

000 X-ECG: EXPLAINABLE FOUNDATION MODEL 001 FOR ELECTROCARDIOGRAM 002

003 **Anonymous authors**

004 Paper under double-blind review

005 ABSTRACT

006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053
Electrocardiography (ECG) is a cost-effective and widely accessible tool for evaluating cardiac health. While numerous machine learning methods have been developed to assist cardiologists in diagnosis, many suffer from lacking explainability, making it difficult to understand why a particular disease is classified. To address this limitation, we introduce X-ECG, an explainable ECG foundation model. To train this model, we first curate wave-level anomalies annotations on public datasets, using a rule-based algorithm that finds abnormal waves, intervals or segments in ECG signal according to established clinical knowledge. To help models learn where to focus, we propose an attention-guided training approach that enables the model to highlight relevant regions. To the best of our knowledge, X-ECG is the first ECG foundation model with built-in explainability. Our experiments show that using our dataset to guide the model not only adds explainability but also improves performance in arrhythmia classification and report generation tasks.

1 INTRODUCTION

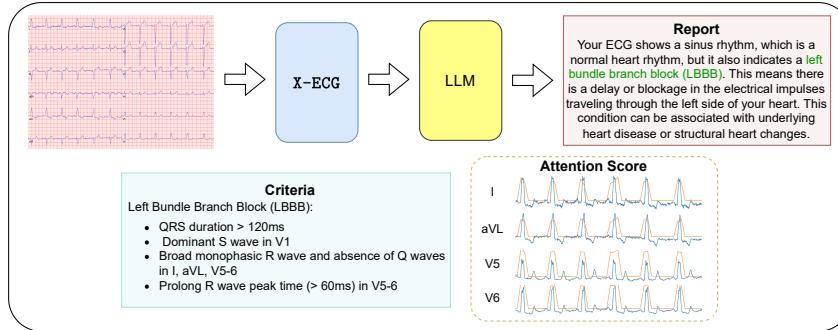


Figure 1: X-ECG is capable of identifying abnormal regions that contribute to its output, thereby enhancing model interpretability. This is achieved by guiding its attention scores during training to focus on these critical regions.

The electrocardiogram (ECG) is a monitoring tool that provides a window into the heart's electrical activity and offers valuable insights into cardiac health. To interpret an ECG sample, cardiologists typically follow a predefined process, examining specific regions and assessing abnormalities. A full and detailed interpretation can take up to five minutes, depending on the cardiologist's level of experience Bortolotti et al. (2025).

To accelerate and support this process, numerous machine learning methods have been developed. Many groups have proposed classification models with excellent performance across different benchmarks Strothoff et al. (2020); Kiranyaz et al. (2016); Wang et al. (2021); Herman et al. (2024), while others have focused on building foundation models that can be adapted to a wide range of downstream tasks Song et al. (2025); McKeen et al. (2025); Wang et al. (2025); Jin et al. (2025); Nguyen et al. (2025). Notably, McKeen et al. (2025) and Nguyen et al. (2025) extend this progress by addressing challenges in interpreting corrupted ECG signals.

054 However, to fully aid doctors in performing a diagnosis, instead of just returning the final disease
 055 class, it is desirable for the model to highlight the regions that contribute to such decision making,
 056 as it helps the doctors to understand the decision. While anomaly detection and localization models
 057 like Bui et al. (2024); Jiang et al. (2023b; 2024) can do exactly that by highlighting irregular beats,
 058 they do not provide diagnosis from those anomalies.

059 Existing ECG datasets, most notably MIMIC-IV-ECG and PTB-XL, typically have the waveform
 060 and metadata containing the disease class(es) associated with each waveform. These datasets how-
 061 ever do not have locations of the anomalies on the waveforms. Models are typically trained following
 062 a black-box training approach on these massive dataset with the waveform as input and class label
 063 as outputs, leading to models with lack of explainability. Therefore, fine-grained annotations of
 064 anomalies associated with each disease class is crucial for building explainable ECG systems.

065 In this study, we curate an algorithm that segments anomalous waves, segments, and intervals that are
 066 associated with each cardiac disease, based on an established clinical knowledge database. Specifi-
 067 cally, we design a two-stage procedure which first finds the abnormal waves, intervals and segments
 068 for each lead in the ECG signal, and then to further finetune the annotations, we employ the Criteria
 069 Feature Retrieval (CFR) proposed in Nguyen et al. (2025) to only highlight the abnormal compo-
 070 nents contribute to the final report. We apply this algorithm on MIMIC-IV-ECG and PTB-XL, cre-
 071 ating fine-grained, wave-level anomaly annotations, which we will refer to as MIMIC-IV-ECG+X
 072 and PTB-XL+X respectively. We then propose a training framework guided by such heatmaps from
 073 those dataset to build X-ECG, an ECG foundation model with explainability. To showcase its utility,
 074 we augment X-ECG with a classification module for cardiac diagnosis and an LLM for automated
 075 report generation. By utilizing this dataset in guiding the attention mechanism of the ECG encoder,
 076 the model not only have a boost in arrhythmia classification and report generation performance, but
 077 also can show the location of anomalies contribute to the final decision.

078 The main contributions of the paper is summarized as below:

- 079 • We construct MIMIC-IV-ECG+X and PTB-XL+X, a dataset for training and benchmark-
 080 ing the model ability to segment abnormal locations respectively.
- 082 • To help the model in efficiently learning the correct attention location, we employ a
 083 Attention-guided mechanism that utilizes that heatmap from X-ECG and model attention
 084 weights.
- 085 • We demonstrate that the guided model have a boost performance in various tasks, such as
 086 arrhythmia classification, report generation and anomaly localization.

088 2 RELATED WORK

090 2.1 FOUNDATION MODELS FOR ECG

092 In recent years, foundation models have emerged as a transformative paradigm across various do-
 093 mains, including ECG analysis. These models are typically pretrained on large-scale datasets and
 094 subsequently adapted to a wide range of downstream tasks through transfer learning or fine-tuning.
 095 To build such models, researchers often adopt pretraining frameworks like multimodal alignment or
 096 self-supervised learninig.

097 METS Li et al. (2023) and MERL Liu et al. (2024) employ the dual-alignment strategy to align
 098 ECG signals with their corresponding text reports, enabling the learning of high-quality ECG rep-
 099 resentations. Building upon METS, MERL introduces an inter-alignment module within the signal
 100 modality to further refine these representations. By leveraging these techniques and training on
 101 large-scale datasets, both METS and MERL demonstrate the capability for zero-shot cardiac clas-
 102 sification—allowing the identification of previously unseen diseases without the need for additional
 103 labeled data.

104 Similarly, ESI Yu et al. (2024), ECG-Chat Zhao et al. (2025) and TolerantECG Nguyen et al. (2025)
 105 also adopt a multimodal alignment pretraining framework. Due to the presence of samples with
 106 highly similar text reports, training noise can arise and hinder representation learning. To mitigate
 107 this issue, ESI incorporates a retrieval-augmented generation pipeline, TolerantECG applies a more
 straightforward feature retrieval method, and ECG-Chat leverages signal interval durations to en-

108 hance report diversity. By aligning ECG signals with semantically enriched reports, both models
 109 demonstrate improved performance on downstream tasks such as arrhythmia diagnosis and ECG-
 110 report retrieval.

111 In contrast to the aforementioned methods, ST-MEM Na et al. (2024) adopts a self-supervised learning
 112 strategy focused on reconstructing masked ECG patches. By training the model to recover the
 113 full 12-lead ECG signal, ST-MEM effectively learns joint spatio-temporal representations. This ap-
 114 proach not only yields strong performance in arrhythmia classification, but also demonstrates adapt-
 115 ability to varying lead configurations—highlighting the spatial awareness embedded in its learned
 116 representations.

117 ECG-FM McKeen et al. (2025) is built upon a transformer-based design. The training process
 118 incorporates domain-specific augmentations tailored to ECG data, along with a dual-objective strat-
 119 egy combining contrastive learning and signal masking. This setup allows the model to capture
 120 rich contextual representations, enhancing its ability to generalize across multiple tasks—including
 121 identifying specific cardiac conditions from ECG readings, detecting reduced left ventricular ejec-
 122 tion fraction, and recognizing abnormal levels of cardiac troponin (cTn).

124 2.2 EXPLAINABLE AI

125 Explainable AI has become a vital area of research, aiming to enhance the transparency and inter-
 126 pretability of complex machine learning models, particularly deep learning systems. This need is
 127 especially pronounced in the medical domain, where trust in model outputs is often limited due to
 128 the lack of clear explanations behind predictions.

129 Among the most widely adopted techniques for visualizing model decision-making is Grad-CAM
 130 Selvaraju et al. (2019), which produces heatmaps that highlight influential input features—such as
 131 image regions—that contribute to a model’s output. Grad-CAM and its variants have been partic-
 132 ularly impactful in computer vision, offering intuitive visual explanations for convolutional neural
 133 networks.

134 For Transformer-based architectures, a complementary technique is attention calibration Lu et al.
 135 (2021); Zhou et al. (2024). This method introduces an auxiliary loss function during training, guided
 136 by predefined attention heatmap labels. By shaping the model’s attention distribution, it helps de-
 137 termine the appropriate level of focus each token should receive, thereby improving interpretability
 138 and aligning attention with human-understandable patterns.

141 3 DATA CURATION

142 To support both the generation of MIMIC-IV-ECG+X and PTB-XL+X with clinical expertise and
 143 the evaluation of final outputs, we collaborated with medical students and doctors.

144 Despite unprecedented advancement in AI for healthcare domain, physicians are cautious about
 145 adopting healthcare AI models that lack explainability because opaque predictions undermine trust
 146 and accountability in clinical decision-making. Without clear reasoning, clinicians cannot assess
 147 whether outputs are based on sound medical evidence or spurious patterns, raising concerns about
 148 patient safety and liability. Additionally, these AI models are usually trained using as a blackbox
 149 with a pre-defined dataset. Since data is required for efficient models, to make them interpretable, we
 150 generate MIMIC-IV-ECG+X and PTB-XL+X that equip with abnormal heatmap region for each
 151 ECG signal. This dataset requires two-stage rule-based procedure: *Abnormal seeking* and *Criteria*
 152 *seeking*

155 3.1 ABNORMAL SEEKING

156 We first identify the locations that represent irregular features. Specifically, we gather all the con-
 157 ditions of abnormal representations across the *Life in the Fastlane* ECG library lit (2008). From
 158 those features, we can find which waves, intervals or segments that need to pay high attention to
 159 and highlight them. To enable this algorithm, the input ECG signal must first be segmented into
 160 three fundamental components: the P wave, QRS complex, and T wave. We achieve this by train-
 161 ing a UNet3+ model Huang et al. (2020) on the LuDB dataset Kalyakulina et al. (2020), following

162

163

Table 1: Abnormal conditions for each component in ECG signal

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

the methodology outlined by Joung et al. (2024). Once segmented, we compute key interval features—namely PR intervals, QT intervals, and ST segments—for each heartbeat. Abnormal regions are then identified based on the criteria specified in Table 1. For QT interval analysis, we apply Bazett’s correction to account for heart rate variability prior to detecting abnormalities:

191

192

$$Q_i T_i^{correct} = \frac{Q_i T_i}{\sqrt{R_{i-1} R_i}},$$

193

194

195

196

197

3.2 CRITERIA SEEKING

198

199

200

201

202

203

204

205

206

207

208

After identifying the initial abnormal locations, we refine them using the CFR module Nguyen et al. (2025), as illustrated in Figure 2. This module provides diagnostic criteria that specify which ECG components contribute to each diagnosis, enabling the removal of redundant or noise-induced false positives. For each diagnosis in the dataset, we retrieve its associated criteria and use the specified abnormal components to extract the relevant segments identified earlier. In cases where certain components were missed during the initial segmentation, we still highlight them if they are referenced in the diagnostic criteria. For further details on the generated heatmap dataset, please refer to Appendix A.

209

210

211

212

213

214

215

4 MODEL ARCHITECTURE

To showcase the advantages of our synthesized X-ECG dataset in arrhythmia classification, anomaly localization, and report generation, we train a foundational model named X-ECG, as illustrated in Figure 3. The model is built upon the CoCa framework Yu et al. (2022), which has demonstrated strong performance on ECG datasets Yu et al. (2024); Zhao et al. (2025). To enrich the model with spatial and temporal context, we incorporate the Spatial-Temporal embedding proposed by Jin et al. (2025), which encodes lead-specific and time-frame information for each token. Our experiments reveal that this embedding not only enhances classification accuracy but also significantly improves the model’s ability to localize abnormal regions within the ECG signal.

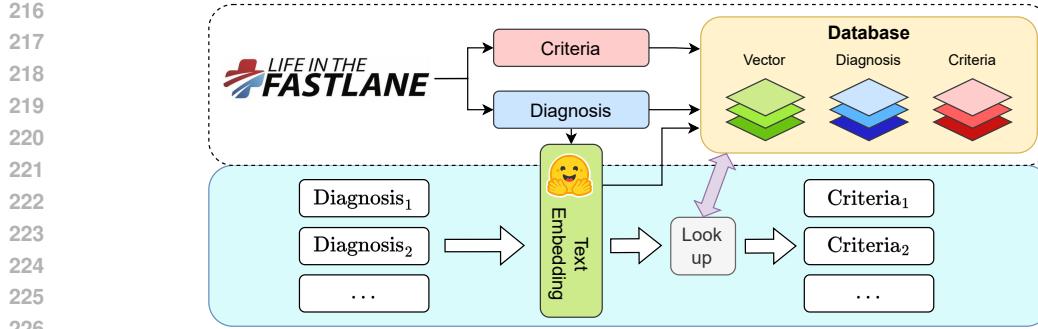


Figure 2: The Criteria Feature Retrieval (CFR) module. It returns diagnostic criteria for each input diagnosis by comparing its embedding with those stored in a pre-constructed vector database.

4.1 ECG SIGNAL AND TEXT REPORT ALIGNMENT

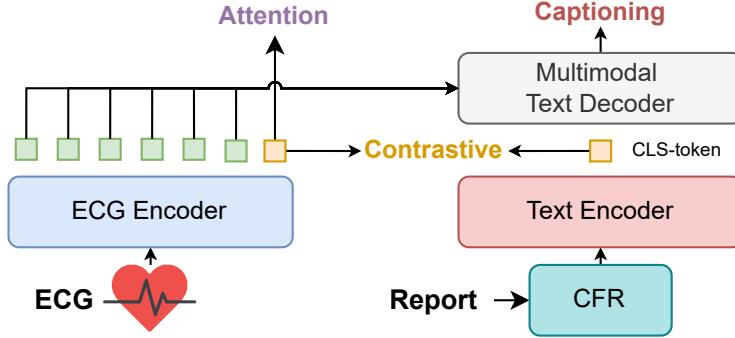


Figure 3: Overview of the X-ECG architecture. It comprises an ECG encoder and a text encoder that process the input ECG waveforms and their corresponding comprehensive augmented reports, respectively. Contrastive learning is performed using the CLS tokens from both encoders to align cross-modal representations. A multimodal text decoder is further employed to reconstruct the input report, conditioned on the ECG token embeddings. To guide focus regions, attention scores between the CLS token and other ECG tokens are optimized using an attention-specific loss function.

Following the previous ECG language pretraining approach Li et al. (2023); Liu et al. (2024); Yu et al. (2024); Nguyen et al. (2025), as shown in Figure 3, we first employ two pretraining objectives for comprehensive learning, including contrastive loss for robust representation learning and captioning loss for semantic alignment. Both encoders aim to project the inputting ECG and text into a unified embedding space. Consequently, the ECG encoder and the Multimodal Text Decoder are jointly optimized by contrasting the paired text against others in the sampled batch, while the Text Encoder is frozen with pretrained weights.

Optimizing the contrastive loss enables alignment between ECG features \mathbf{E} and corresponding text reports \mathbf{T} . This process is illustrated across a batch B of training samples:

$$\mathcal{L}_{contrastive} = -\frac{1}{2} \left(\underbrace{\sum_{i=1}^B \log \frac{\exp(\mathbf{E}_i^\top \mathbf{T}_i / \tau)}{\sum_{j=1}^B \exp(\mathbf{E}_i^\top \mathbf{T}_j / \tau)}}_{\text{ecg-to-text}} + \underbrace{\sum_{i=1}^B \log \frac{\exp(\mathbf{T}_i^\top \mathbf{E}_i / \tau)}{\sum_{j=1}^B \exp(\mathbf{T}_i^\top \mathbf{E}_j / \tau)}}_{\text{text-to-ecg}} \right)$$

Optimizing the captioning loss encourages the ECG features to accurately predict the tokenized text t using an auto-regressive mechanism, conditioned on the ECG tokens e generated by the ECG encoder and the Multimodal Text Decoder parameters θ :

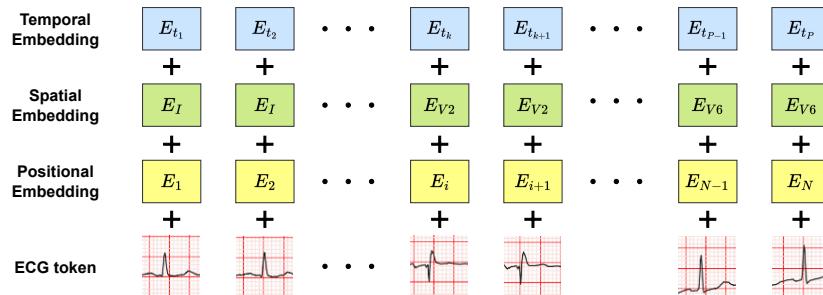
270
271
272
273
274
275
276

$$\mathcal{L}_{captioning} = - \sum_{l=1}^L \log P_{\theta}(t_l | t_{<l}, e)$$

4.2 SPATIAL-TEMPORAL EMBEDDING

In addition to the Transformer’s positional embedding $PE = \{E_1, E_2, \dots, E_N\}$, we incorporate the Spatial-Temporal (ST) embedding proposed by Jin et al. (2025) to enhance the model’s ability to capture lead-specific and temporal information for each ECG token, as illustrated in Figure 4. For spatial embedding, given the input is a 12-lead ECG signal, we define a set $SE = \{E_I, E_{II}, \dots, E_{V6}\}$, assigning a unique embedding to each corresponding lead. For temporal embedding, we use $TE = \{E_{t_1}, E_{t_2}, \dots, E_{t_P}\}$, where P denotes the number of patches obtained after splitting the original 10-second signal.

Figure 4: Prior to being processed by the ECG encoder, each ECG token is enriched with Spatial and Temporal embeddings in addition to standard Positional embeddings. This augmentation provides the model with explicit information about the lead identity and time frame associated with each token, enhancing its ability to capture spatial-temporal patterns in the ECG signal.



4.3 ATTENTION-GUIDING FOR EXPLAINABLE MODEL

To enable the model to learn which regions to attend to, we leverage the Transformer architecture Vaswani et al. (2023) and its inherent attention mechanism. Specifically, we introduce an Attention-Guiding procedure that incorporates an auxiliary loss function targeting the attention scores. To preserve the model’s ability to initially consider all components, the input is processed in the standard manner. However, during training, the attention scores from the final Transformer layer are regularized using a Kullback–Leibler (KL) divergence loss, formulated as follows:

$$\mathcal{L}_{attention} = \frac{1}{B} \sum_{i=1}^B \sum_{n=1}^N p_{i,n} \frac{p_{i,n}}{a_{i,n}^{CLS}},$$

where $a_{i,n}^{CLS}$ denotes the attention score between the CLS token and the n -th token of the i -th sample, while $p_{i,n}$ represents the constructed target attention heatmap for token n of sample i . The variables B , N refer to the batch size and the total number of ECG tokens, respectively. Since the target heatmap highlights anomalous regions, this mechanism guides the model to focus on diagnostically relevant areas, thereby improving its decision-making capability. For samples exhibiting normal heart conditions, where no abnormal heatmap can be constructed, the attention loss $\mathcal{L}_{attention}$ is set to zero. The overall training loss is defined as:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{contrastive} + \beta \mathcal{L}_{captioning} + \gamma \mathcal{L}_{attention}$$

where α , β and σ are hyperparameter weights corresponding to the contrastive, captioning, and attention-guiding loss terms, respectively.

324

5 EXPERIMENT

325

5.1 DATASET

326 To ensure a fair comparison with baseline models, we pretrain our model using the MIMIC-IV-ECG
 327 dataset Gow et al. (2023), which comprises over 800,000 12-lead ECG recordings of 10-second
 328 duration from approximately 160,000 unique patients. During preprocessing, any “NaN” or “Inf”
 329 values in the ECG signals are replaced with zero. For attention-guided supervision, we incorporate
 330 the synthesized MIMIC-IV-ECG+X as an auxiliary training heatmap for X-ECG.
 331

332 For evaluation, we employ the PTB-XL dataset Strodthoff et al. (2020), a clinical 12-lead ECG
 333 dataset containing 21,837 recordings from 18,885 patients. This dataset is used for both arrhythmia
 334 classification and the generation of PTB-XL+X. We benchmark the classification task across four
 335 data variants: Diagnostic, Rhythm, Form, and All. For the report generation task, we adopt the LLM
 336 pretraining, fine-tuning, and benchmarking protocols established by ECG-Chat Zhao et al. (2025).
 337

338

5.2 CONFIGURATION

339 During the pretraining phase, we utilize a 1D 12-layer Vision Transformer with a patch size of 50
 340 as the ECG encoder. The text encoder is initialized with the pretrained MedCPT model Jin et al.
 341 (2023) and remains frozen throughout this process. To ensure that the CLS token attends to all input
 342 tokens, we first flatten the 12-lead ECG signal before feeding it into the encoder.
 343

344 The model is trained with a learning rate of $1e - 4$ over 20 epochs, with the first 10,000 steps
 345 designated for warm-up. Training is conducted using a batch size of 64 across 4× NVIDIA A100
 346 GPUs, each with 80 GB of memory. For the loss weights, we adopt $\alpha = 1$ and $\beta = 2$, following the
 347 configuration in Zhao et al. (2025). The attention-guiding weight γ is set to 0.5 to prevent it from
 348 exerting excessive influence during the training of X-ECG.
 349

350

5.3 METRICS

351 **Arrhythmia Classification.** We adopt the FMax score Strodthoff et al. (2020), Area Under the
 352 Precision-Recall Curve (AUPRC), and Area Under the Receiver Operating Characteristic Curve
 353 (AUROC) as evaluation metrics for this classification task. To ensure balanced assessment across all
 354 diagnostic categories, we report the macro-averaged versions of these metrics, which compute the
 355 mean performance across all classes regardless of class imbalance.
 356

357 **Anomaly localization.** To evaluate whether the attention scores of X-ECG accurately highlight rel-
 358 evant regions within each data sample, we frame this as a binary segmentation task. For assessment,
 359 we employ instance-averaging metrics including AUPRC and AUROC.
 360

361 **Report generation.** We evaluate the output of the LLM conditioned on ECG representations using
 362 BLEU-2, BLEU-4, ROUGE-L, and METEOR to assess lexical similarity between the generated and
 363 reference reports—both at the word level and across contiguous word sequences. Additionally, we
 364 employ BERT-score to assess semantic similarity, capturing how closely the generated report aligns
 365 with the meaning of the reference report.
 366

367

6 RESULTS AND ANALYSIS

368

6.1 ARRHYTHMIA CLASSIFICATION

369 Upon completing the pretraining phase of X-ECG, we attach a linear classification head to the ar-
 370 chitecture. This layer utilizes the CLS token embedding as input. To demonstrate the quality of
 371 the learned representations, we freeze the ECG encoder and train only the linear head. For base-
 372 line comparisons, we utilize the official pretrained weights and apply the same procedure. Table 2
 373 presents the classification results across multiple PTB-XL tasks Strodthoff et al. (2020).
 374

375 Overall, X-ECG significantly outperforms all baseline methods across the four classification
 376 tasks—Diagnostic, Rhythm, Form, and All—on every evaluation metric. This highlights the model’s
 377 strong ability to capture clinically relevant features from ECG input signals, resulting in more ac-

378
 379 Table 2: Linear probing classification result evaluated on PTB-XL dataset. **Bold** indicates the best
 380 result, and underline indicates the second-best result

382 Task	383 Methods	384 Explainability	385 FMax	386 AUPRC	387 AUROC
388 <i>Diagnostic</i>	389 ST-MEM Na et al. (2024)	390 \times	391 27.24	392 21.55	393 86.20
	394 MERL _{Resnet} Liu et al. (2024)	395 \times	396 25.88	397 19.46	398 82.92
	399 MERL _{ViT} Liu et al. (2024)	400 \times	401 29.99	402 <u>24.64</u>	403 86.43
	404 ECG-FM McKeen et al. (2025)	405 \times	406 <u>30.42</u>	407 24.58	408 <u>88.07</u>
	409 X-ECG (Ours)	410 \checkmark	411 41.34	412 34.72	413 92.01
388 <i>Rhythm</i>	389 ST-MEM Na et al. (2024)	390 \times	391 56.64	392 49.18	393 95.32
	394 MERL _{Resnet} Liu et al. (2024)	395 \times	396 40.87	397 35.43	398 85.61
	399 MERL _{ViT} Liu et al. (2024)	400 \times	401 27.98	402 21.43	403 74.44
	404 ECG-FM McKeen et al. (2025)	405 \times	406 52.97	407 46.11	408 85.37
	409 X-ECG (Ours)	410 \checkmark	411 60.72	412 55.46	413 95.88
388 <i>Form</i>	389 ST-MEM Na et al. (2024)	390 \times	391 <u>33.21</u>	392 <u>27.10</u>	393 <u>81.57</u>
	394 MERL _{Resnet} Liu et al. (2024)	395 \times	396 23.47	397 17.61	398 68.79
	399 MERL _{ViT} Liu et al. (2024)	400 \times	401 27.98	402 21.43	403 74.44
	404 ECG-FM McKeen et al. (2025)	405 \times	406 27.83	407 20.57	408 78.74
	409 X-ECG (Ours)	410 \checkmark	411 38.85	412 31.87	413 86.24
388 <i>All</i>	389 ST-MEM Na et al. (2024)	390 \times	391 30.79	392 <u>24.89</u>	393 <u>87.16</u>
	394 MERL _{Resnet} Liu et al. (2024)	395 \times	396 26.01	397 19.61	398 81.01
	399 MERL _{ViT} Liu et al. (2024)	400 \times	401 29.58	402 24.02	403 83.87
	404 ECG-FM McKeen et al. (2025)	405 \times	406 <u>31.07</u>	407 24.26	408 85.20
	409 X-ECG (Ours)	410 \checkmark	411 41.68	412 34.76	413 91.50

405 Table 3: Anomaly Localization result evaluated on PTB-XL+X. **Bold** indicates the best result, and
 406 underline indicates the second-best result

407 Methods	408 AUPRC	409 AUROC
410 ECGAD Jiang et al. (2023a)	411 <u>33.59</u>	412 <u>65.56</u>
413 ECG-Chat Zhao et al. (2025)	414 <u>18.37</u>	415 <u>50.57</u>
416 X-ECG (Ours)	417 43.78	418 80.05

419 curate diagnostic predictions. Notably, X-ECG is the only approach that explicitly incorporates
 420 explainability, a crucial attribute for real-world clinical deployment. These results validate the effec-
 421 tiveness of the proposed model architecture and training strategy, particularly in leveraging attention
 422 mechanisms and pretrained ECG representations.

423 6.2 ANOMALY LOCALIZATION

424 For the evaluation of our model attention score, to see what components contribute to the final
 425 decision output, we perform inference on the PTB-XL test set and benchmark it as a binary segmen-
 426 tation task, as shown in Table 3. To obtain the ground truth label for comparison, we use the same
 427 two-stage mechanism as before to generate the desired abnormal criteria heatmap in PTB-XL+X.

428 The attention scores produced by ECG-Chat exhibit limited reliability, as indicated by an AUROC
 429 value close to 50%, suggesting that its focus regions are nearly random. In contrast, our model,
 430 X-ECG, achieves the highest scores in both AUPRC and AUROC—even without exposure to any
 431 sample signals from the PTB-XL dataset—demonstrating strong generalization and robust represen-
 432 tation learning. Interestingly, although ECGAD is specifically designed for anomaly localization,
 433 it only achieves the second-best performance on both metrics. As shown in Table 2 and Table 3,
 434 X-ECG is capable of performing both arrhythmia classification and anomaly localization, whereas
 435 ECGAD is limited to identifying abnormal regions without the ability to classify the underlying
 436 diagnosis.

432

433

434

435

436

437

438

439

440

441

Table 4: Report generation evaluation using the English-translated PTB-XL report as label

Methods	BLEU-2	BLEU-4	ROUGE-L	METEOR	BERT-score
ECG-Chat Zhao et al. (2025)	21.18	11.96	33.83	32.97	89.00
X-ECG (Ours)	21.46	12.28	34.19	33.07	89.08

6.3 REPORT GENERATION

We extend the LLaVA framework Liu et al. (2023), which was originally proposed as an end-to-end trained large multimodal model aligning a vision encoder with an LLM for joint visual-language understanding, to the ECG domain by connecting a pretrained ECG encoder with an LLM for automated report generation. To enable cross-modal alignment, a learnable projection layer is used to transform ECG embeddings into the text embedding space of the LLM. During this stage, both the ECG encoder and the LLM are kept frozen to preserve their pretrained knowledge, while the projection layer is optimized to establish an effective mapping between modalities. Once the projection layer is trained, we proceed with continual fine-tuning of the entire pipeline. In this stage, the ECG encoder remains unchanged, while the feed-forward layers of the LLM are updated using the LoRA framework Hu et al. (2021), and the projection layer is updated concurrently.

Table 4 shows a comparison of report generation quality between our model and the baseline Zhao et al. (2025), measured against the ground truth report. Our model consistently outperforms the baseline across all metrics, including BLEU-2, BLEU-4, ROUGE-L, METEOR, and BERT-score. These improvements reflect enhanced lexical precision, structural coherence, and semantic fidelity in the generated reports.

7 CONCLUSION

This paper presents X-ECG, the first *explainable* ECG foundation model along with two abnormal heatmap locations dataset MIMIC-IV-ECG+X and PTB-XL+X. With careful design of the architecture and loss function, X-ECG not only outperform all baselines that don't have explainability in both arrhythmia classification and report generation task across various metrics. The explainability of our model is represented by a heatmap that provides detailed, wave-level segmentation of anomalies in the ECG waveforms. Our generated heatmaps are consistent with manual inspection by a licensed cardiac specialist. The contribution of our paper also includes the release of wave-level annotations of anomalies in ECG waveforms on the largest published ECG database, facilitating further development of explainability paradigm in ECG research.

LIMITATION

The data curation process for MIMIC-IV-ECG+X and PTB-XL+X relies on external clinical knowledge to identify abnormal conditions. However, this approach may overlook certain cases due to dependencies on patient-specific factors such as gender and age. As illustrated in Figure 6, nearly half of the heatmaps in the test set are only partially correct, indicating limitations in the current annotation strategy. To improve the accuracy and reliability of abnormal region identification, a more comprehensive algorithm could be adopted for enhanced heatmap generation. Additionally, the attention-guiding mechanism currently utilizes KL divergence as its loss function. However, alternative formulations such as cross-entropy could be employed to enforce stricter alignment between predicted attention distributions and ground truth.

480

481

482

483

484

485

486 REFERENCES
487488 Life in the fast lane, 2008. URL <https://litfl.com/>.489 Alessandro Bortolotti, Fabrizio Ricci, Carmelita Cieri, Federica Cocco, Chiara Martini, Mar-
490 cello Panunzi, Davide Rossi, Anna Sorella, Silvio Sarullo, Davide Scordo, Giulia Renda,
491 Sabina Gallina, and Riccardo Palumbo. Tracing visual expertise in ecg interpretation: An
492 eye-tracking pilot study. *Annals of Noninvasive Electrocardiology*, 30(3):e70082, 2025. doi:
493 <https://doi.org/10.1111/anec.70082>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/anec.70082> ANEC-25-5171.R1.495 Nhat-Tan Bui, Dinh-Hieu Hoang, Thinh Phan, Minh-Triet Tran, Brijesh Patel, Donald Adjeroh,
496 and Ngan Le. Tsrnet: Simple framework for real-time ecg anomaly detection with multimodal
497 time and spectrogram restoration network, 2024. URL <https://arxiv.org/abs/2312.10187>.498 Brian Gow, Tom Pollard, Larry A. Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fer-
500 nandes, Nathaniel Greenbaum, Jonathan W. Waks, Parastou Eslami, Tanner Carbonati, Ashish
501 Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven
502 Horng. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset (version 1.0), 2023. URL
503 <https://doi.org/10.13026/4nqg-sb35>.504 Robert Herman, Anthony Demolder, Boris Vavrik, Michal Martonak, Vladimir Boza, Viera Kres-
505 nakova, Andrej Iring, Timotej Palus, Jakub Bahyl, Olivier Nelis, Monika Beles, Davide Fab-
506 bricatore, Leor Perl, Jozef Bartunek, and Robert Hatala. Validation of an automated artificial
507 intelligence system for 12-lead ecg interpretation. *Journal of Electrocardiology*, 82:147–154,
508 2024. ISSN 0022-0736. doi: <https://doi.org/10.1016/j.jelectrocard.2023.12.009>. URL <https://www.sciencedirect.com/science/article/pii/S0022073623002698>.509 Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,
510 and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.511 Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua
512 Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image
513 segmentation, 2020. URL <https://arxiv.org/abs/2004.08790>.514 Aofan Jiang, Chaoqin Huang, Qing Cao, Shuang Wu, Zi Zeng, Kang Chen, Ya Zhang, and Yanfeng
515 Wang. Multi-scale cross-restoration framework for electrocardiogram anomaly detection, 2023a.
516 URL <https://arxiv.org/abs/2308.01639>.517 Aofan Jiang, Chaoqin Huang, Qing Cao, Shuang Wu, Zi Zeng, Kang Chen, Ya Zhang, and Yanfeng
518 Wang. Multi-scale cross-restoration framework for electrocardiogram anomaly detection, 2023b.
519 URL <https://arxiv.org/abs/2308.01639>.520 Aofan Jiang, Chaoqin Huang, Qing Cao, Yuchen Xu, Zi Zeng, Kang Chen, Ya Zhang, and Yanfeng
521 Wang. Anomaly detection in electrocardiograms: Advancing clinical diagnosis through self-
522 supervised learning, 2024. URL <https://arxiv.org/abs/2404.04935>.523 Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart:
524 Learning ecg words and sentences via pre-training ecg language model, 2025. URL <https://arxiv.org/abs/2502.10707>.525 Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong
526 Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-
527 shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.528 Chankyu Joung, Mijin Kim, Taejin Paik, Seong-Ho Kong, Seung-Young Oh, Won Kyeong Jeon, Jae-
529 Hu Jeon, Joong-Sik Hong, Wan-Joong Kim, Woong Kook, Myung-Jin Cha, and Otto van Koert.
530 Deep learning based ecg segmentation for delineation of diverse arrhythmias. *PLOS ONE*, 19
531 (6):e0303178, 2024. doi: 10.1371/journal.pone.0303178. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0303178>.

540 Alena I. Kalyakulina, Igor I. Yusipov, Viktor A. Moskalenko, Alexander V. Nikolskiy, Konstantin A.
 541 Kosonogov, Grigory V. Osipov, Nikolai Yu. Zolotykh, and Mikhail V. Ivanchenko. Ludb: A
 542 new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access*, 8:
 543 186181–186190, 2020. doi: 10.1109/ACCESS.2020.3029211.

544 Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj. Real-time patient-specific ecg classification
 545 by 1-d convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 63(3):
 546 664–675, 2016. doi: 10.1109/TBME.2015.2468589.

547 Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps
 548 ecg zero-shot learning, 2023. URL <https://arxiv.org/abs/2303.12311>.

549 Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot
 550 ecg classification with multimodal learning and test-time clinical knowledge enhancement, 2024.
 551 URL <https://arxiv.org/abs/2403.06659>.

552 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL
 553 <https://arxiv.org/abs/2304.08485>.

554 Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. Attention calibration for transformer in
 555 neural machine translation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.),
 556 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the
 557 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*,
 558 pp. 1288–1298, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/
 559 v1/2021.acl-long.103. URL <https://aclanthology.org/2021.acl-long.103/>.

560 Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An
 561 open electrocardiogram foundation model, 2025. URL <https://arxiv.org/abs/2408.05178>.

562 Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation
 563 learning to capture spatio-temporal relationship of electrocardiogram, 2024. URL <https://arxiv.org/abs/2402.09450>.

564 Huynh Dang Nguyen, Trong-Thang Pham, Ngan Le, and Van Nguyen. Tolerantecg: A founda-
 565 tion model for imperfect electrocardiogram, 2025. URL <https://arxiv.org/abs/2507.09887>.

566 Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh,
 567 and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based lo-
 568 calization. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN
 569 1573-1405. doi: 10.1007/s11263-019-01228-7. URL <http://dx.doi.org/10.1007/s11263-019-01228-7>.

570 Junho Song, Jong-Hwan Jang, DongGyun Hong, Joon myoung Kwon, and Yong-Yeon Jo. Crema:
 571 A contrastive regularized masked autoencoder for robust ecg diagnostics across clinical domains,
 572 2025. URL <https://arxiv.org/abs/2407.07110>.

573 Nils Strothoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg
 574 analysis: Benchmarks and insights from ptb-xl, 2020. URL <https://arxiv.org/abs/2004.13701>.

575 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,
 576 Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL <https://arxiv.org/abs/1706.03762>.

577 Jikuo Wang, Xu Qiao, Changchun Liu, Xinpei Wang, YuanYuan Liu, Lianke Yao, and Huan
 578 Zhang. Automated ecg classification using a non-local convolutional block attention module.
 579 *Computer Methods and Programs in Biomedicine*, 203:106006, 2021. ISSN 0169-2607. doi:
 580 <https://doi.org/10.1016/j.cmpb.2021.106006>. URL <https://www.sciencedirect.com/science/article/pii/S016926072100081X>.

594 Yue Wang, Xu Cao, Yaojun Hu, Haochao Ying, Hongxia Xu, Ruijia Wu, James Matthew Rehg,
595 Jimeng Sun, Jian Wu, and Jintai Chen. Anyecg: Foundational models for multitask cardiac
596 analysis in real-world settings, 2025. URL <https://arxiv.org/abs/2411.17711>.

597 Han Yu, Peikun Guo, and Akane Sano. Ecg semantic integrator (esi): A foundation ecg model
598 pretrained with llm-enhanced cardiological text, 2024. URL <https://arxiv.org/abs/2405.19366>.

600 Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu.
601 Coca: Contrastive captioners are image-text foundation models, 2022. URL <https://arxiv.org/abs/2205.01917>.

602 Yubao Zhao, Jiaju Kang, Tian Zhang, Puyu Han, and Tong Chen. Ecg-chat: A large ecg-language
603 model for cardiac disease diagnosis, 2025. URL <https://arxiv.org/abs/2408.08849>.

604 Peilin Zhou, Qichen Ye, Yueqi Xie, Jingqi Gao, Shoujin Wang, Jae Boum Kim, Chenyu You, and
605 Sunghun Kim. Attention calibration for transformer-based sequential recommendation, 2024.
606 URL <https://arxiv.org/abs/2308.09419>.

607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647

A DEMONSTRATION OF PTB-XL+X

Figure 5 showcases representative examples of the generated heatmaps alongside their corresponding original text reports from PTB-XL+X.

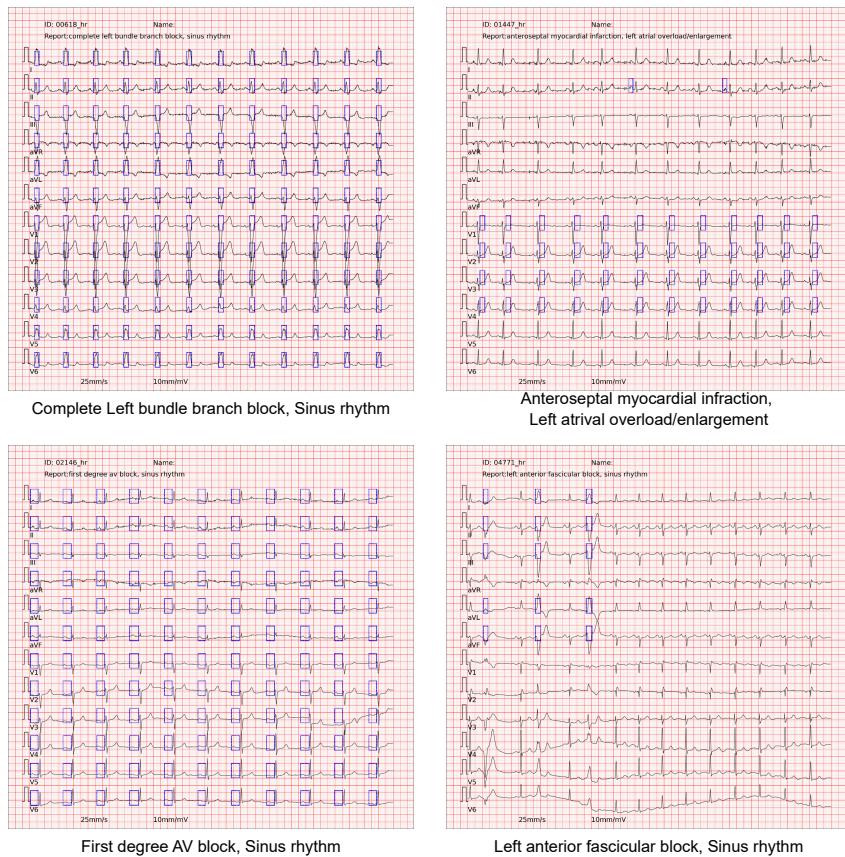


Figure 5: Example outputs from the two-stage rule-based procedure applied to the PTB-XL dataset. Blue boxes highlight the label regions of detected anomalies identified by our algorithm.

To statistically evaluate the quality of our dataset, we submitted 100 samples to cardiologists for expert review. The assessment focused on verifying the accuracy of the abnormal heatmap annotations. The results of this evaluation are presented in Figure 6.

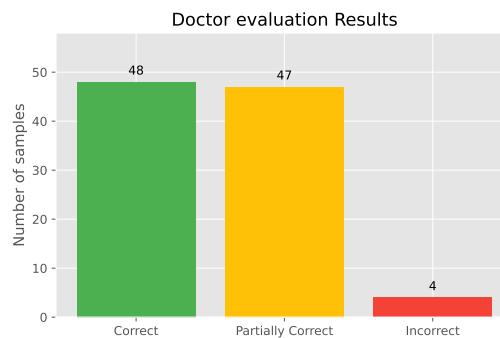


Figure 6: Evaluation results of abnormal heatmaps from 100 samples in PTB-XL+X, reviewed by a clinical expert

702 **B DETAILS OF LARGE LANGUAGE MODELS USAGE**
703704 The use of Large Language Models (LLMs) in this work is limited solely to grammar correction and
705 stylistic refinement. All core aspects—including the formulation of main contributions, experimen-
706 tal design, and data analysis—were conducted independently without LLM involvement.
707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755