Supplemental Material:

1 Datasets:

1.1 Human participant dataset 1

A dataset was collected by Shandhi et al. and Hersek et al. for the purpose of two studies: (1) to study the accuracy of PEP estimation based on the data collected with accelerometers and gyroscopes individually or in combination, and (2) to train a U-Net model for mapping SCG signals to BCG signals to augment BCG to datasets with only SCG recordings [1, 2]. In this protocol, a total of 26 healthy participants (10 females, and 16 males) went through a protocol (H18452) approved by the IRB, designed to non-invasively induce changes in PEP [1, 2]. The protocol begins with a rest baseline period of 5 minutes followed by a 3 minute walk on a treadmill and a 1.5 minute squatting exercise ending in a 5 minute recovery period [1, 2]. One accelerometer, and one gyroscope were placed on the mid-sternal line respectively above and below the midpoint of the supersternal notch and the xiphoid process [1, 2]. For the purpose of this study we only used the z-axis accelerometer data in this dataset which was collected from the mid-sternum.

1.2 Human participant dataset 2

A dataset was collected by Hersek et al. for the purpose of testing the U-Net model trained for SCG to BCG mapping [2]. In this protocol, a total of 10 healthy participants (5 females, and 5 males) went through a protocol (H18452) approved by the IRB, designed to non-invasively induce changes in PEP [2]. The protocol begins with a rest baseline period of 5 minutes standing on a BCG scale, followed by a 20 seconds Valsalva maneuver, a 2 minutes rest period, a 2 minutes walk on a treadmill, and a 1.5 minute squatting exercise ending in a 5 minute recovery period on the BCG scale [2]. One accelerometer was placed on the mid-sternum and a wireless 3-lead ECG sensor was used for ECG data collection [2]. For the purpose of this study we only used the z-axis accelerometer data in this dataset which was collected from the mid-sternum.

1.3 Human participant dataset 3

A dataset was collected by Ashouri et al. to study the effect of sensor placement for SCG recordings on PEP estimation accuracy [3]. In this study, a total of 10 healthy participants (5 females, and 5 males) participated in a two day protocol (H13512) approved by the IRB, with the same activity tasks on both days but different sensor placements. The protocol was designed to study the effect of sensor placement on signal morphology over a wide range of PEP values. The protocol for each day starts with a 1 minute baseline, followed by a 1 minute stepping exercise, and ending in a 5 minute recovery period [3]. On each day, three accelerometers were placed on the chest, day 1, on the mid-sternum, 7.5 cm to the left, and 7.5 cm to the right, and day 2, on the mid-sternum, 5 cm above and 5 cm below [3]. For the purpose of this study we only used the z-axis accelerometer data of this dataset which was collected from the sensor placed on the mid-sternum.

1.4 Human participant dataset 4

A dataset was collected by Gurel et al. [4] to design a classifier that uses non-invasively sensed physiological signals to differentiate between three protocol tasks: rest, mental arithmetic, and N-back tasks. In this study, a total of 16 healthy participants (6 females, and 10 males) went through a protocol (H13512) approved by the IRB, consisting of 6 phases consisting of mental arithmetic [5] and N-back (a measure of working memory function [6]) [7] tasks designed to induce mental stress resulting in cardiovascular responses. The protocol starts with a 3 minute baseline followed by 6 tasks (three arithmetic and three N-back) with 3 minute rest periods in between each two task ending with a final 3 minute rest period [4]. During the entirety of this protocol, the participants were told to keep their eyes closed. For each participant, one SCG sensor was placed on the mid-sternum, and a wireless 3-lead ECG sensor was used for ECG data collection [4]. For the purpose of this study, we only used the z-axis accelerometer data collected from the mid-sternum.

1.5 Human participant dataset 5

A dataset was collected by Chan et al. [8] to demonstrate oxygen saturation (SpO_2) estimation using a chest-worn wearable patch biosensor. In this study, 20 healthy participants (6 females, and 14 males) went through a breath-hold protocol (H21100) approved by the IRB, where each participant was asked to to hold their breath for as long as they can for 10 times separated by 1 minute rest periods between breath-holds [8]. For each participant, a wearable patch biosensor was attached to the mid-sternum (Collecting SCG, single-lead ECG, and PPG signals), and a wireless 3-lead ECG sensor was used for ECG data collection [8]. For the purpose of this study we only used the z-axis accelerometer data from this dataset which was collected from the patch placed on the mid-sternum.

1.6 Animal subject dataset 6

A dataset was collected by Zia et al. [9] to study how cardiovascular features change during hemorrhage and develop a globalized model for blood volume status estimation. In this study, 6 pigs went through an experimental protocol (A100276) approved by the IACUC, to induce hypovolemia where blood was progressively drawn from the pig until blood volume reached one of four levels (7, 14, 21, and 28 percent of total blood volume). For each subject, an SCG sensor was placed on the mid-sternum, PPG sensors were placed on the femoral artery branches, and a 3-lead ECG sensor was used for ECG data collection [9]. For the purpose of this study, we only used the z-axis accelerometer data collected from the mid-sternum during baseline (before any blood was drawn) to allow for comparison to healthy subject SCG waveforms. This animal subject dataset was solely used for validating the generative model described in section **??**. The pig dataset was to test the synthetic signals are acquired signals of the same modality but from a demographic with completely different physiology.

2 Pre-processing

2.1 Heartbeat separation

Using the R-peaks of the ECG signals collected concurrently with the SCG signals, the filtered signals were heartbeat-separated from 80 ms before the R-peak to 560 ms after the R-peak. This region was chosen to make sure the whole SCG beat is present in the region while keeping the interval short for computational time reduction.

2.2 Signal quality indexing

For signal quality indexing (SQI) a method introduced in prior work for SCG beats was used to identify and exclude the beats contaminated with noise above a certain threshold [10]. The exponential factor (λ) of 25 was chosen, and beats with SQI values below 0.5 were rejected [11]. Finally, the remaining beats were down-sampled to 250 Hz to reduce the computational cost while keeping the SCG frequency components.

Finally, the beats are min-max normalized and centered around 0.5 to ensure model generalization.

2.3 Skeleton signal

The skeleton signal consists of two Gaussian waveforms in the AO and AC locations. The Gaussian waveform at the AO location has a fixed width of 160 ms (variance=28 ms), and the Gaussian waveform at the AC location has a fixed width of 112 ms (variance=20 ms). These widths are chosen to cover the majority of S1 and S2 portions of a real SCG beat respectively. This skeleton signal is a simple representation of the SCG features.

3 Model architecture:

3.1 Scaled positional encoding

Transformers use multi-head attention blocks and, unlike RNNs, process input tokens in parallel. This results in accelerated training; however, the model doesn't capture the information about the order of the input tokens. Thus, for sequence modeling tasks, the transformer model needs to receive positional information in some format. This is usually done by adding positional information to the token embeddings themselves. One method, which we used in this work and is often used by literature, is triangular positional embedding [12,13]. The triangular positional embedding is defined as follows:

$$PE(pos, 2i) = sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{1}$$

$$PE(pos, 2i+1) = cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{2}$$

Where pos is the token step, i is the embedding index, and d_{model} is the dimension of embedding. For natural machine translation (NMT), the positional embedding for the encoder and decoder have the same scales. However, in this work since the encoder side receives SCG features, and the decoder side generates an SCG waveform, having embeddings with the same scales might cause constraints on these two separate spaces. To overcome this, as suggested by prior work, we multiply a trainable weight to relax the constraints and let the embeddings adapt to the scaling of the two sides [13].

3.2 Embeddings

To make the input readable for the transformer network, inputs need to be mapped (encoded) into embedding vectors. Samples are discretized to tokens and then a token embedding block will map each token to an embedding vector that can be fed into the transformer model. Thus, the performance is determined by (1) how these samples are split into tokens, and (2) the embedding that maps these tokens to embedding vectors. For instance, in NLP tasks, a string of text is usually split into words, each given a token ID integer using a tokenizer. Each token is then mapped to an embedding vector. This embedding can be as simple as a trainable embedding layer added to the network, or more complex such as Word2Vec where the embeddings are a 1-to-1 mapping, or BERT where embeddings are contextual and depend on the other tokens in the input sentence [14, 15].

To use the transformer architecture for SCG signals, a tokenization and embeddings similar to word tokenization and embeddings but relevant to the SCG signals subspace is necessary. In this work we propose tokenization as splitting the SCG beat into time windows and embeddings as features for each time window. Three different fixed embeddings were explored in this work: (1) Spectrogram with each time window (128 ms window with 75% overlap) as one token with an embedding vector length equal to the number of frequency bins (128 bins), (2) Maximal overlap discrete wavelet transform (MODWT) using the Haar wavelet with each time window as one token with embedding vector length equal to the number of decomposition levels (8 levels), (3) A pre-trained embedding block which is the encoder block extracted from an auto-encoder with a 4-layer CNN encoder, and a 4-layer CNN decoder with ReLu activation trained on the training set for 60 epochs. This pre-trained encoder converts the input beat sample of size 160, to 20 tokens (through the CNN network with receptive field of 8) with embedding vector length of 64 (resulting in a 20*64 image). These embedding blocks were added to both encoder and decoder sides for the skeleton input waveform embeddings, and the SCG beat embeddings, respectively.

Note that MODWT is a type of wavelet transform and was chosen instead of continuous wavelet transform (CWT) as it enables perfect reconstruction of the signal which is required in this work and is not possible using CWT [16].

Fig. 2(b) represents the tokenizer and embedding block, and Fig. 2(h) represents the SCG reconstruction block that is responsible for reconstructing the SCG beat from the predicted embedding vectors.

To compare the performance between these three embeddings we replaced the tokenizer/embedding block in the architecture (Fig. 2(b)) while keeping the rest of the architecture unchanged, and trained the model on the same training dataset and evaluated on the test dataset.

3.3 Encoder and decoder

Both encoder and decoder blocks of the transformer architecture (Fig. 2(d, f)) consist of two main layers: attention layers and fully connected layers. The attention layer receives a query as input, and returns an output based on a set of key-value pairs that it has as its memory. This output is a weighted sum of values where each value has a weight assigned as a function of the query and the key [12]. The encoder block (Fig. 2(d)) learns the relations in the input sequence. The self-attention block in the encoder attends to all tokens coming from the previous layer. The decoder block (Fig. 2(f)) is responsible for generating the SCG beat using the information it gains from the input SCG samples combined with the information coming from the encoder. The self-attention block in the decoder can only attend to the same token or the previous tokens. Attending to the tokens after the current token is blocked using attention masks. The encoder-decoder cross-attention block in the decoder receives the key-value pairs from the encoder and the query from the previous layer of the decoder.

3.4 Encoder and decoder pre-net

In the architecture proposed here, the source signal is a skeleton signal and the target signal is a real SCG beat. Because, the two subspaces are not the same, adaptation steps are required to project the signals from the source and target domains into the same subspace. For this, we implemented a pair of trainable pre-net blocks responsible for projecting the fixed embeddings of both encoder and decoder inputs to a more flexible subspace to increase compatibility between the input and output domains. These blocks consist of two layer fully connected networks with ReLU activation similar to the TTS transformer architecture in prior work. For both pre-net blocks, we added a linear projection layer for center consistency after the ReLU activation of the final layer , This projection layer resolves the problems that may cause from adding a positional embedding in range of [-1,1] (equation 1, 2) to the output of ReLU which is in range of $[0,+\infty]$ [13].

3.5 Decoder post-net

Because the pre-net transforms the tokens from the SCG domain to a different subspace, we use a decoder post-net block (Fig. 2(g)) to revert this transformation on the generator tokens.

We implemented a post-net block (Fig. 2(g)) responsible for this task that predicts the token embeddings from the output tokens generated by the transformer. This block consists of a linear projection to predict the embeddings and a 5-layer CNN to refine the prediction of these embeddings using a ResNet architecture similar to the TTS transformer architecture in prior work [13, 17].

4 Hyperparameter tuning:

To optimize the performance of the model we tuned some of the hyperparameters using a grid search. These hyperparameters include: learning rate warm-up parameters (warm-up steps), number of encoder/decoder blocks in the transformer, number of attention heads, batch size, and number of layers in the pre-net and post-net blocks. To evaluate the performance we trained the model on the training set and measured its performance using the validation set and the distance metrics explained in section **??**. Hyperparameter tuning results are shown in the Supplementary Tables (Tables S2, S3, and S4).

5 Distance metrics for model evaluation:

5.1 Maximum Mean Discrepancy (MMD)

Maximum Mean Discrepancy (MMD) test statistic measures whether two distributions are different based on the samples drawn from each. It does so by computing the norm of the difference between the feature distributions' means in the reproducing kernel Hilbert space [18]. MMD was calculated between each pair of datasets (that contain SCG beat samples) using equation (3) [18].

$$MMD^2{}_{U,U}(P^n||Q^m) = \frac{1}{n(n-1)}\sum_{i\neq j} k(x_i, x_j) + \sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i, y_j) + \frac{1}{m(m-1)}\sum_{i\neq j} k(y_i, y_j) \tag{3}$$

Where P is the true distribution, and Q is the approximated distribution. n and m are the number of samples drawn from P and Q respectively. x and y are 1-D SCG beat samples drawn from P and Q respectively.

5.2 Sliced-Wasserstein Distance (SWD)

The Wasserstein distance calculates the cost of transforming one distribution to another [19]. The Sliced-Wasserstein distance is the cost between multi-dimensional distributions calculated using the 1-D Wasserstein distance between projections of these higher-dimensional distributions to random 1-D spaces [19]. To calculate the SWD, we obtain 1-D representations of the SCG beat samples in the datasets by projecting them to random 1-D spaces and aggregate the 1-D Wasserstein distances between these 1-D representations using equation (4) [20]. We used the same algorithm validated in prior work by Karras et al. for SWD calculations where inverse empirical cumulative distribution function is calculated using the 1-D representations [21].

$$W_1(P^n||Q^m) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt \tag{4}$$

Where P is the true distribution, and Q is the approximated distribution. n and m are the number of samples drawn from P and Q respectively. And, F^{-1}, and G^{-1} are quantile functions of P, and Q respectively.

5.3 Kullback–Leibler divergence (KLD)

The Kullback-Leibler divergence is a metric to measure the distance between two distributions [22]. The KL divergence is calculated using equation 5 [23]. Note that KL divergence is asymmetric and measures how far an approximated distribution is from the true distribution. To estimate KL divergence for data consisting of continuous vectors, we can add an extra step that first estimates the two distributions P and Q by computing their nearest neighbor estimates, and then calculate the KL divergence [23–25].

$$D_{KL}(P^n||Q^m) = -\frac{d}{n}\sum_{i=1}^{n} log(\frac{r_k(x_i)}{\hat{s_k}(x_i)}) + log(\frac{m}{n-1}) \tag{5}$$

Where P is the true distribution, and Q is the approximated distribution. n and m are the number of samples in P and Q respectively, and d is the length of each sample (length of the SCG beat). And, $r_k(x_i)$ and $s_k(x_i)$ are the Euclidean distances of x_i drawn from P to the k^{th} nearest-neighbor of x_i in P and Q samples respectively. For $r_k(x_i)$ we chose k=2 because the nearest-neighbor of x_i to P would be itself, so we calculate the distance to the second nearest neighbor. For $r_k(x_i)$ we chose k=1.

6 Morphological variability:

In order to validate that the random ID token induces inter-participant variability, we generated 16 subjects where each has a unique random ID with linearly varying PEP and LVET parameters. Using dynamic time warping (DTW), we measured the average distance between the waveforms belonging to each participant and the distance between the waveforms of two separate participants. The results showed that the DTW distance is 0.51 ± 0.06 within beats of each participant and 0.77 ± 0.24 between the beats of two different participants. The DTW distances within the beats of each participant and between participants imply that on average the intra-participant variability is less than inter-participant variability. The higher variance in inter-participant distances indicates the fact that some patient pairs can have more similar SCG morphologies compared to other pairs.

7 Demystifying the transformer model:

The attention weights of the transformer block are visualized by heat-maps in Fig. S3. The self-attention blocks in the encoder and decoder indicate how different parts of an embedding sequence (from the previous layer) interact with itself. And the cross-attention blocks in the decoder indicate how an embedding sequence from the previous layer of the decoder interacts with the output sequence of the encoder (memory). In the transformer architecture, the encoder block encodes the skeleton signal input, and passes key-value pairs (memory) to the decoder block. The decoder will use these key-value pairs that hold information learned from the encoder input to generate the output sequence.

We hypothesize that for generating realistic SCG beats with controllable features, the decoder block fuses information from the decoder input (real SCG beats) and information encoded through the encoder from the skeleton signal (desired features). Fig. S3 shows a visualization of the self-attention weights for the first encoder and decoder layer in form of heat-maps. The encoder self-attention weights (Fig. S3(a)) show that input tokens which are parts of the skeleton signal are focusing on the two Gaussian peaks representing the AO, and AC location and amplitude of the SCG beat. The decoder self-attention block shows diagonal attentions for most of the heads meaning the real SCG signal tokens attend strongly either to themselves, or to a token with a constant offset, especially in the S1 and S2 regions (Fig. S3(b)). In addition, in head 8 of the decoder self-attention we can observe that the S1 and S2 complexes are attending to the random ID token which holds global information such as the morphology. The encoder-decoder attention weight (Fig. S3(c)), visualizes how the information from the encoder are fused with the information from the previous decoder layer. It can be observed that the S1 and S2 portions of the real SCG beats are attending to the S1 and S2 portions of the skeleton signal. This means that the generated SCG output is build using the information from the skeleton input which is what we expect the model learn in this translation task. For example, attention from S1 portion of the real SCG to S1 portion of the skeleton can be observed in heads 1, 2, 3, 4, 5, and 7, and attention from S2 portion of the real SCG to S2 portion of the skeleton can be observed in heads 1, 2, 4, 5, 7, and 8. Some other combinations of attention is also observed through different attention heads. For example, head 3 shows the S1 portion of the real SCG is attending to both S1 and S2 portions of the skeleton signal, and head 8 shows the S2 portion of the real SCG is attending to both AO and AC portions of the skeleton signal.

References

[1] M. M. H. Shandhi, B. Semiz, S. Hersek, N. Goller, F. Ayazi, and O. T. Inan, "Performance analysis of gyroscope and accelerometer sensors for seismocardiography-based wearable pre-ejection period estimation," *IEEE journal of biomedical and health informatics*, vol. 23, no. 6, pp. 2365–2374, 2019.

[2] S. Hersek, B. Semiz, M. M. H. Shandhi, L. Orlandic, and O. T. Inan, "A globalized model for mapping wearable seismocardiogram signals to whole-body ballistocardiogram signals based on deep learning," *IEEE journal of biomedical and health informatics*, vol. 24, no. 5, pp. 1296–1309, 2019.

[3] H. Ashouri and O. T. Inan, "Automatic detection of seismocardiogram sensor misplacement for robust pre-ejection period estimation in unsupervised settings," *IEEE sensors journal*, vol. 17, no. 12, pp. 3805–3813, 2017.

[4] N. Z. Gurel, H. Jung, S. Hersek, and O. T. Inan, "Fusing near-infrared spectroscopy with wearable hemodynamic measurements improves classification of mental stress," *IEEE sensors journal*, vol. 19, no. 19, pp. 8522–8531, 2018.

[5] R. Li-Mei Liao and M. G. Carey, "Laboratory-induced mental stress, cardiovascular response, and psychological characteristics," *Rev. Cardiovasc. Med*, vol. 16, pp. 28–35, 2015.

[6] A. M. Owen, K. M. McMillan, A. R. Laird, and E. Bullmore, "N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies," *Human brain mapping*, vol. 25, no. 1, pp. 46–59, 2005.

[7] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during n-back task—quantified in the prefrontal cortex using fnirs," *Frontiers in human neuroscience*, vol. 7, p. 935, 2014.

[8] M. Chan, V. G. Ganti, J. A. Heller, C. A. Abdallah, M. Etemadi, and O. T. Inan, "Enabling continuous wearable reflectance pulse oximetry at the sternum," *Biosensors*, vol. 11, no. 12, p. 521, 2021.

[9] J. Zia, J. Kimball, C. Rolfes, J.-O. Hahn, and O. T. Inan, "Enabling the assessment of trauma-induced hemorrhage via smart wearable systems," *Science Advances*, vol. 6, no. 30, Jul. 2020.

[10] J. Zia, J. Kimball, S. Hersek, M. M. H. Shandhi, B. Semiz, and O. T. Inan, "A unified framework for quality indexing and classification of seismocardiogram signals," *IEEE journal of biomedical and health informatics*, vol. 24, no. 4, pp. 1080–1092, 2019.

[11] A. H. Gazi, S. Sundararaj, A. B. Harrison, N. Z. Gurel, M. T. Wittbrodt, M. Alkhalaf, M. Soudan, O. Levantsevych, A. Haffar, A. J. Shah *et al.*, "Transcutaneous cervical vagus nerve stimulation inhibits the reciprocal of the pulse transit time's responses to traumatic stress in posttraumatic stress disorder," in *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 1444–1447.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] Z. Zhang, Q. K. Telesford, C. Giusti, K. O. Lim, and D. S. Bassett, "Choosing wavelet methods, filters, and lengths for functional brain network construction," *PloS one*, vol. 11, no. 6, p. e0157243, 2016.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[19] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, "Optimal mass transport: Signal processing and machine-learning applications," *IEEE signal processing magazine*, vol. 34, no. 4, pp. 43–59, 2017.

[20] A. Ramdas, N. G. Trillos, and M. Cuturi, "On wasserstein two-sample testing and related families of nonparametric tests," *Entropy*, vol. 19, no. 2, p. 47, 2017.

[21] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[22] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[23] F. Pérez-Cruz, "Kullback-leibler divergence estimation of continuous distributions," in *2008 IEEE international symposium on information theory*. IEEE, 2008, pp. 1666–1670.

[24] Q. Wang, S. R. Kulkarni, and S. Verdú, "A nearest-neighbor approach to estimating divergence between continuous random vectors," in *2006 IEEE International Symposium on Information Theory*. IEEE, 2006, pp. 242–246.

[25] N. Leonenko, L. Pronzato, and V. Savani, "A class of rényi information estimators for multidimensional densities," *The Annals of Statistics*, vol. 36, no. 5, pp. 2153–2182, 2008.

Supplemental Tables:

Table. S1: Model Summary

Layer	**Description**
SCG Embedding	MODWT
Encoder/Decoder Prenet	(2x) FCN + ReLU linear projection
Positional Encoding	Scaled
Transformer Encoder Block	(3x) Encoders (8-heads, 20 tokens)
Transformer Decoder Block	(3x) Decoders (8-heads, 20 tokens)
Decoder Post-net	Linear projection 5-layer CNN (Res-Net)
SCG Reconstruction	Inverse MODWT

Table. S2: Number of Encoder Decoder Blocks (N)

Layer Number	SWD	MMD	KLD
2-layer	0.40	0.10	113.85
3-layer	**0.39**	**0.09**	**106.49**
4-layer	0.46	0.13	118.54

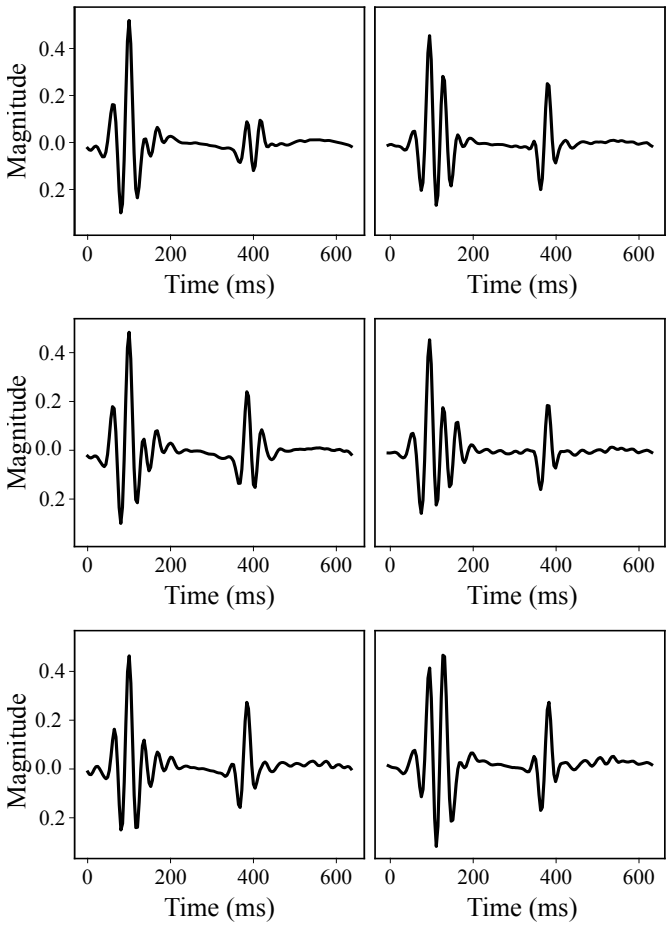
Table. S3: Number of Attention Heads

Heads	SWD	MMD	KLD
4-head	0.42	0.10	116.83
8-head	**0.39**	**0.09**	**106.49**
16-head	0.51	0.16	133.17

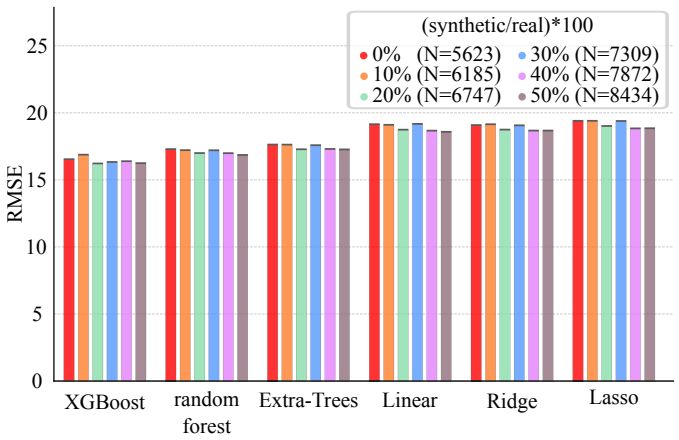
Table. S4: Comparison of Scaled and Original Positional Encoding

Type	SWD	MMD	KLD
Original	0.40	0.10	112.04
Scaled	**0.39**	**0.09**	**106.49**

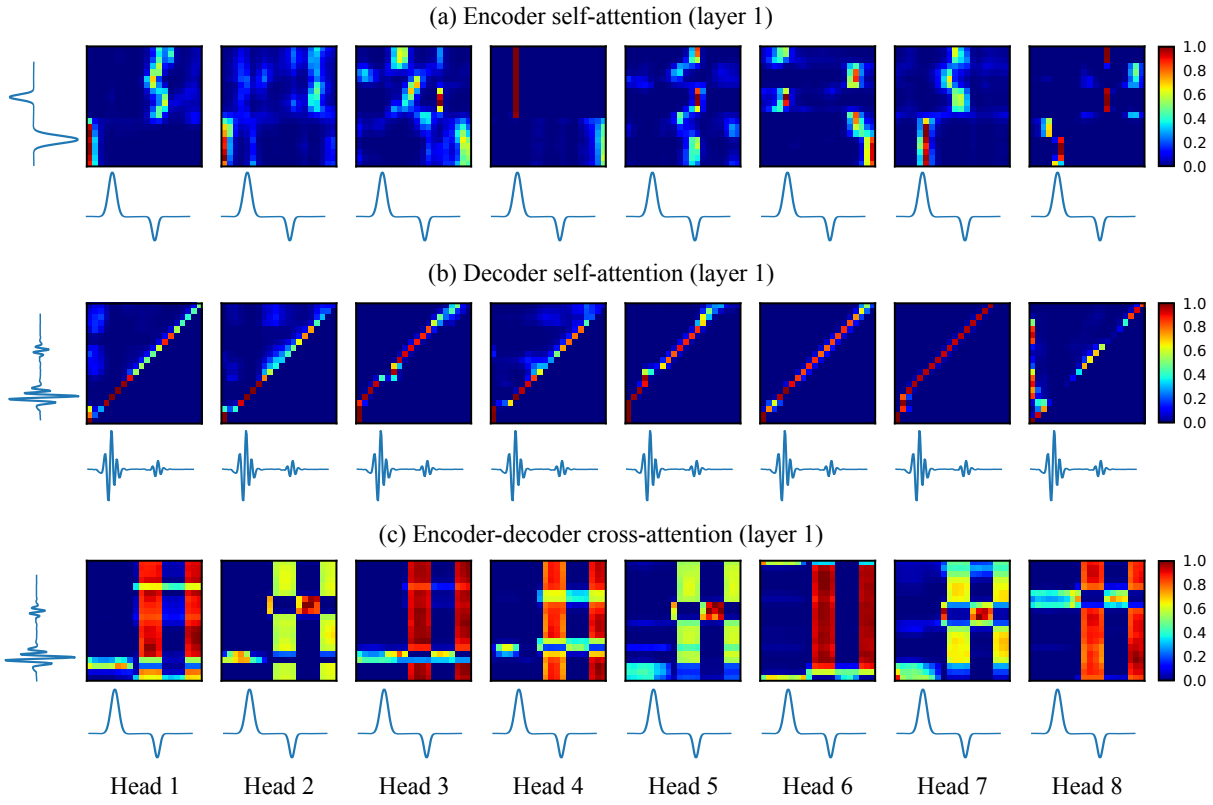
Supplemental Figures:



Supp. Figure. S1: SCG signals generated with different morphologies by varying the synthetic ID token (Fig. 1(a)) and keeping other input parameters constant (PEP, LVET, AO amplitude, and AC amplitude). We can observe that the generated beats have different morphologies for the S1 and S2 portions, which suggests inter-participant variability of the generated beats controlled by the random ID token.



Supp. Figure. S2: Comapring the RMSE for PEP estimates (using the same models as in prior work [31]) with dataset being augmented with different amounts of synthetic data without varying PEP and LVET parameters to experiment the effect of augmenting data without diversity.



Supp. Figure. S3: Attention plots from the trained SCG generator model after 60k steps to demystify the model, visualizing how the model learns the SCG representation. Attention matrices visualize the regions of the inputs that the model focuses at each layer to generate the output. The heat-maps show significant attention on S1 and S2 regions of the SCG that contain the most information about SCG which is what the model is expected to learn. Note that raw input output waveforms are shown for better interpretation only while the actual input and output for each layer is a transformed version of the raw input and output. (a) Encoder self-attention matrix, (b) Decoder self-attention matrix, (c) Encoder-decoder cross-attention matrix.