performance across nearly all augmentation levels, culminating in a superior Average Rank of 1.50 compared to the baseline's 3.00. This demonstrates that even when data is plentiful, the high-quality signals from our model provide the most beneficial contribution, solidifying its standing as the most robust and effective generator.

Table 4: Impact of augmenting a complete real dataset (8 folds) with synthetic data, measured by AUROC.

| | **Number of Synthetic Folds Added** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Synthetic Models** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **Avg Rank** |
| SSSD-ECG (Baseline) | 0.928 ± 0.002 | 0.930 ± 0.002 | 0.929 ± 0.004 | 0.928 ± 0.003 | 0.928 ± 0.004 | 0.928 ± 0.003 | 0.926 ± 0.004 | 0.928 ± 0.002 | 3.00 |
| MIST-ECG (Ours) | **0.930 ± 0.001** | **0.931 ± 0.003** | **0.930 ± 0.004** | **0.929 ± 0.003** | **0.929 ± 0.005** | **0.928 ± 0.004** | **0.929 ± 0.004** | **0.931 ± 0.004** | **1.50** |

# 5 Conclusion

In this work, we argue that closing the morphological fidelity gap is a prerequisite for the responsible use of synthetic ECGs in healthcare care. We addressed this foundational challenge by introducing **MIST-ECG**, a principled training paradigm that enhances a state-of-the-art diffusion model by imposing a strong, clinically relevant structural prior in the time-frequency domain. Our comprehensive evaluation demonstrated the profound impact of this approach. The MIST-ECG framework proved instrumental, clearly improving physiological coherence (4%-8% gain) and an average of 74% reduction in interlead correlation error. We also show that classifiers trained on our supplemented synthetic data can achieve performance comparable to those trained on real data in low-data regimes, establishing our method's ability to create a viable surrogate for real data. This research provides a robust and trustworthy methodology for generating high-fidelity medical time series and offers a scalable and privacy-conscious foundation for advancing ECG-based cardiac research on larger or different waveform public physiological datasets. (Code and evaluation scripts will be released upon paper acceptance.)

# References


[1] Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based conditional ecg generation with structured state space models. *Computers in biology and medicine*, 163:107115, 2023.

[2] Anne Marie Delaney, Eoin Brophy, and Tomas E Ward. Synthesis of realistic ecg using generative adversarial networks. *arXiv preprint arXiv:1909.09150*, 2019.

[3] Huan He, Shifan Zhao, Yuanzhe Xi, and Joyce C Ho. Meddiff: Generating electronic health records using accelerated denoising diffusion model. *arXiv preprint arXiv:2302.04355*, 2023.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[5] VV Kuznetsov, VA Moskalenko, DV Gribanov, and Nikolai Yu Zolotykh. Interpretable feature generation in ecg using a variational autoencoder. *Frontiers in genetics*, 12:638191, 2021.

[6] Lujundong Li, Tong Xia, Haojie Zhang, Dongchen He, Kun Qian, Bin Hu, Yoshiharu Yamamoto, Björn W. Schuller, and Cecilia Mascolo. Ecg-dpm: Electrocardiogram generation via a spectrogram-based diffusion probabilistic model. In *2024 IEEE Smart World Congress (SWC)*, pages 300–305, 2024.

[7] Manuel Martínez-Sellés and Manuel Marina-Breysse. Current and future use of artificial intelligence in electrocardiography. *Journal of Cardiovascular Development and Disease*, 10(4):175, 2023.

[8] Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, Shujie Hu, Yanqing Liu, Jinyu Li, Sheng Zhao, Xixin Wu, et al. Autoregressive speech synthesis without vector quantization. *arXiv preprint arXiv:2407.08551*, 2024.

[9] George A Mensah, Valentin Fuster, Christopher JL Murray, Gregory A Roth, Global Burden of Cardiovascular Diseases, and Risks Collaborators. Global burden of cardiovascular diseases and risks, 1990-2022. *Journal of the American College of Cardiology*, 82(25):2350–2473, 2023.

[10] Yuling Sang, Marcel Beetz, and Vicente Grau. Generation of 12-lead electrocardiogram with subject-specific, image-derived characteristics using a conditional variational autoencoder. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

[11] Weijie Sun, Sunil Vasu Kalmady, Nariman Sepehrvand, Amir Salimi, Yousef Nademi, Kevin Bainey, Justin A Ezekowitz, Russell Greiner, Abram Hindle, Finlay A McAlister, et al. Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms. *NPJ Digital Medicine*, 6(1):21, 2023.

[12] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

[13] Tong Zhou, Xuhang Chen, Yanyan Shen, Martin Nieuwoudt, Chi-Man Pun, and Shuqiang Wang. Generative ai enables eeg data augmentation for alzheimer's disease detection via diffusion model. In *2023 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-ASIA)*, pages 1–6. IEEE, 2023.

[14] Fei Zhu, Fei Ye, Yuchen Fu, Quan Liu, and Bairong Shen. Electrocardiogram generation with a bidirectional lstm-cnn generative adversarial network. *Scientific reports*, 9(1):6734, 2019.

# Appendix A: Experimental Setup Details

### Dataset and Cohort

All experiments presented in the main paper were conducted on the PTB-XL dataset [12]. This public dataset contains 21,837 clinical 12-lead ECG recordings from 18,885 patients. Each 10-second recording was sampled at 100 Hz (1,000 time steps per lead, i.e. 10 second). We used the standard patient-level data splits to ensure no data leakage, resulting in 17,441 training, 2,193 validation, and 2,203 test samples. Patient demographic information (age, BMI: calculated by height, weight) was extracted from the metadata to enable personalized conditioning.

# Appendix B: Detailed Inter-lead Correlation Analysis

A fundamental property of clinically valid 12-lead ECGs is the complex set of physiological correlations between different leads, which reflect the three-dimensional propagation of the heart's electrical wavefront. A high-fidelity generative model must successfully capture these spatio-temporal relationships. To visually and quantitatively assess this, we computed Pearson correlation matrices for real and synthetic data and visualized them as heatmaps. The following figures provide a detailed comparison.

Figure 1a shows the ground-truth correlation matrix computed from real ECGs in the PTB-XL test set. It displays well-known clinical patterns, such as the strong positive correlation between adjacent precordial leads (e.g., V1-V2) and the characteristic negative correlation between limb leads I and III. This serves as the reference against which the synthetic models are compared.

Figures 1b and 1c show the correlation matrices for the synthetic data generated by the baseline SSSD-ECG model and our proposed MIST-ECG framework, respectively. A visual inspection reveals that while the SSSD-ECG model captures the general structure, the MIST-ECG's matrix is a much closer match to the ground truth in Figure 1a.

The superiority of the MIST-ECG framework is confirmed by the difference heatmaps in Figures 1d and 1e. The difference matrix for the SSSD-ECG model (Figure 1d) shows large error patches (darker reds and blues), indicating a significant deviation from the real data's physiological structure. In stark contrast, the difference matrix for the MIST-ECG framework (Figure 1e) is substantially more muted and closer to the neutral zero-centered color, indicating a much smaller error. This visual evidence provides a clear intuition for the quantitative results reported in the main paper, where the MIST-ECG framework reduced the average absolute correlation error by 70%. This analysis provides compelling evidence that the frequency-domain supervision of the mel-spectrogram loss is crucial for generating ECGs that are not only morphologically accurate but also physiologically coherent.

# Appendix C: Outlier and Failure Mode Analysis

To better understand model limitations, we conducted an outlier analysis based on reconstruction error. We compared our MIST-ECG with demographically conditioned baselines.

### Conditioning Feature Encoding

To enable multimodal conditioning, we structured the input features as follows:

- **Clinical Labels:** The 71 SCP statement labels were decomposed into three clinically meaningful groups: Diagnostic (40 labels, e.g., `MI`), Form (19 labels, e.g., `HVOLT`), and Rhythm (12 labels, e.g., `AFIB`). Each group was one-hot encoded separately.
- **Demographic Features:** Continuous demographic variables were discretized into clinically relevant bins before being one-hot encoded:
  - **Age:** 6 bins derived from cutoffs [12, 17, 34, 54, 74].
  - **BMI:** 6 bins derived from standard clinical cutoffs [18.5, 25, 30, 35, 40].

Each one-hot encoded vector was then projected into a 32-dimensional embedding space, as described in the Methods section.

While demographically conditioned variants produces more extreme outliers, disproportionately contributing to overall error, our proposed MSIT-ECG (mel) preserves low outlier rates as the baseline model. A clinical feature analysis of these outlier cases (Figures 2) revealed that the models struggle most with **atypical physiological states, such as bradycardia (low heart rate) and low-voltage ECGs**. These cases are often under-represented in the training data and represent a key challenge for generative models, highlighting the importance of evaluating models not just on average performance but also on their robustness to rare events.

# Appendix D: Additional Visualizations

This section provides supplementary visualizations that offer qualitative support for our quantitative findings and illustrate key aspects of our methodology and its practical application.

### Qualitative Comparison of Real and Synthetic ECGs

Figure 3 provides a qualitative, side-by-side comparison of a real 12-lead ECG from the PTB-XL test set and a synthetic counterpart generated by our MIST-ECG model for the same clinical condition ('norm-sn'). This visualization serves as a visual Turing test, demonstrating the model's ability to capture not only the fundamental P-QRS-T morphology and timing but also the subtle interlead relationships and overall rhythm characteristic of a real physiological signal. The high degree of visual similarity provides qualitative support for the strong quantitative performance reported in the main paper.

### Illustrating the Mel-Spectrogram Loss Mechanism

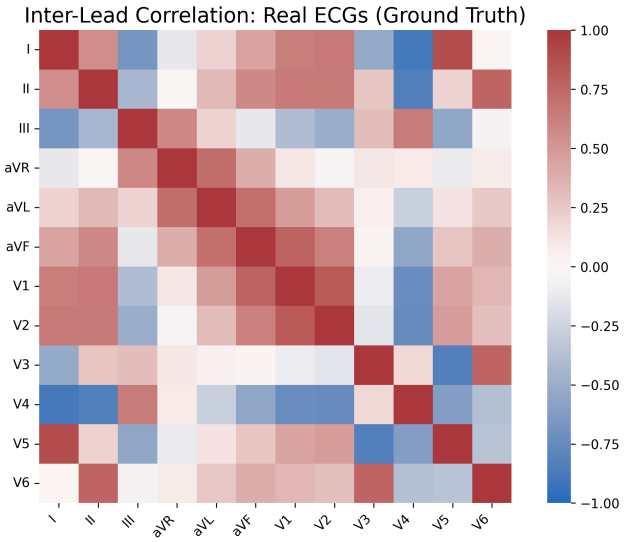
Figures 4 and 5 illustrate the core mechanism behind our mel-spectrogram loss function. They display the time-frequency representations (mel-spectrograms) of the real and synthetic ECGs shown in Figure 3, respectively. The loss function works by minimizing the pixel-wise difference between these two representations during training. The visual congruence between the two spectrograms—in terms of energy distribution across frequency bands and consistent temporal patterns—highlights how this frequency-domain supervision guides the model to reproduce the complex structural characteristics of the original signal. This directly leads to the improved morphological fidelity reported in our results.

# Appendix E: Additional tables

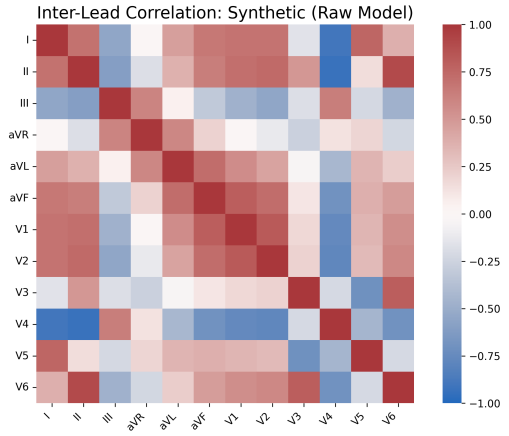
This section provides supplementary tables offer more quantitative evaluations of synthesized ECGs for downstream utility.

Table 6: Full results for the data substitution experiment, measured by AUROC (mean ± 95% CI). *p < 0.05 vs. best model in that column. Supplement to table 3 as it only highlight the low data scenario (real-data folds 1-3)

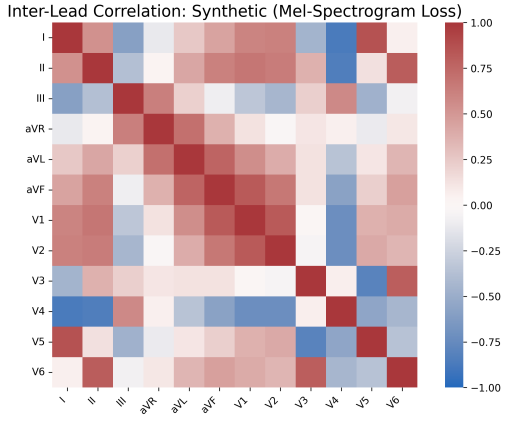
| | Number of Real Data Folds Added | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Data Type** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **Avg Rank** |
| Real Data Only | — | 0.901 ± 0.009 | 0.912 ± 0.003 | 0.916 ± 0.003 | **0.922 ± 0.005** | **0.924 ± 0.003** | **0.927 ± 0.002** | 0.926 ± 0.003 | 0.927 ± 0.005 | 2.62 |
| Synthetic (SSSG-ECG) | 0.541 ± 0.074* | 0.901 ± 0.007 | **0.914 ± 0.002** | 0.917 ± 0.004 | 0.920 ± 0.004 | 0.923 ± 0.005 | 0.926 ± 0.005 | **0.928 ± 0.003** | 0.927 ± 0.005 | 2.89 |
| Synthetic (MIST-ECG) | **0.640 ± 0.094** | **0.902 ± 0.004** | 0.911 ± 0.002* | **0.919 ± 0.004** | 0.920 ± 0.005 | 0.923 ± 0.004 | 0.925 ± 0.002 | 0.926 ± 0.004 | **0.928 ± 0.002** | **2.78** |



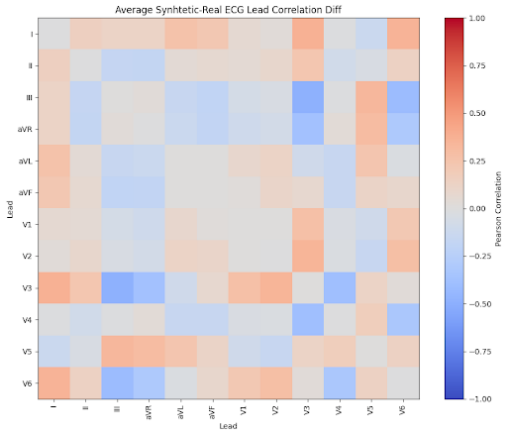

(a) The ground-truth inter-lead correlation matrix for real ECG data.



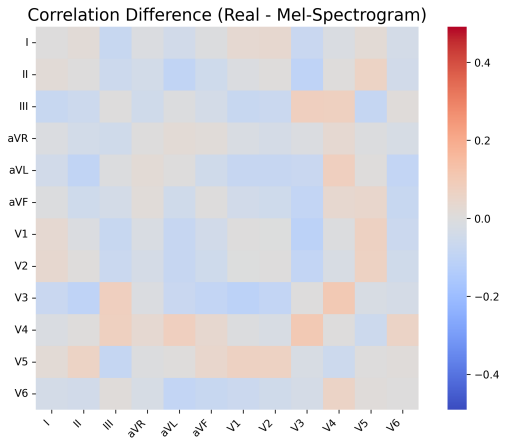

(b) Inter-lead correlation matrix for synthetic data from the baseline `SSSD-ECG` model.




(c) Inter-lead correlation matrix for synthetic data from our MIST-ECG framework.




(d) Difference matrix (Real - SSSG-ECG). Darker colors indicate larger errors.




(e) Difference matrix (Real - MIST-ECG). The significantly paler colors demonstrate the superior performance of our method.

Figure 1: Comparison of correlation difference matrices for the baseline and proposed models.

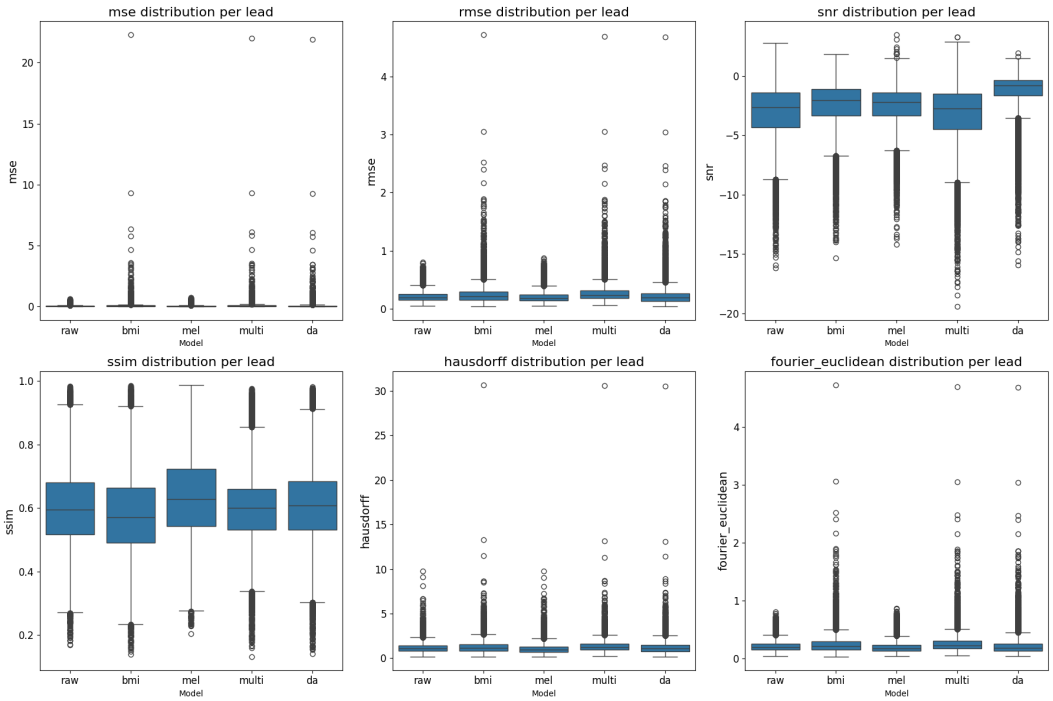



Figure 2: Statistical and Morphological metric distribution across baseline SSSD-ECG (raw) and 4 variants: mel - MIST-ECG variant, multi - Disease + BMI + Age conditioned variant, bmi - Disease + BMI conditioned variant, da - Disease + Age conditioned variant. Boxplots show that while demographic conditioning models achieve higher SNRs compared to baseline, they exhibit a larger number of extreme outliers in error metrics (MSE, RMSE, Hausdorff distance, Fourier Transform distance), indicating greater variability and consistent occasional failure cases. MIST-ECG variant shows persisting lower outliers in error metrics.

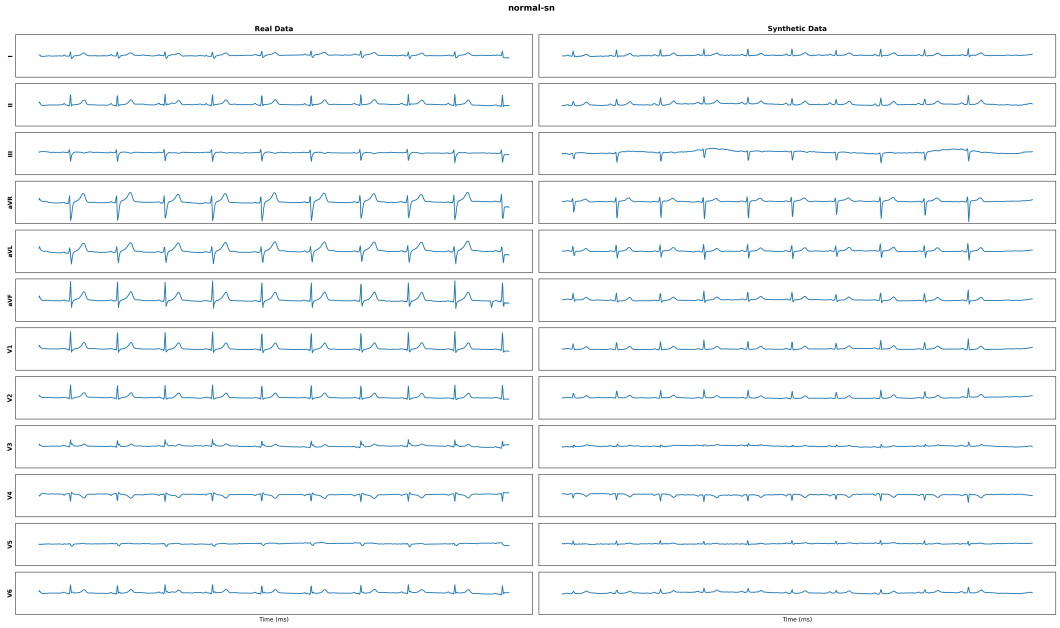



Figure 3: Comparison of real and synthetic 12-lead ECG signals for disease code 'norm-sn', with the synthetic sample generated by the MSIT-ECG model described in Table 1.

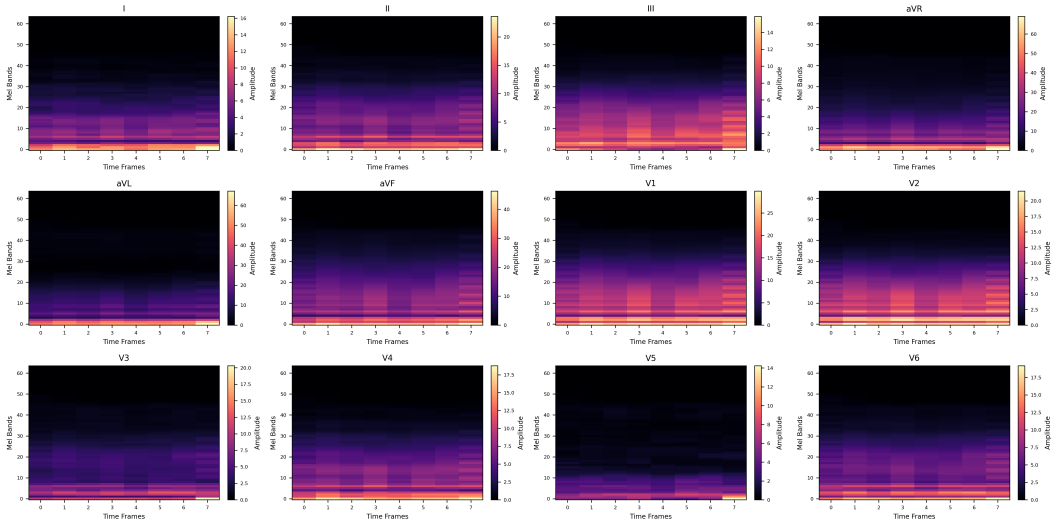



Figure 4: Mel-spectrogram visualization of the real 12-lead ECG signal (shown in Figure 3) after applying the Short-Time Fourier Transform (STFT)

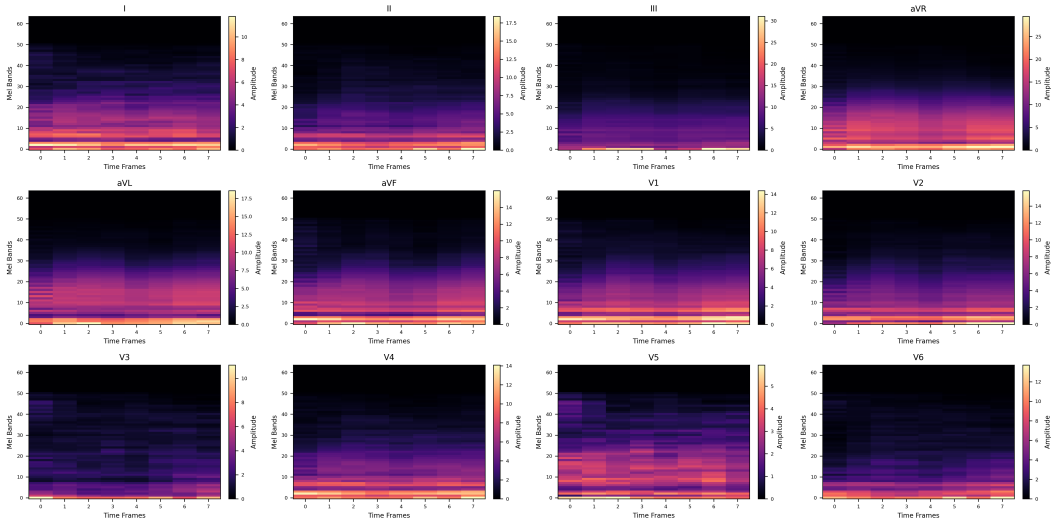



Figure 5: Mel-spectrogram visualization of the synthetic 12-lead ECG signal for disease code 'norm-sn' (shown in Figure 3) after applying the Short-Time Fourier Transform (STFT).