CARP: A FRAMEWORK FOR CLINICALLY-ALIGNED ECG GENERATION WITH WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electrocardiograms (ECGs) are recordings of the heart's electrical activity over time. Text-to-ECG generation enables critical applications in rare-label data synthesis, robust model evaluation, and complex disease simulation. This generation problem is challenging since ECGs associated with clinically-distinct text largely share the same waveform structure and differ only in subtle morphological features and measurements. Recent diffusion-based generators produce plausible ECGs that are only partially clinically-aligned with the report, missing subtle morphologies and implied findings, particularly for rare conditions. To address these limitations, we introduce CARP, which adapts a pretrained world model to text-to-ECG generation and augments rectified-flow training with direct, sample-specific supervision over morphological measurements. Because these measurements are typically computed via non-differentiable waveform segmentation, we develop a novel differentiable soft-segmentation objective that aligns generated and target ECGs and enables training. Across MIMIC-IV-ECG and the external datasets PTB-XL and HEEDB, CARP has the lowest error on 21/24 clinical measurements. Ablations show that the world model pretraining is essential: without it, real-to-synthetic AUPRC falls 55.2%. Furthermore, morphology supervision lowers all nine measurement errors and nearly halves FID (6.6 to 3.5). Real-to-synthetic macro-AUPRC on CARP's ECGs reaches 80.3%, 24.0% above the best prior generator, and augmenting rare-condition classifiers with CARP ECGs improves AUPRC over real-only training by 10.7% in-domain and by 20.1% and 15.3% on the out-of-domain datasets.

1 INTRODUCTION

Electrocardiograms (ECGs) are multi-channel recordings of cardiac activity. As a low-cost, non-invasive, and widely available modality, the ECG has become a critical tool for triaging and diagnosing cardiac diseases across wearables, cardiology clinics, and hospital ICUs. Consequently, ECG classifiers are being developed to enable early disease detection and improve patient outcomes (Attia et al., 2019; Poterucha et al., 2025).

However, training robust ECG classifiers for disease diagnosis is limited by data scarcity. ECGs are considered protected health information (PHI) and difficult to share publicly, and this scarcity compounds for populations already underrepresented in clinical data, such as patients with rare or congenital conditions (Saleh et al., 2026). For instance, the Wolff-Parkinson-White syndrome appears in just 0.08% of MIMIC-IV-ECG recordings (Appendix C.4.1). Generating synthetic ECGs from clinical text addresses these gaps by improving rare-disease classification, with further applications in pre-deployment evaluation and multi-disease simulation for education (Jin et al., 2026).

Despite its promise, generating *clinically-aligned* ECGs, i.e., ECGs consistent with the diagnoses in the conditioning text, is a challenging problem, since ECGs with clinically distinct text often share most of their waveform and differ only in subtle morphology. For instance, as shown in the first row of Fig. 1, the right bundle branch block (RBBB) condition can be distinguished only by a small notch in the QRS complex in what would otherwise be considered a healthy ECG. In order to quantitatively describe these subtle differences, medical practitioners are trained to understand various measurements of the rhythm (timing), morphology (shape), and axis (direction) of an ECG.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

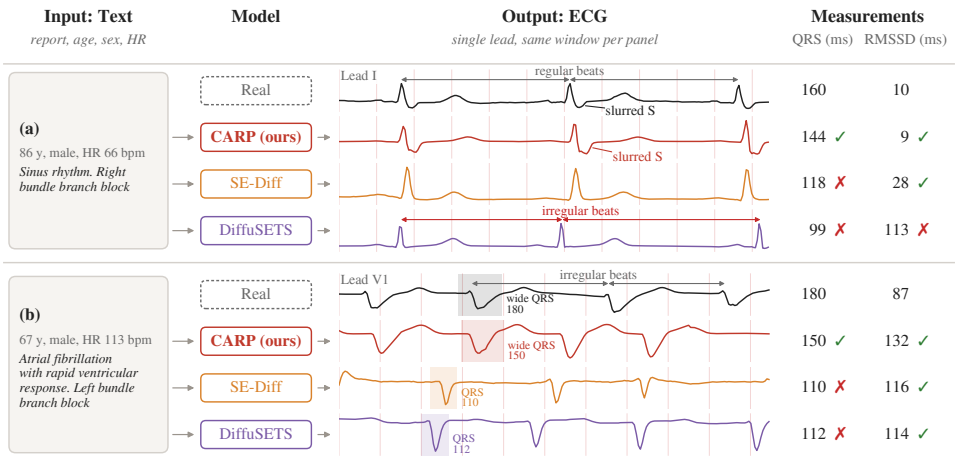


Figure 1: **CARP generates ECGs that better match the diagnoses in the conditioning text than baseline generators.** All models produce 10-second, 12-lead ECG, but we show the leads relevant to the abnormality for convenience. **(a)** CARP ECG has a slurred terminal S wave and a widened QRS (144 vs. 160 ms in the real ECG), whereas SE-Diff and DiffuSETS produce narrow QRS complexes (118 and 99 ms). **(b)** All three generators correctly produce the irregular rhythm for atrial fibrillation, but only CARP's produced ECG has the correct wide QRS pattern. We show in §C.4 that CARP produces clinically-aligned ECGs across both common and rare diagnoses.

Diffusion models have driven recent progress in text-to-ECG generation. Building on their success in text-to-image and text-to-video synthesis, generators such as DiffuSETS (Lai et al., 2025) and SE-Diff (Wang et al., 2026) learn to denoise ECGs conditioned on clinical text, producing realistic waveforms that capture the overall structure of the heartbeat. Nevertheless, these generators often produce ECGs that are misaligned with the conditioning text: the waveform looks plausible but contradicts the report on precisely the subtle features that define the diagnosis. For instance, for the RBBB report in Fig. 1a, the real ECG has a QRS duration of 160 ms, whereas SE-Diff and DiffuSETS produce narrow QRS complexes of 118 and 99 ms, below the 120 ms clinical threshold for RBBB. Similarly, in Fig. 1b, both generators reproduce the irregular rhythm of atrial fibrillation but miss the wide QRS pattern stated in the report. This misalignment has practical consequences: as shown in Section 3, ECGs generated by these methods struggle to be effective as training data for classifiers for rare or multiple labels.

We identify two key limitations of diffusion-based generators that underlie this misalignment. First, they are trained from scratch on text-ECG pairs, and must therefore learn both generic signal structure and its subtle text-dependent variations from limited paired data. In contrast, in other modalities, large-scale pretrained generators provide transferable generative priors. In particular, world models such as Cosmos 3 (NVIDIA et al., 2026) are pretrained as many-to-many generators across heterogeneous modalities, enabling mappings such as text-to-video and image-to-video with a single shared backbone. Second, these generators do not directly supervise the morphological measurements specific to each text-ECG sample. Their denoising objectives, even when augmented with generic physiological priors as in SE-Diff, do not target the specific measurements, such as QRS duration, that distinguish one diagnosis from another.

In this work, we introduce CARP: Clinically-Aligned ECG Rectified Flow with Pretrained World Priors, which addresses both limitations. First, rather than training from scratch, CARP adapts a pretrained world model to text-to-ECG generation for the first time, transferring its generative prior to a physiological modality it was never trained on. Second, CARP augments rectified-flow training with direct, sample-specific supervision over morphological measurements. Because these measurements are typically computed by segmenting the waveform into its constituent waves, a non-differentiable operation, we develop a novel differentiable soft-segmentation objective that aligns the measurements of generated and target ECGs.

Extensive clinically-oriented evaluations across three datasets show that CARP consistently outperforms existing ECG generators, both in-domain (MIMIC-IV-ECG) and out-of-domain (PTB-XL,

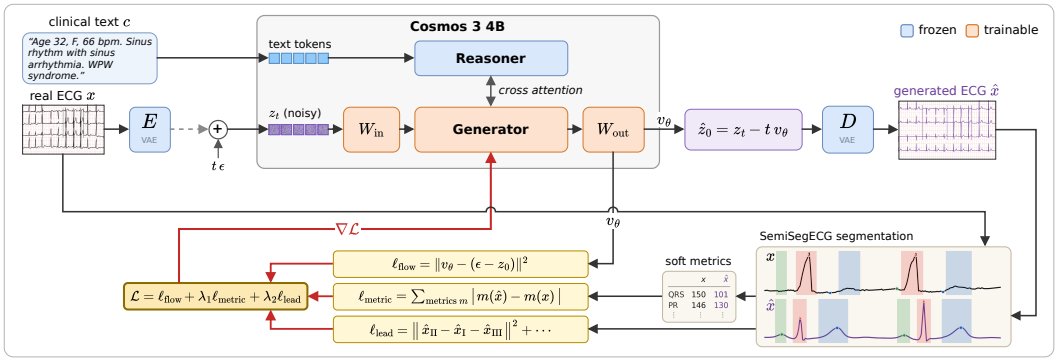


Figure 2: **Overview of CARP's training pipeline.** We encode the clinical report and patient metadata with the world model reasoner; the attached generator predicts the velocity v_θ for the noisy ECG latent z_t. During training, z_0 is produced from real ECG x and noised to yield z_t, which the generator is trained to denoise. To clinically align the generator, we delineate the P, QRS, and T waves of $\hat{x}$ and x, and compare their clinical measurements (ℓ_{metric}) and lead-wise identities (ℓ_{lead}).

HEEDB). CARP-generated ECGs better augment classifiers across 14 diagnostic label families, improving average AUPRC over real-only training by 14.2 points in-domain, and improve AUPRC on rare conditions by 10.7 points in-domain and by 20.1 and 15.3 points on PTB-XL and HEEDB, where existing generators underperform real-only training. CARP also better follows what clinical reports state and imply: real-to-synthetic macro-AUPRC on its ECGs reaches 80.3%, 24.0 points above the best prior generator, and its ECGs match the cardiac axis implied by a diagnosis in 92.2% of reports, against 59.5% for the best prior generator. Finally, we show that the pretrained world-model generator is necessary: without it, real-to-synthetic AUPRC falls from 73.0% to 17.8%.

In summary, we make the following contributions: **First,** we present CARP, the first framework to adapt a large-scale, multimodal world model prior (Cosmos 3) to text-to-ECG generation, demonstrating that pretrained world priors can successfully transfer to text-to-ECG synthesis. **Second,** we introduce a novel, end-to-end differentiable loss for fine-grained ECG morphological measurements directly during rectified-flow training. **Third,** we design a comprehensive, clinically-oriented framework for evaluating text-to-ECG models, which tests whether synthetic ECGs improve classification across 14 diagnostic label families and five rare conditions, both in-domain (MIMIC-IV-ECG) and out-of-domain (PTB-XL, HEEDB), and how faithful they are to their reports, using real-to-synthetic and synthetic-to-real classification, clinical measurements, and report-alignment case studies.

2 Method

This section describes how CARP adapts a pretrained world model to generate ECGs and trains it with a clinical measurements loss. Following Wang et al. (2026), we formalize text-to-ECG generation as follows. For recording i, let $x_i \in \mathbb{R}^{m \times n}$ be the real 10-second, 12-lead ECG, r_i its clinical report, and m_i its metadata (age, sex, and heart rate). We form the conditioning text $c_i = (r_i, m_i)$ and learn $p_\theta(x \mid c_i)$ to generate $\hat{x}_i \sim p_\theta(\cdot \mid c_i)$.

2.1 Architectural adaptation

In order to adapt world models for generating ECGs, we begin by training a VAE (E_ϕ, D_ψ) that can encode and decode ECG segments. Let each ECG be divided into K equally-sized segments. The CNN-based VAE jointly encodes each segment's leads into L latent tokens of width d_z:

$$x_i = \text{Concat}(x_i^{(1)}, \ldots, x_i^{(K)}), \qquad h_i^{(j)} = E_\phi(x_i^{(j)}) \in \mathbb{R}^{L \times d_z}. \tag{1}$$

The flow runs in standardized ECG-VAE latent space. We standardize the ECG latents along their channels using training-set latent statistics (μ, σ) and then map each d_z-dimensional token to the d_C-dimensional generator tower's representation space using a learned linear projection P_{in}. Let $z_{i,t}$ denote the noisy latent at flow time $t \in [0, 1]$, with noiseless ECG latents at $t = 0$ and added

Gaussian noise at $t = 1$:

$$z_i^{(j)} = \frac{h_i^{(j)} - \mu}{\sigma}, \qquad z_i = \mathrm{Concat}(z_i^{(1)}, \ldots, z_i^{(K)}), \qquad u_{i,t} = P_{\mathrm{in}}(z_{i,t}), \tag{2}$$

where $N = KL$ is the total number of latent tokens, $z_i \in \mathbb{R}^{N \times d_z}$, and $u_{i,t} \in \mathbb{R}^{N \times d_C}$. Let G and R denote Cosmos's generator tower and reasoner tower, respectively. We pass the conditioning text c_i through R and the noised latents z_i through G. At layer ℓ, tokens in $G^{(\ell)}$ jointly attend to each other and the text tokens in $R^{(\ell)}$ through dual-stream joint attention (NVIDIA et al., 2026). At the end of the tower, a learned projection P_{out} maps the resultant output prediction to an ECG latent velocity. The flow solver uses this velocity to update $z_{i,t}$ at each step. At any flow time t, we can estimate the clean latent as $\hat{z}_{i,t} = z_{i,t} - t\, v_\theta(z_{i,t}, t, c_i)$. When necessary, we obtain the generated ECG $\hat{x}_{i,t}$ by reversing latent standardization, decoding each segment latent with the VAE decoder D_ψ, and concatenating the ECG segments:

$$v_\theta(z_{i,t}, t, c_i) = P_{\mathrm{out}}(G(u_{i,t}, t; R(c_i)))\,, \qquad \hat{x}_{i,t} = \mathrm{Concat}\Big(\big[D_\psi(\sigma \odot \hat{z}_{i,t}^{(j)} + \mu)\big]_{j=1}^{K}\Big)\,, \tag{3}$$

where $\hat{z}_{i,t}^{(j)}$ denotes the estimated clean latent tokens for segment j at flow time t. During generation, CARP solves the full flow from Gaussian noise to obtain $\hat{z}_i$ and decodes it as above to obtain $\hat{x}_i$.

2.2 Clinical metrics loss

Fig. 1 and Fig. 6 show that physiological measurements used by cardiologists can be effective evaluators of misalignment. While these measurements are simple, their definitions often rely on non-differentiable operations whose gradients cannot be backpropagated. For example, the RR-interval between two consecutive heartbeats is $\mathrm{RR} = s_2 - s_1$, where `argmax` locates the peak samples s_1 and s_2. Since these are discrete indices, RR provides no gradient for backpropagation.

To address this issue, we use SemiSegECG (Park et al., 2025), an ECG segmentation model that labels each time sample as P wave, QRS complex, T wave, or background, and processes each lead independently. For one lead, let $p_a(s)$ be the probability that sample s belongs to wave $a \in \{\mathrm{P}, \mathrm{QRS}, \mathrm{T}\}$. These probabilities are differentiable but do not directly yield measurements. A measurement such as the QT interval requires locating wave boundaries, pairing them within the same beat, and comparing beats between two ECGs that may differ by heart rate or phase shifts. We replace each non-differentiable step with a differentiable counterpart computed from $p_a(s)$, and compare the distributions of the resulting measurements between x and $\hat{x}_{i,t}$.

Timing measurements. Timing measurements capture how electrical impulses are generated and conducted through the heart, such as the RR interval between beats and the P, QRS, PR, and QT durations within a beat. To compute them, we require *soft* onset and offsets for each of P, Q, R, S and T waves:

$$\mathrm{on}_a(s) = p_a(s)\big(1 - p_a(s-1)\big), \qquad \mathrm{off}_a(s) = p_a(s)\big(1 - p_a(s+1)\big), \qquad a \in \{\mathrm{P}, \mathrm{QRS}, \mathrm{T}\}, \tag{4}$$

where $\mathrm{on}_a(s)$ is the probability that wave a begins at sample s, and $\mathrm{off}_a(s)$ is the probability that a ends at s. Unlike an argmax, these peaks move continuously with the boundaries, so even a slightly misplaced onset or offset receives a gradient. However, a generated ECG differs in beat count and alignment and may lack some P or T waves, so their measurements cannot be matched beat by beat.

To address this, we compare histograms of each measurement instead of comparing beat by beat. To build the histogram of a timing measurement, which runs from a start boundary b_1 to a later end boundary b_2 of the same beat (e.g., $\mathrm{on}_{\mathrm{QRS}}$ to $\mathrm{off}_{\mathrm{QRS}}$), we weight each pair of samples $(s, s+k)$ by $b_1(s), b_2(s+k)$ and sum over s:

$$H(k) = \sum_s b_1(s)\, b_2(s+k)\, e^{-\alpha\, n(s, s+k)}, \qquad n(s, s') = \sum_{s<u<s'} \mathrm{on}_{\mathrm{QRS}}(u),\ \alpha = 2. \tag{5}$$

Amplitude measurements. Amplitude measurements capture the strength and direction of the heart's electrical activity in each lead, such as R-wave height, S-wave depth, and T-wave inversion. To compute them, we first estimate the baseline from samples labeled as background (p_{bg}) and subtract it to obtain $\tilde{x}$. As with wave boundaries, we replace the extreme voltage of each wave with

a soft version, a weighted average over the wave $\mathcal{W}(s) = [s, \min\{s' \geq s : \hat{y}(s') = a \neq \hat{y}(s'+1)\}]$ that begins at sample s, with detached argmax labels $\hat{y}$:

$$A(s) = \sum_{s' \in \mathcal{W}(s)} w(s')\, \tilde{x}(s'), \qquad w(s') \propto p_a(s')^2\, e^{\tilde{x}(s')/\tau}, \tag{6}$$

where the weights favor high voltages and samples likely to belong to wave a, and the temperature τ sets how sharply they concentrate. To measure downward deflections such as S-wave depth or T-wave inversion, we negate $\tilde{x}$ in the exponent, which turns $A(s)$ into a soft minimum. To compare these amplitudes between ECGs without matching their beats, we again collect them into a histogram over voltage bins. To obtain one amplitude per wave rather than one per sample, we weight each $A(s)$ by the onset probability $\mathrm{on}_a(s)$. We defer computation of other metric types to Appendix C.5.

Comparing measurement distributions. Now that we have differentiable procedures for measuring metrics based on both symbolic procedures and neural classification, we can develop soft histograms for each metric for a given ECG. During training, we produce these histograms for the generated $\hat{x}$ and the real target x. We turn these histograms into cumulative histograms $F_m(b)$, which represents the fraction of measurement m's mass at or below bin b. Let $\hat{F}_m(b)$ and $F_m(b)$ represent $\hat{x}$ and x, respectively, and $\rho(u) = \sqrt{u^2 + \delta^2} - \delta$, which is a smoothing function for the sharp corners of the cumulative histograms. We define the subloss for a metric m in lead l as:

$$\ell_{m,l} = \Delta_m \sum_b \left| \hat{F}_{m,l}(b) - F_{m,l}(b) \right|, \qquad \ell_{\text{metric}} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{1}{|\mathcal{L}_m|} \sum_{l \in \mathcal{L}_m} \ell_{m,l} \tag{7}$$

Finally, we aggregate over leads and metrics to produce ℓ_{metric}. To avoid evaluating CARP with the same segmentation that supervises its training (SemiSegECG), we conduct evaluation of clinical measurements with ECGDeli (Pilia et al., 2021) instead, which is a separate, non-neural framework.

2.3 Combined objective and training

We train using rectified flow in standardized ECG latent space. Letting $\epsilon \sim \mathcal{N}(0, I)$ and $t \in [0, 1]$, we construct $z_{i,t} = (1-t)z_i + t\epsilon$ and train the generator to predict $\epsilon - z_i$ with $\ell_{\text{flow}} = \frac{1}{Nd_z} \|v_\theta(z_{i,t}, t, c_i) - (\epsilon - z_i)\|_F^2$. Let t_c be a threshold denoting low noise. When $t < t_c$, we decode $\hat{x}_{i,t}$ using Eq. (3) and apply the clinical loss, allowing us to enforce physiological supervision on latents within proximity to the target latent. Following Wang et al. (2026), we also apply ℓ_{lead}, which exploits known algebraic relations (Wilson et al., 1946) to ensure consistent beat-timing across leads. Letting the λ coefficients weight the auxiliary losses, $\eta(t) = 1/\max(t, 0.1)$, and $\gamma_B = |B|/|\{i \in B : t_i < t_c\}|$ for batch B, our training objective is therefore:

$$\mathcal{L} = \mathbb{E}_{(x,c),\epsilon,t}[\ell_{\text{flow}} + \lambda_{\text{lead}}\ell_{\text{lead}} + \eta(t)\,(\lambda_{\text{metric}}\gamma_B \mathbf{1}[t < t_c]\ell_{\text{metric}})]\,. \tag{8}$$

3 Experiments

3.1 Experimental setup

Datasets and training. We train CARP on approximately 556,000 MIMIC-IV-ECG recordings (Gow et al., 2023), with disjoint patients for training, validation, and testing. The input for the generation is conditioned on a clinical report, age, sex, and heart rate. Training hyperparameters are specified in Appendix C.2. We use a randomly drawn subset of the MIMIC-IV-ECG test split for evaluation, and also evaluate on two external, out-of-domain datasets: PTB-XL (Wagner et al., 2020) and HEEDB (Koscova et al., 2026). Appendix C.1 provides further cohort details.

Baselines. We evaluate CARP against two current state-of-the-art ECG generators, SE-Diff (Wang et al., 2026) and DiffuSETS (Lai et al., 2025). Both are the only prior methods that generate 10-second, 12-lead ECGs from natural-language clinical reports, and outperform GAN (Adib et al., 2022) and SSM-based generators (Alcaraz & Strodthoff, 2023). We ensured baselines and CARP were trained and tested on the same split (details in Appendix C.3).

Evaluation. We consider three research questions: (1) **Classifier Performance.** Does the synthetic ECG produced by CARP improve classification performance, (a) particularly for rare labels, with

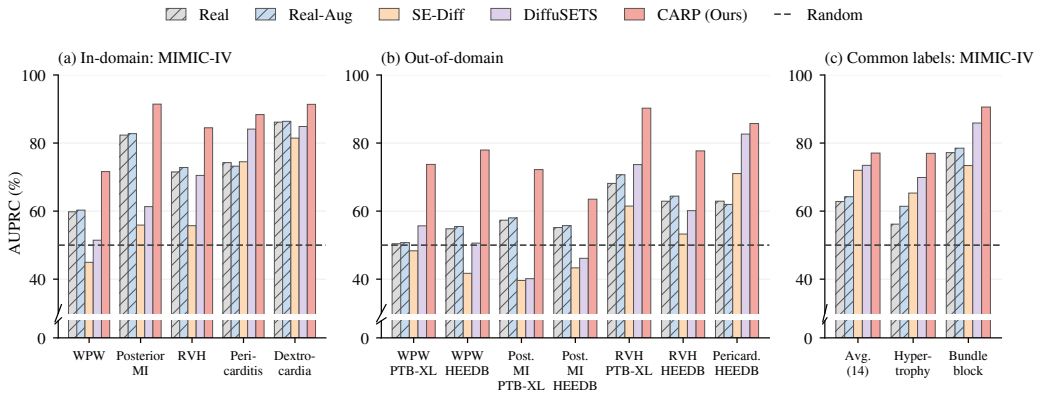


Figure 3: **RQ1: Augmentation with CARP improves downstream classifiers, both in-domain and out-of-domain.** Random chance AUPRC is 50% (dashed line). **(a)** In-domain results on MIMIC-IV-ECG for five rare labels (WPW, posterior MI, RVH, pericarditis, dextrocardia). **(b)** Out-of-domain results on PTB-XL and HEEDB, evaluated without retraining. **(c)** In-domain AUPRC averaged over the 14 diagnostic label families, and for two common labels where baseline generators also improve over Real (all labels and datasets in Appendix Table 5).

(b) generalizability across datasets? (2) **Quality of Synthetic ECG.** Do generated ECGs capture the full diversity of real data, including subtle morphologies and implicit knowledge in clinical text? (3) **Contribution of CARP's components.** Are the pretrained world model and fidelity-based objective necessary for CARP's performance?

3.2 RQ1: Does CARP Improve Downstream Classifiers?

We evaluate the utility of synthetic ECG augmentation on five critical, low-prevalence conditions, such as WPW (0.08%) and pericarditis (0.31%) (Appendix C.4.1), in a low-resource setting with only ten real positive training samples. We compare three training regimes: (a) *Real-only* (no augmentation, imbalanced), (b) *Real-augmented, latently balanced* (Gaussian noise and jitter on real positives), and (c) *synthetic ECG augmentation* using SE-Diff, DiffuSETS, and our CARP model (Appendix C.4.1). Following prior work (Wang et al., 2026), we train a binary classifier per condition using 200 real negatives and 200 positives (10 real, 190 augmented), and we evaluate on a 50/50 balanced test set of real positives and real negatives. As an empirical reference, we also train a setting where 200 real positive and 200 real negative samples are used. All classifiers are trained only on MIMIC-IV and evaluated in-domain on MIMIC, as well as out-of-domain on PTB-XL and HEEDB. Overall, classifiers augmented with CARP-generated ECGs outperform both real-only and baseline-augmented models on most labels, both in-domain and out-of-domain (Fig. 3).

In-Domain Performance. Across the five rare labels, CARP achieves an 85.5% mean AUPRC, substantially outperforming Real (74.8%) and Real-Aug (75.1%) (Fig. 3a). The marginal improvement of Real-Aug over Real highlights that upsampling via Gaussian noise is insufficient. CARP performs best at WPW (71.6% vs. 60.3% for the next-best model), which requires a modified PR-interval and delta-wave morphology. We find that baseline synthetic generators can even *underperform* real-only classifiers, such as on WPW, across datasets. This indicates that morphologically imprecise synthetic samples hurt classification with imbalanced labels.

Common Labels. Like rare labels, CARP outperforms SE-Diff and DiffuSETS on common labels, most consistently on hypertrophy, where it leads the best baseline on all three datasets (by 7.1, 16.0 and 6.3 points on MIMIC-IV, PTB-XL and HEEDB). Across 14 diagnostic label families on MIMIC-IV, SE-Diff (72.0%) and DiffuSETS (73.4%) improve over Real (62.8%) and Real-Aug, with CARP outperforming them (77.1%). In particular, we find best performance on shape-related labels such as bundle blocks; margins are large on hypertrophy (77.0% vs. 69.9% for DiffuSETS and 56.2% for Real) and bundle-branch block (90.6% vs. 85.9% and 77.2%) (Fig. 3c). These results hold out-of-domain on PTB-XL and HEEDB too (Appendix Table 5). We find that these results hold when classifying age and sex from the ECG as well (Fig. 4a, Table 6).

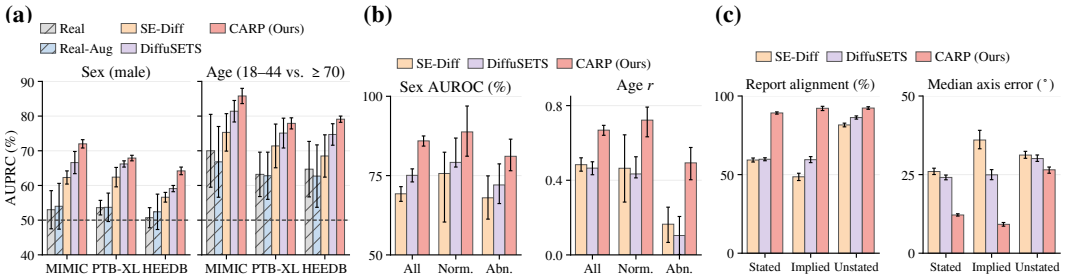


Figure 4: **(a) Results for age- and sex classifiers**, showing that CARP is the best augmenter for under-represented (young, male) patients. **(b) CARP generates age- and sex-consistent ECGs.** We predict sex and age from generated ECGs with models trained on real ECGs; on abnormal reports, which skew older, baselines ignore the input age. **(c) CARP follows stated and implied cardiac axes.** Reports are grouped by whether they state an axis deviation, imply one through a diagnosis, or leave the axis unstated.

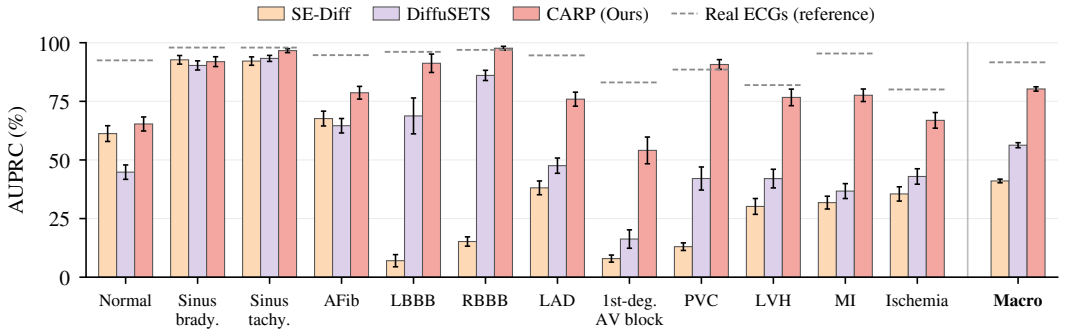


Figure 5: **Real-to-synthetic performance on MIMIC-IV-ECG, showing that CARP ECGs are more clinically aligned to their conditioning reports**. We train a multilabel classifier on real MIMIC-IV-ECG, and evaluate it on test ECGs from each generator. Dashed grey lines indicate test performance on (real ECGs) and serve as upper-bounds. Full results in Table 9.

Out-of-Domain Rare-Label Performance. We first observe that classifiers transfer poorly to out-of-domain datasets (Fig. 3b), a key challenge found in deployed ECG classifiers (Yagi et al., 2022; Perez Alday et al., 2020). Real model AUPRC drops from 74.8% in-domain to 58.6% on PTB-XL and 59.1% on HEEDB, with near-random performance for WPW on PTB-XL (50.4%). The poor generalization is often due to institutional and geographic variation in machine-hardware and patient demographics (Yagi et al., 2022). For both external datasets, standard augmentation (59.8%, 58.0%) and baseline generators (SE-Diff: 49.8%, 49.3%; DiffuSETS: 56.5%, 56.4%) fail to bridge this generalizability gap for rare labels. In contrast, CARP substantially mitigates transfer loss, achieving 78.7% on PTB-XL and 74.4% on HEEDB (+20.1% and +15.3% over Real), with substantial gains on WPW (+23.3% PTB-XL; +23.1% HEEDB).

3.3 RQ2: Assessing CARP ECG Quality

A text-to-ECG generator should produce ECGs whose clinical content matches the clinical text, and whose measurements match those of the corresponding real ECGs. We assess this in two ways: whether a classifier trained on real data classifies generated ECGs correctly (real-to-synthetic), and how closely generated ECGs match real ones on clinical measurements and distributional metrics. We also include a detailed qualitative study on three rare labels (Fig. 6, elaborated in Appendix B).

Real-to-Synthetic Performance. To evaluate the clinical realism of generated ECGs, we follow prior work (Shmelkov et al., 2018; Alcaraz & Strodthoff, 2023) by testing a classifier trained on real data directly on synthetic outputs. Specifically, we train a multilabel classifier on the MIMIC-IV train set ($n = 556$k), which achieves a reference macro-AUPRC of 91.6% on the real test-set ECGs.

Table 1: **CARP has the lowest error on most clinical measurements, and the best FID and recall.** We compute the MAE (↓) of measurements between each generated ECG and its paired real ECG. We use the DiffuSETS VAE embedding for FID, k-NN precision and recall (k=3, %). Full results in Table 10

	MIMIC-IV			PTB-XL (external)		
Metric	SE-Diff	DiffuSETS	CARP	SE-Diff	DiffuSETS	CARP
HR (bpm) ↓	1.7 ±0.2	1.8 ±0.3	**1.1** ±0.2	1.4 ±0.1	7.1 ±0.1	**1.0** ±0.1
PR (ms) ↓	27.7 ±1.7	26.2 ±1.8	**23.5** ±1.7	26.7 ±0.8	23.8 ±0.8	**23.1** ±0.7
QRS (ms) ↓	14.0 ±0.7	11.0 ±0.7	**10.9** ±0.6	12.4 ±0.4	**10.8** ±0.3	11.0 ±0.3
QT (ms) ↓	27.0 ±1.8	25.8 ±2.5	**24.3** ±2.1	24.0 ±0.7	24.5 ±0.7	**23.5** ±0.7
P duration (ms) ↓	13.7 ±0.6	12.4 ±0.5	**11.8** ±0.6	14.4 ±0.4	12.9 ±0.4	**12.3** ±0.3
T duration (ms) ↓	14.8 ±1.0	14.3 ±0.8	**13.4** ±0.9	12.6 ±0.4	12.8 ±0.4	**11.9** ±0.4
ST J+60 (μV) ↓	56.3 ±3.9	54.0 ±4.3	**51.7** ±4.4	52.4 ±1.2	59.6 ±1.5	**52.0** ±1.4
Cardiac axis (°) ↓	39.8 ±2.6	38.1 ±2.1	**30.0** ±1.9	36.6 ±1.1	44.2 ±1.3	**35.4** ±1.1
FID ↓	17.54	6.68	**2.72**	28.47	22.92	**15.93**
Precision ↑	88.0	**90.4**	89.3	45.2	54.4	**57.0**
Recall ↑	80.4	83.9	**90.5**	75.7	85.5	**89.8**

When evaluated on synthetic ECGs, CARP achieves a macro-AUPRC of 80.3%, significantly outperforming DiffuSETS (56.3%) and SE-Diff (41.0%; Fig. 5). CARP outperforms baseline models particularly on structural morphology such as bundle blocks, capturing fine-grained features like RBBB and PVC far more reliably (e.g., 97.7% vs. 86.1% and 15.2% on RBBB; 90.7% vs. 42.1% and 13.0% on PVC). Conversely, all models achieve similar performance on heart rate-related labels such as sinus bradycardia and tachycardia (90–97%), indicating that baselines succeed at modeling overall heart rate but miss subtle morphology.

Does the ECG follow the conditioning age and sex? We predict the age and sex of each generated ECG with a regressor and a classifier trained on real ECGs, and find that CARP's ECGs match the conditioning age and sex closest (Fig. 4b). We find that for abnormal reports the baseline-produced ECGs ignore the age conditioning almost entirely ($r = 0.16$ and 0.10 against 0.49 for CARP).

ECG Metric Performance. We next evaluate fidelity with the clinical measurements that motivate CARP. As argued by prior work (Wang et al., 2026), using pointwise metrics (such as MSE and MAE) to measure clinical alignment is uninformative, since even trivial models such as a flatline and a fixed normal-sinus-rhythm ECG (which ignore the conditioned text) can obtain lower or comparable MAE and RMSE than every generator (Table 13). Instead we evaluate fidelity via a suite of 8 metrics, which are used by cardiologists and in deployed algorithms (GE , 2019), covering rate, rhythm, timing, hypertrophy and conduction (Dubin, 2000). We find that CARP has the lowest error on most clinical measurements across in-domain and external datasets (Table 1). In particular, CARP has the lowest heart-rate, PR, QT, P-duration, T-duration, ST-level, and QRS-axis errors on both MIMIC-IV and the external PTB-XL, and is comparable on QRS duration. In embedding space, CARP achieves the highest recall and the lowest FID on both datasets.

Do text-to-ECG models reproduce features that a report implies but does not state? ECG reports are *under-specified*: they omit normal findings such as a normal cardiac axis ($-30°$ to $90°$), and they leave out abnormalities that a diagnosis implies (e.g. left anterior fascicular block (LAFB) implies left axis). We test whether generators reproduce these features using the cardiac axis, which can be measured directly from the signal (Appendix C.5). We group reports by how they convey the axis: *Stated* reports explicitly state the axis, *Implied* reports contain a diagnosis that implies the axis, and *Unstated* reports name neither, which usually implies a normal axis. A generated ECG aligns with its report when its measured axis falls in the expected category. On MIMIC-IV-ECG, CARP has the highest alignment and the lowest axis error in every group (Fig. 4c, Table 11), with the largest gap on *Implied* reports.

3.4 RQ3: Are CARP's components necessary?

We ablate two components: (a) randomly initializing the Cosmos generator, and (b) training w/o the clinical metric loss, and train all three variants for 8 epochs (Table 2). Without the pretrained generator, we observe that the model does not learn rhythm: mean RR error rises from 12.6 to 157.4 ms, and real-to-synthetic AUPRC falls from 73.0% to 17.8%. When ablating $\mathcal{L}_{metric}$, we find

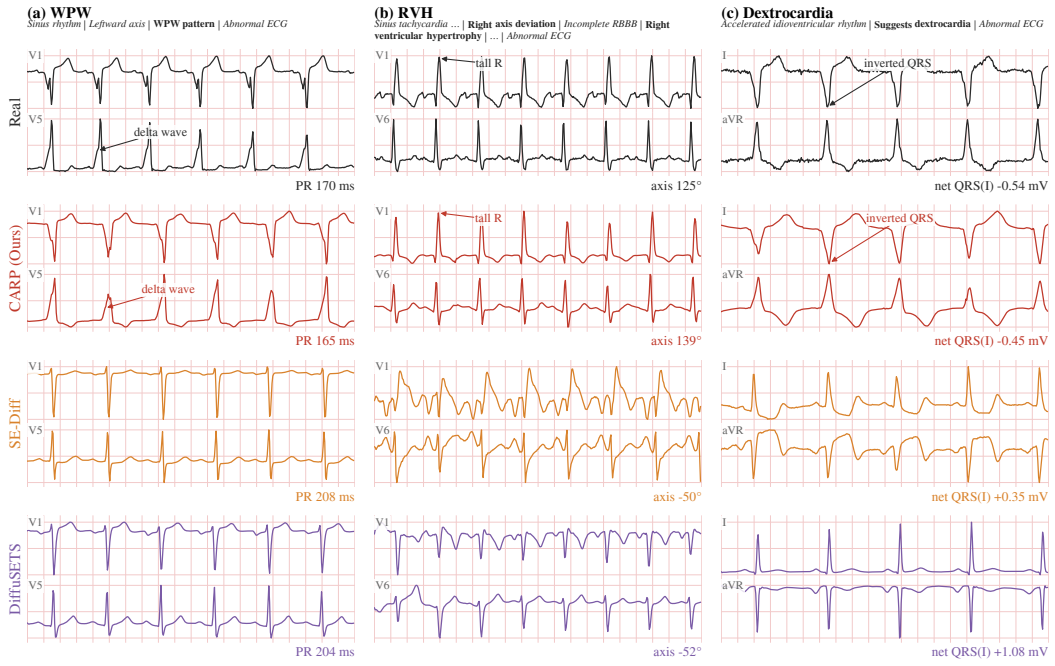


Figure 6: **Qualitative study of ECGs generated for rare conditions, showing that CARP is more clinically aligned with the morphology described in free-text reports, compared to baselines.** We show the label-relevant leads over the first 4s of each ECG. Arrows mark the label-defining morphology on the real ECG and CARP: the notched QRS ("delta wave") in V5 **(a)**, the tall R wave in V1 **(b)**, and the inverted QRS in lead I **(c)**; neither baseline reproduces these. Below each row is the relevant measurement, where CARP is closest to the real ECG in all three cases. Both baselines produce a left axis in **(b)** despite the stated right axis deviation (details in Appendix B).

that rate labels are unaffected (sinus bradycardia 90.2% vs. 90.3%), while morphology-related labels degrade (morphology AUPRC 67.0% to 65.3%; LBBB 75.2% to 66.2%). All seven morphology errors of Table 12 rise, significantly for PR, QRS, QT, T duration and ST level. Axis alignment is lower or equal in all three report groups on both datasets, and FID nearly doubles (3.5 to 6.6).

4 Conclusion

In this work, we introduced CARP, which adapts a pretrained world model to text-to-ECG generation and supervises morphological measurements through a differentiable soft-measurement objective. Its ECGs follow the report, age and sex more faithfully than prior generators, and they improve rare-disease classifiers. Despite its contributions, limitations remain. The first deals with evaluation; our model is evaluated on three datasets from high-income countries which share labeling conventions; results may shift in datasets from low-resource settings. Furthermore, important rare conditions such as Brugada are not present in the analyzed datasets, limiting evaluation. Finally, akin to prior work, our conditioning metadata is limited to age, sex and heart rate. Future work could incorporate single-lead ambulatory signals, or richer longitudinal text from electronic health records (EHR). Another option is to condition generation on a patient's prior ECGs in addition to their clinical text. Our approach, with world model priors and a morphology objective, is suited to such extensions.

Table 2: **Ablations on CARP's components: pretrained generator and soft segmentation loss ($\mathcal{L}_{metric}$).** Full results in Table 12

	CARP	w/o $\mathcal{L}_{metric}$	w/o gen.
Real-to-synthetic AUPRC (%) ↑			
All labels	**73.0**	71.7	17.8
Morphology	**67.0**	65.3	19.2
Rate/rhythm	**86.2**	85.4	17.1
Measurement error (ms) ↓			
PR	**18.2**	19.0	21.3
QRS	**9.7**	10.5	12.9
QT	**20.9**	21.7	38.4
Mean RR	**12.6**	13.3	157.4
Other			
Axis alignment, stated (%) ↑	**91.7**	88.8	50.6
FID ↓	**3.5**	6.6	26.2

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance (grammar, wording, and clarity), basic method ablating/debugging, and for code implementation (training, evaluation, and plotting scripts). We did not use generative AI tools to formulate the research questions, design the evaluation, or interpret the results. The authors reviewed and edited all AI-assisted text, and reviewed all AI-assisted code and checked its outputs, including by plotting generated ECGs against their real references. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We adhere to the ICLR Code of Ethics. This study uses only de-identified, publicly available ECG datasets: MIMIC-IV-ECG (Gow et al., 2023) and HEEDB (Koscova et al., 2026), accessed under their respective credentialed-access data use agreements, and PTB-XL (Wagner et al., 2020). No direct interaction with human subjects occurred, no personally identifiable information was accessed or released, and we made no attempt at re-identification. In line with PhysioNet's policy on credentialed data (PhysioNet, 2023), we replaced the proprietary LLM in the SE-Diff baseline with an open-source model (Appendix C). CARP is intended for research only and is not a medical device.

REPRODUCIBILITY STATEMENT

We release the code, training and evaluation configurations, and a subset of generated ECGs at an anonymous repository: `https://anonymous.4open.science/r/carp-iclr/`. We will release trained checkpoints if accepted. Appendix C.1 describes the train, validation, and test cohorts and the patient-level splits; Appendix C.2 lists the training hyperparameters; the Baselines subsection of Appendix C describes how SE-Diff and DiffuSETS were run; and Appendix C.4 and the RQ2 subsection of Appendix C give the label definitions and evaluation protocols. Inference latency is reported in Appendix C.8.

REFERENCES

Edmond Adib, Fatemeh Afghah, and John J. Prevost. Arrhythmia classification using CGAN-augmented ECG signals. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1865–1872. IEEE, 2022. doi: 10.1109/BIBM55620.2022.9995088.

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. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2023.107115. URL `https://www.sciencedirect.com/science/article/pii/S0010482523005802`.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Zachi I. Attia, Peter A. Noseworthy, Francisco Lopez-Jimenez, Samuel J. Asirvatham, Abhishek J. Deshmukh, Bernard J. Gersh, Rickey E. Carter, Xiaoxi Yao, Alejandro A. Rabinstein, Brad J. Erickson, Suraj Kapa, and Paul A. Friedman. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201):861–867, September 2019. doi: 10.1016/S0140-6736(19)31721-0.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video gen-

eration models as world simulators. OpenAI technical report, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Ed Burns and Robert Buttner. Killer ECG patterns: Part 1. Life in the Fast Lane (LITFL), ECG Library, September 2021. URL https://litfl.com/killer-ecg-patterns-part-1/. Last updated 8 October 2024. Accessed 20 September 2026.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL https://www.jair.org/index.php/jair/article/view/10302.

Jintai Chen, Kuanlun Liao, Kun Wei, Haochao Ying, Danny Z. Chen, and Jian Wu. ME-GAN: Learning panoptic electrocardio representations for multi-view ECG synthesis conditioned on heart diseases. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3360–3370. PMLR, 2022. URL https://proceedings.mlr.press/v162/chen22n.html.

Dale Dubin. *Rapid Interpretation of EKG's: An Interactive Course*. Cover Publishing Company, Tampa, FL, 6 edition, 2000.

Marquette 12SL ECG Analysis Program: Physician's Guide. GE Healthcare, Milwaukee, WI, USA, 3 2019. Document no. 2056246-002C, Revision C.

Tomer Golany and Kira Radinsky. PGANs: Personalized generative adversarial networks for ECG synthesis to improve patient-specific deep ECG classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 557–564, 2019. doi: 10.1609/aaai.v33i01.3301557.

Tomer Golany, Kira Radinsky, and Daniel Freedman. SimGANs: Simulator-based generative adversarial networks for ECG synthesis to improve deep ECG classification. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3597–3606. PMLR, 2020. URL https://proceedings.mlr.press/v119/golany20a.html.

Brian Gow, Tom Pollard, Larry A. Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W. Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset. *PhysioNet*, 2023. doi: 10.13026/4nqg-sb35. Version 1.0.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G. Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19622–19635, 2023. doi: 10.52202/075280-0861.

Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1): 65–69, 2019. doi: 10.1038/s41591-018-0268-3.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Jiarui Jin, Haoyu Wang, Xiang Lan, Jun Li, Hongyan Li, and Shenda Hong. Uniecg: Understanding and generating ecg in one unified model, 2026. URL https://arxiv.org/abs/2509.18588.

Zuzana Koscova, Valdery Moura Junior, Matthew Reyna, Shenda Hong, Aditya Gupta, Manohar Ghanta, Reza Sameni, Aaron Aguirre, Qiao Li, Sahar Zafar, Gari Clifford, and M. Brandon Westover. Harvard-emory ecg database. Brain Data Science Platform, 2026. URL https://bdsp.io/content/heedb/5.0/.

Yongfan Lai, Jiabo Chen, Qinghao Zhao, Deyun Zhang, Yue Wang, Shijia Geng, Hongyan Li, and Shenda Hong. DiffuSETS: 12-lead ECG generation conditioned on clinical text reports and patient-specific information. *Patterns*, 6(10):101291, 2025. doi: 10.1016/j.patter.2025.101291.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017. doi: 10.1109/ICCV.2017.324.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Frozen pretrained transformers as universal computation engines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7628–7636, 2022. doi: 10.1609/aaai.v36i7.20729.

P.W. Macfarlane, B. Devine, and E. Clark. The university of glasgow (uni-g) ecg analysis program. In *Computers in Cardiology, 2005*, pp. 451–454, 2005. doi: 10.1109/CIC.2005.1588134.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

NVIDIA, :, Aditi, Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, Aarti Basant, Mukesh Beladiya, Mohammad Qazim Bhat, Zaid Pervaiz Bhat, Dan Blick, Vanni Brighella, Han Cai, Tiffany Cai, Eric Cameracci, Jiaxin Cao, Yulong Cao, Mark Carlson, Carlos Casanova, Ting-Yun Chang, Yan Chang, Yu-Wei Chao, Prithvijit Chattopadhyay, Roshan Chaudhari, Chieh-Yun Chen, Junyu Chen, Ke Chen, Qizhi Chen, Wenkai Chen, Xiaotong Chen, Yu Chen, An-Chieh Cheng, Click Cheng, Xiu Chia, Jeana Choi, Chaeyeon Chung, Wenyan Cong, Yin Cui, Magdalena Dadela, Nalin Dadhich, Wenliang Dai, Joyjit Daw, Alperen Degirmenci, Rodrigo Vieira Del Monte, Robert Denomme, Sameer Dharur, Marco Di Lucca, Ke Ding, Wenhao Ding, Yifan Ding, Yuzhu Dong, Nicole Drumheller, Yilun Du, Aigul Dzhumamuratova, Aleksandr Efitorov, Hamid Eghbalzadeh, Naomi Eigbe, Imad El Hanafi, Hassan Eslami, Benedikt Falk, Jiaojiao Fan, Jim Fan, Amol Fasale, Sergiy Fefilatyev, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Vikram Fugro, Prashant Gaikwad, TJ Galda, Katelyn Gao, Yihuai Gao, Wenhang Ge, Sreyan Ghosh, Arushi Goel, Vivek Goel, Akash Gokul, Rama Govindaraju, Jinwei Gu, Miguel Guerrero, Elfie Guo,

Aryaman Gupta, Siddharth Gururani, Hugo Hadfield, Song Han, Ankur Handa, Zekun Hao, Mohammad Harrim, Ali Hassani, Nathan Hayes-Roth, Yufan He, Chris Helvig, Cyrus Hogg, Madison Huang, Michael Huang, Sophia Huang, Yufan Huang, Jacob Huffman, DeLesley Hutchins, Suneel Indupuru, Boris Ivanovic, Arihant Jain, Joel Jang, Ryan Ji, Yanan Jian, Dongfu Jiang, Jingyi Jin, Atharva Joshi, Nikhilesh Joshi, Pranjali Joshi, Andy Ju, Jaehun Jung, Weiwei Kang, Scott Kassekert, Jan Kautz, Ashna Khetan, Julia Kiczka, Slawek Kierat, Gwanghyun Kim, Kuno Kim, Sunny Kim, Kezhi Kong, Xin Kong, Zhifeng Kong, Tomasz Kornuta, Egor Krivov, Hui Kuang, Saurav Kumar, Chia-Wen Kuo, George Kurian, Wojciech Kutak, JF Lafleche, Himangshu Lahkar, Omar Laymoun, Jayjun Lee, Sanggil Lee, Gabriele Leone, Boyi Li, Freya Li, Jiajun Li, Jinfeng Li, Ling Li, Pengcheng Li, Shangru Li, Tingle Li, Xiaolong Li, Xuan Li, Zhaoshuo Li, Zhiqi Li, Hao Liang, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Ming-Yu Liu, Sifei Liu, Zihan Liu, Hai Loc Lu, Xiangyu Lu, Alice Luo, Ruipu Luo, Wenjie Luo, Jiangran Lyu, Martin Ding Ma, Nic Ma, Qianli Ma, Dawid Majchrowski, Louis Marcoux, Miguel Martin, Qing Miao, Ashkan Mirzaei, Shreyas Misra, Kaichun Mo, Durra Mohsin, Hyejin Moon, Pawel Morkisz, Saeid Motiian, Kirill Motkov, Seungjun Nah, Yashraj Narang, Deepak Narayanan, Thabang Ngazimbi, Julian Ouyang, Shubham Pachori, David Page, Yatian Pang, Sehwi Park, Mahesh Patekar, Mostofa Patwary, Marco Pavone, Trung Pham, Wei Ping, Soha Pouya, Shrimai Prabhumoye, Varun Praveen, Delin Qu, Hesam Rabeti, Morteza Ramezanali, Marilyn Reeb, Xuanchi Ren, Kristen Rumley, Wojciech Rymer, Jun Saito, Yeongho Seol, John Shao, Piyush Shekdar, Tianwei Shen, Humphrey Shi, Min Shi, Stella Shi, Kevin Shih, Mohammad Shoeybi, Mateusz Sieniawski, Shuran Song, Alexander Sotelo, Amir Sotoodeh, Sunil Srinivasa, Vignesh Srinivasakumar, Bartosz Stefaniak, Rahul Heinrich Steiger, Shangkun Sun, Jiaxiang Tang, Shitao Tang, Yangyang Tang, Yue Tang, Tolou Tavakkoli, Kayley Ting, Krzysztof Tomala, Wei-Cheng Tseng, Jibin Varghese, Sergei Vasilev, Thomas Volk, Raju Wagwani, Roger Waleffe, Andrew Z. Wang, Boxiang Wang, Haoxiang Wang, Qiao Wang, Shihao Wang, Shijie Wang, Ting-Chun Wang, Yan Wang, Yu Wang, Rohit Watve, David Wehr, Fangyin Wei, Xinshuo Weng, Jay Zhangjie Wu, Kedi Wu, Hongchi Xia, Summer Xiao, Tianjun Xiao, Kevin Xie, Daguang Xu, Jiashu Xu, Mengyao Xu, Ruqing Xu, Xingqian Xu, Yao Xu, Dinghao Yang, Dong Yang, Hans Yang, Xiaodong Yang, Xuning Yang, Yichu Yang, Yurong You, Zhiding Yu, Hao Yuan, Simon Yuen, Xiaohui Zeng, Pengcuo Zeren, Cindy Zha, Haotian Zhang, Jenny Zhang, Jing Zhang, Liangkai Zhang, Paris Zhang, Shun Zhang, Xuanmeng Zhang, Zhizheng Zhang, Ann Zhao, Yilin Zhao, Yuliya Zhautouskaya, Charles Zhou, Fengzhe Zhou, Shilin Zhu, Yuke Zhu, Dima Zhylko, and Artur Zolkowski. Cosmos 3: Omnimodal world models for physical ai, 2026. URL https://arxiv.org/abs/2606.02800.

Minje Park, Jeonghwa Lim, Taehyung Yu, and Sunghoon Joo. SemiSegECG: A multi-dataset benchmark for semi-supervised semantic segmentation in ECG delineation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pp. 5099–5104. ACM, 2025. doi: 10.1145/3746252.3760790.

Erick A. Perez Alday, Annie Gu, Amit J. Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D. Clifford, and Matthew A. Reyna. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement*, 41(12):124003, 2020. doi: 10.1088/1361-6579/abc960.

PhysioNet. Responsible use of MIMIC data with online services like GPT. https://physionet.org/news/post/gpt-responsible-use/, 2023. MIT Laboratory for Computational Physiology. Posted 2023-04-18. Accessed 2026-09-25.

Nicolas Pilia, Claudia Nagel, Gustavo Lenis, Silvia Becker, Olaf Dössel, and Axel Loewe. ECGdeli – an open source ECG delineation toolbox for MATLAB. *SoftwareX*, 13:100639, 2021. doi: 10.1016/j.softx.2020.100639.

Timothy J. Poterucha, Linyuan Jing, Ramon Pimentel Ricart, Michael Adjei-Mosi, Joshua Finer, Dustin Hartzel, Christopher Kelsey, Aaron Long, Daniel Rocha, Jeffrey A. Ruhl, David vanMaanen, Marc A. Probst, Brock Daniels, Shalmali D. Joshi, Olivier Tastet, Denis Corbin, Robert Avram, Joshua P. Barrios, Geoffrey H. Tison, I-Min Chiu, David Ouyang, Alexander Volodarskiy, Michelle Castillo, Francisco A. Roedan Oliver, Paloma P. Malta, Siqin Ye, Gregg F. Rosner, Jose M. Dizon, Shah R. Ali, Qi Liu, Corey K. Bradley, Prashant Vaishnava, Carol A. Waksmonski,

Ersilia M. DeFilippis, Vratika Agarwal, Mark Lebehn, Polydoros N. Kampaktsis, Sofia Shames, Ashley N. Beecy, Deepa Kumaraiah, Shunichi Homma, Allan Schwartz, Rebecca T. Hahn, Martin Leon, Andrew J. Einstein, Mathew S. Maurer, Heidi S. Hartman, John Weston Hughes, Christopher M. Haggerty, and Pierre Elias. Detecting structural heart disease from electrocardiograms using AI. *Nature*, 644(8075):221–230, 2025. doi: 10.1038/s41586-025-09227-0.

Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira Jr., Thomas B. Schön, and Antonio Luiz P. Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1):1760, 2020. doi: 10.1038/s41467-020-15432-4.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Keenan Saleh, Raaif Hadadi, Yixiu Liang, Hong Wong, Arunashis Sau, James Howard, Evan Brittain, Jeffrey Annis, Majd El-Harasis, Matthew Shun-Shin, Jagdeep Mohal, Akriti Naraen, Jack Samways, Jessica Artico, James Ware, Prapa Kanagaratnam, Fu Siong Ng, Massoud Zolgharni, Wenjia Bai, Amanda Varnava, Zachary Whinnett, and Ahran Arnold. AI-ECG classification for Brugada syndrome: A study of machine learning techniques to optimise for limited datasets. *PLOS Digital Health*, 5(2):e0001222, February 2026. doi: 10.1371/journal.pdig.0001222. URL https://doi.org/10.1371/journal.pdig.0001222.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my gan? In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, pp. 218–234, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01215-1. doi: 10.1007/978-3-030-01216-8_14. URL https://doi.org/10.1007/978-3-030-01216-8_14.

Zitao Shuai, Zongzhe Xu, Yuntian Wu, Sirui Li, Tianhong Li, and Yuzhe Yang. Signal or noise? understanding generative models for real-world sensor time series. *arXiv preprint arXiv:2607.04245v1*, 2026. URL https://arxiv.org/abs/2607.04245v1.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Nils Strodthoff, Temesgen Mehari, Claudia Nagel, Philip Aston, Ashish Sundar, Claus Graff, Jørgen Kanters, Wilhelm Haverkamp, Olaf Doessel, Axel Loewe, Markus Bär, and Tobias Schaeffter. PTB-XL+, a comprehensive electrocardiographic feature dataset. *PhysioNet*, June 2023. doi: 10.13026/g6h6-7g88. URL https://doi.org/10.13026/g6h6-7g88. Version 1.0.1.

Vajira Thambawita, Jonas L. Isaksen, Steven A. Hicks, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Christina Ellervik, Morten Salling Olesen, Torben Hansen, Claus Graff, Niels-Henrik Holstein-Rathlou, Inga Strümke, Hugo L. Hammer, Mary M. Maleckar, Pål Halvorsen, Michael A. Riegler, and Jørgen K. Kanters. DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. *Scientific Reports*, 11(1):21896, 2021. doi: 10.1038/s41598-021-01295-2.

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*, 7:154, 2020. doi: 10.1038/s41597-020-0495-6.

Xiaoda Wang, Kaiqiao Han, Yuhao Xu, Xiao Luo, Yizhou Sun, Wei Wang, and Carl Yang. SE-Diff: Simulator and experience enhanced diffusion model for comprehensive ECG generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/43f5f6c5cb333115914c8448b8506411-Abstract-Conference.html.

Frank N. Wilson, Franklin D. Johnston, Francis F. Rosenbaum, and Paul S. Barker. On einthoven's triangle, the theory of unipolar electrocardiographic leads, and the interpretation of the precordial electrocardiogram. *American Heart Journal*, 32(3):277–310, 1946. ISSN 0002-8703. doi: 10.1016/0002-8703(46)90791-0. URL https://www.sciencedirect.com/science/article/pii/0002870346907910.

Ryuichiro Yagi, Shinichi Goto, Yoshinori Katsumata, Calum A. MacRae, and Rahul C. Deo. Importance of external validation and subgroup analysis of artificial intelligence in the detection of low ejection fraction from electrocardiograms. *European Heart Journal – Digital Health*, 3(4): 654–657, November 2022. doi: 10.1093/ehjdh/ztac065.

Jenny Yang, Nguyen Thanh Dung, Pham Ngoc Thach, Nguyen Thanh Phong, Vu Dinh Phu, Khiem Dong Phu, Lam Minh Yen, Doan Bui Xuan Thy, Andrew A. S. Soltan, Louise Thwaites, and David A. Clifton. Generalizability assessment of AI models across hospitals in a low-middle and high income country. *Nature Communications*, 15(1):8270, September 2024. doi: 10.1038/s41467-024-52618-6. URL https://www.nature.com/articles/s41467-024-52618-6.

Beatrice Zanchi, Giuliana Monachino, Francesca Dalia Faraci, Matteo Metaldi, Pedro Brugada, Georgia Sarquella-Brugada, Elijah R Behr, Josep Brugada, Lia Crotti, Bernard Belhassen, and Giulio Conte. Synthetic electrocardiograms for brugada syndrome: from data generation to expert cardiologists evaluation. *European Heart Journal - Digital Health*, 6(4):683–687, 07 2025. ISSN 2634-3916. doi: 10.1093/ehjdh/ztaf039. URL https://doi.org/10.1093/ehjdh/ztaf039.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all: Power general time series analysis by pretrained LM. In *Advances in Neural Information Processing Systems*, volume 36, pp. 43322–43355, 2023. doi: 10.52202/075280-1877.

A Related Work

A.1 ECG Generators

Early ECG generators were GANs that synthesized single heartbeats to augment arrhythmia classifiers (Golany & Radinsky, 2019; Adib et al., 2022). Later, GANs were trained with a cardiac ODE simulator in the generator (Golany et al., 2020), and were extended to synthesize multi-lead 10-s ECGs (Thambawita et al., 2021; Chen et al., 2022). GANs are, however, unstable to train and prone to mode collapse in multi-lead, multi-label settings. Diffusion and score-based models (Ho et al., 2020; Song et al., 2021), and their latent (Rombach et al., 2022) flow-matching (Lipman et al., 2023) and rectified-flow (Liu et al., 2023) variants, train stably and cover modes better. These models condition on a fixed label set, if at all, and cannot express clinical text that states multiple diagnoses or names conditions differently across datasets (Appendix C.4.1).

Only two recent works address text-to-ECG generation, i.e., generating 12-lead ECGs conditioned on clinical text. DiffuSETS (Lai et al., 2025) trains a latent diffusion model from scratch on text-ECG pairs from MIMIC-IV-ECG, conditioned on the clinical report and patient metadata (age, sex and heart rate). It uses neither a pretrained generative prior nor a physiological prior, and must learn ECG structure from the text-ECG pairs alone. SE-Diff (Wang et al., 2026) adds a generic physiological prior: it fits the parameters of an ODE-based ECG simulator to each diagnostic category, penalizes generated beats that deviate from the simulated template, and augments the report with retrieval-based LLM text. However, a key limitation is that the simulator is fitted only for the 20 most frequent categories, and each sample is mapped to a single category template. A report with multiple diagnoses therefore receives a template for at most one of them, and a report whose diagnoses all lie outside the top 20 receives no simulator supervision; hence its likely failure in Figs.1 and 6. CARP instead adapts a pretrained world model to transfer its generative prior and directly supervises the morphological measurements of each text-ECG sample rather than matching a per-category template.

A.2 World Models

Whereas autoregressive models learn a distribution over (often text) sequences, world models learn *dynamics*, i.e., they predict future states conditioned on past states and, optionally, on an agent's actions. Recent world models are trained on large video and text corpora and predict future states directly as video (Brooks et al., 2024; Bruce et al., 2024; NVIDIA et al., 2025) or as video representations (Assran et al., 2025). World models are now emerging as general-purpose backbones: Cosmos 3 (NVIDIA et al., 2026) jointly models language, image, video, audio and action as a many-to-many generator, and its reasoner and generator towers share joint attention, so that a single shared backbone can enable text-to-video, image-to-video and many other mappings.

Prior work has shown that pretrained models can surprisingly transfer to modalities they were never trained on: frozen language-model layers perform well on image and protein tasks (Lu et al., 2022), and pretrained language models forecast and analyze time series with little or no fine-tuning (Gruver et al., 2023; Zhou et al., 2023). ECG generators, in contrast, are trained from scratch (Alcaraz & Strodthoff, 2023; Lai et al., 2025; Wang et al., 2026), as are all six generator families in a recent benchmark of generative models for sensor time series, which includes text-to-ECG (Shuai et al., 2026). To our knowledge, CARP is the first to transfer the generative prior of a world model to a physiological modality.

A.3 ECG Classification

Prior work has demonstrated that deep networks trained on sufficient ECG data reach cardiologist-level arrhythmia detection (Hannun et al., 2019; Ribeiro et al., 2020) and detect conditions that clinicians cannot (Attia et al., 2019; Poterucha et al., 2025). Their performance drops, however, on external hospitals and patient subgroups (Yagi et al., 2022; Yang et al., 2024), and data scarcity is most severe for rare or congenital conditions (Saleh et al., 2026). Class-imbalance remedies such as loss reweighting (Lin et al., 2017), oversampling and interpolation (Chawla et al., 2002; Zhang et al., 2018) are often insufficient (Appendix 8), so prior work augments training data with synthetic ECGs from task-specific GANs (Golany et al., 2020; Adib et al., 2022; Zanchi et al., 2025). SE-Diff evaluates this augmentation on two tasks (gender and SVT) (Wang et al., 2026) on a total of 200 test samples. We instead evaluate across 14 diagnostic label families, including five rare conditions, both in-domain (MIMIC-IV-ECG) and out-of-domain (PTB-XL, HEEDB) data.

B Qualitative Study

We examine ECGs produced for clinical text containing three rare conditions: Wolff-Parkinson-White syndrome, right ventricular hypertrophy, and dextrocardia, from the MIMIC-IV-ECG dataset (Fig. 6).

For the WPW report **(a)**, the ECG has a notched QRS, called a "delta wave", because of an extra electrical pathway that activates the ventricles early. CARP generates this delta-wave QRS pattern, but the baselines produce a regular, narrow QRS instead. The PR interval (165 ms) for the CARP ECG is also better aligned with the real ECG (170 ms) than SE-Diff (208 ms) or DiffuSETS (204 ms).

For the RVH report **(b)** that states right axis deviation explicitly, CARP places the frontal QRS axis at $139°$, close to the real ECG's $125°$. However, SE-Diff and DiffuSETS imply an axis of approximately $-50°$, which is a leftward axis that contradicts the report. Neither baseline shows a clear tall R wave in V1, unlike the real ECG and CARP.

A dextrocardia report, as in Fig. 6**(c)**, arises either from a true congenital defect where the heart points toward the right side of the chest or, anomalously, from misplaced ECG leads. In this instance, the report also indicates an "idioventricular" rhythm, implying a widened QRS complex. The report does not mention an inverted lead I, but CARP generates one, as expected in dextrocardia. Both baselines instead generate an upright, near-normal lead I, which implies a normally positioned heart. Despite this incorrect morphology, which can be measured by the axis inversion, we note that pointwise error metrics such as MAE or RMSE fail to capture these structural differences: SE-Diff and DiffuSETS exhibit lower MAE (107 and 104 μV) and RMSE (208 and 222 μV) than CARP (171 and 318 μV) (Appendix C.7).

C Additional Experimental Set-up

C.1 Train and Test Cohorts

We use the standard train/val/test split of MIMIC-IV-ECG (Gow et al., 2023) for training CARP, as with prior work (Wang et al., 2026; Lai et al., 2025). This results in 556,309 ECGs from 111,791 patients, 79,103 validation ECGs from 15,890 patients, and 158,956 test ECGs from 31,901 patients respectively. All splits are made at the patient-level, i.e. no two ECGs from a patient are in different splits. For external validation, we use all 21,799 records in PTB-XL and randomly drawn 1,000 test records (`seed` $= 2026$) from HEEDB.

C.2 Training Hyperparameters

We outline the full training hyperparameters and configuration for CARP below. We run for 32 epochs and use the validation loss to pick the best checkpoint. We train on 10 NVIDIA A100s with 80GB memory each; a full training run lasts 8 hours, including validation and test runs.

Table 3: Training settings for CARP.

Setting	Value
Epochs	32
Optimizer	AdamW, $\beta = (0.9, 0.95)$
Learning rate	10^{-4}; 50-step warm-up
Gradient clipping	Global norm 0.5
Batch size	640 (10 GPUs x 64)
Training seed	20260820
L_{lead} weight	1000
λ_{metric} (clinical metric loss)	0.25
t_c (low-noise cutoff for the clinical metric loss)	0.2
Rhythm term weight	20

C.3 Baselines

We train and test baselines on the same MIMIC-IV-ECG splits specified in Appendix C.1. We retrain both SE-Diff and DiffuSETS according to their released code and paper-specific implementation details. For SE-Diff, we use an open-source model `gemma-4-26B-A4B-it` instead of the original `gpt-4.1-mini` to reduce costs and comply with MIMIC-IV's credentialed usage policy for protected health data (PhysioNet, 2023). Reports generated by `gpt-4.1-mini` are not released in the SE-Diff code repository, and while a significant difference is not expected, we cannot regenerate those reports to make the direct comparison.

Classification and Inference. All ECGs (Real, SE-Diff, DiffuSETS, and CARP) are encoded via the same VAE released by DiffuSETS. We re-use the same sampling parameters ($n = 1000$) for SE-Diff and DiffuSETS, and provide inference latency times in Table C.8. All binary classifier experiments use the same set-up as SE-Diff (Wang et al., 2026); we ablate the choice of the VAE encoder in Table 7. For multilabel classification experiments, we use the *XResNet1d50* classifier from prior work (Wagner et al., 2020; Alcaraz & Strodthoff, 2023).

Classic augmentation baselines. We also evaluate against classic augmentation techniques, such as SMOTE (Chawla et al., 2002), Mixup (Zhang et al., 2018), and simply adding Gaussian noise, which we term *Real-Aug*; we find that the latter performs at par with the other techniques (Table 8). SMOTE and Mixup each add 190 augmented positives to the ten real positives. *Real-Aug* copies each real positive 19 times, rescales it by a factor drawn from $\mathcal{N}(1, 0.05^2)$, and adds Gaussian noise at 5% of each input coordinate's standard deviation over the training set. *SMOTE* (Chawla et al., 2002) interpolates linearly between a real positive and one of its five nearest real positives, with a uniformly sampled weight. *Mixup* (Zhang et al., 2018) takes a convex combination of two random real positives, with weight $\lambda \sim \text{Beta}(0.4, 0.4)$, and keeps the positive label. SMOTE and Mixup

Table 4: **Five rare conditions used for evaluation**: Wolff-Parkinson-White (WPW), right ventricular hypertrophy (RVH), posterior myocardial infarction (PMI), pericarditis (PCT) and dextrocardia (DCT). Prevalence is the share (and count) of positive ECGs among all MIMIC-IV-ECG recordings. RAD: right axis deviation; MI: myocardial infarction; ↑/↓: elevation/depression.

Condition	Prevalence	Cause	ECG findings	Clinical implication
WPW	0.077% (613)	Extra conduction pathway	Short PR, delta wave	Fast arrhythmias
RVH	0.58% (4,632)	Thick right ventricle	RAD, tall R in V1	Heart/lung disease
PMI	0.070% (555)	Posterior-wall MI	ST↓, tall R in V1–V3	Large infarct
PCT	0.31% (2,484)	Inflamed pericardium	ST↑, PR↓	Risk of tamponade
DCT	0.039% (306)	Heart on the right side	Inverted beats in lead I	Linked heart defects

operate on the raw waveform, which is then encoded like every other input. *Focal loss* (Lin et al., 2017) (γ=2) instead trains on the ten real positives alone.

C.4 RQ1 Details

C.4.1 Rare Conditions

To identify rare conditions for evaluation, we focus on conditions with (a) high clinical severity (b) low population prevalence (c) at least 50 records across datasets, to enable statistically-significant evaluation. To determine severity, we use prior work by cardiologists that identifies "killer" ECG patterns, which are conditions that doctors are trained to triage and escalate (Burns & Buttner, 2021). We then filter for conditions with low prevalence ($<0.6\%$), but have at least 50 ECGs across datasets to enable statistically-significant evaluation; this excludes very-rare conditions such as Brugada type 1 (a congenital disorder that increases risk of cardiac arrest), since only 2 ECG reports across the 3 datasets (MIMIC, PTB-XL and HEEDB) mention the condition. Our criteria results in five conditions with low prevalence: Wolff-Parkinson-White syndrome (WPW), Right Ventricular Hypertrophy (RVH), Posterior Myocardial Infarct (PMI), pericarditis (PCT) and Dextrocardia (DCT). Table 4 summarizes the cause, ECG findings, and clinical implication of each condition. Per-dataset prevalence and absolute counts are provided in Table 5.

Benefits of Free-Text Conditioning. We further note that these rare findings are documented inconsistently across datasets, which makes label-based conditioning brittle, and motivates the free-text conditioning that prior work (Wang et al., 2026; Lai et al., 2025) introduces. As an illustration, consider how WPW is indicated across datasets: `Ventricular preexcitation(WPW)` (MIMIC-IV), `wolff-parkinson-white syndrome` (PTB-XL), and `Ventricular pre-excitation, WPW pattern type A` (HEEDB).

C.4.2 Full Results

Table 5 reports AUPRC for the five rare conditions and the 14 diagnostic label families of Appendix C.4.1, in-domain on MIMIC-IV-ECG and out-of-domain on PTB-XL and HEEDB.

Table 6 evaluates age and sex, as conditioning inputs (real-to-synthetic) and as imbalanced classification tasks (synthetic-to-real).

C.5 RQ2 Details

Real-to-synth experiment. We train a ResNet classifier (Appendix C.4) on 556k real ECGs from MIMIC (train split); we produce synthetic ECGs from the reports and metadata of the test set (2,055 ECGs), and test the classifier on those. Full results are in Table 9.

Table 5: **CARP augmented ECG classifiers have the highest average AUPRC over the 14 diagnostic label families on all three datasets, with pronounced gains for rare conditions.** We split the misc. rare family into the five rare conditions of §3.2. Labels with fewer than 50 test positives are omitted. *Real (200 pos.)* is a reference, i.e. trained on 200 real MIMIC-IV-ECG positives instead of 10 real and 190 synthetic ones.

Label	Prevalence (Count)	Real	Real-Aug	SE-Diff	DiffuSETS	CARP	Real (200 pos.)
MIMIC-IV							
Normal / sinus rhythm	82.3% (654,093)	68.2 ±7.6	66.5 ±8.7	74.8 ±3.6	79.0 ±1.3	**80.4** ±3.3	84.5 ±2.0
Sinus bradycardia	12.4% (98,180)	81.5 ±13.8	80.3 ±13.8	**94.0** ±0.8	92.9 ±1.9	92.9 ±1.1	93.9 ±1.3
Sinus tachycardia	8.7% (69,154)	69.3 ±7.9	72.2 ±8.0	89.7 ±2.7	92.9 ±0.4	**95.6** ±2.3	96.0 ±1.3
AF / flutter	11.9% (94,841)	58.6 ±7.2	58.0 ±5.8	73.7 ±3.7	72.0 ±1.8	**77.4** ±2.5	85.0 ±5.2
Bundle-branch / IV block	14.8% (117,382)	77.2 ±12.4	78.5 ±10.4	73.4 ±3.0	85.9 ±2.5	**90.6** ±1.0	91.6 ±2.1
AV / fascicular / axis	26.6% (211,669)	61.9 ±9.0	62.6 ±6.8	75.6 ±2.8	76.2 ±2.0	**78.8** ±1.9	84.0 ±1.7
Atrial / ventricular ectopy	10.1% (80,460)	51.1 ±1.0	54.1 ±1.0	**58.8** ±1.2	54.0 ±2.3	58.5 ±1.6	62.1 ±0.5
Hypertrophy / enlargement	14.1% (111,983)	56.2 ±5.4	61.4 ±9.9	65.3 ±7.0	69.9 ±4.0	**77.0** ±1.4	78.3 ±2.3
Myocardial infarction	21.5% (170,406)	53.6 ±2.8	55.4 ±4.1	64.4 ±1.7	66.0 ±1.2	**66.9** ±1.1	71.4 ±3.1
Ischemia / ST-T	39.2% (311,772)	57.9 ±5.2	60.7 ±4.5	67.0 ±1.9	62.3 ±0.4	**71.5** ±1.9	74.8 ±1.4
Low voltage	9.3% (74,214)	61.9 ±3.0	60.6 ±3.4	66.4 ±2.1	**66.7** ±0.5	65.7 ±3.8	77.6 ±4.0
Paced rhythm	3.8% (30,130)	73.6 ±7.5	76.6 ±3.4	83.9 ±0.9	85.6 ±0.6	**87.2** ±1.1	85.2 ±1.5
QT prolongation	5.0% (39,410)	58.0 ±4.2	59.0 ±2.0	67.9 ±2.6	71.4 ±0.3	**79.2** ±1.5	75.7 ±4.1
Misc. rare conditions							
WPW	0.077% (613)	59.8 ±8.5	60.3 ±8.8	44.9 ±4.8	51.4 ±4.0	**71.6** ±3.8	78.8 ±2.0
Posterior MI	0.070% (555)	82.3 ±2.6	82.8 ±3.3	55.9 ±4.1	61.3 ±5.9	**91.4** ±1.7	94.5 ±0.5
RVH	0.58% (4,632)	71.5 ±3.8	72.8 ±3.5	55.7 ±8.1	70.5 ±2.8	**84.5** ±1.2	88.2 ±2.0
Pericarditis	0.31% (2,484)	74.2 ±12.2	73.2 ±14.9	74.5 ±0.8	84.1 ±0.4	**88.4** ±1.7	90.5 ±1.6
Dextrocardia	0.039% (306)	86.2 ±3.5	86.4 ±6.5	81.5 ±4.2	84.9 ±5.4	**91.4** ±1.7	92.3 ±3.6
Other rare rhythms	4.9% (39,230)	50.6 ±6.1	53.0 ±4.2	53.0 ±2.7	53.6 ±2.1	**56.9** ±3.6	62.6 ±1.9
Average (14 families)		62.8	64.2	72.0	73.4	**77.1**	80.2
PTB-XL (external)							
Normal / sinus rhythm	88.4% (19,266)	60.7 ±4.9	58.1 ±5.8	67.5 ±4.5	**75.1** ±2.5	60.1 ±8.8	68.7 ±11.9
Sinus bradycardia	2.9% (637)	63.4 ±9.3	62.0 ±8.2	**80.0** ±0.6	73.0 ±6.0	75.0 ±5.0	73.5 ±5.2
Sinus tachycardia	3.8% (826)	57.6 ±10.6	62.0 ±13.5	78.5 ±1.6	85.7 ±3.9	**92.8** ±2.1	96.6 ±0.3
AF / flutter	7.2% (1,570)	52.2 ±12.2	54.5 ±10.6	65.2 ±1.0	58.5 ±8.7	**67.5** ±8.2	75.6 ±9.1
Bundle-branch / IV block	13.9% (3,037)	67.9 ±14.6	70.8 ±11.6	70.4 ±1.4	77.1 ±1.3	**82.9** ±0.7	84.0 ±0.7
AV / fascicular / axis	12.4% (2,713)	59.1 ±5.7	59.9 ±6.4	60.4 ±2.5	59.5 ±5.3	**63.5** ±1.2	71.7 ±1.0
Atrial / ventricular ectopy	7.0% (1,521)	54.7 ±1.8	57.0 ±3.5	52.3 ±3.3	51.2 ±5.3	**64.1** ±8.7	65.6 ±2.1
Hypertrophy / enlargement	10.4% (2,258)	54.4 ±4.7	58.5 ±5.4	56.7 ±9.6	61.1 ±3.3	**77.1** ±6.9	71.8 ±1.2
Myocardial infarction	18.0% (3,926)	50.4 ±3.8	53.6 ±2.6	57.6 ±5.4	54.3 ±6.7	**61.8** ±3.1	62.6 ±3.9
Ischemia / ST-T	26.5% (5,783)	54.9 ±10.0	52.0 ±7.4	60.1 ±5.6	63.0 ±5.4	**74.0** ±1.5	70.5 ±5.5
Misc. rare conditions							
WPW	0.36% (79)	50.4 ±3.8	50.7 ±3.8	48.3 ±7.9	55.7 ±4.2	**73.7** ±4.2	70.0 ±2.0
Posterior MI	0.46% (101)	57.3 ±8.2	58.0 ±10.3	39.6 ±2.5	40.1 ±1.4	**72.2** ±1.3	73.1 ±2.6
RVH	0.25% (55)	68.1 ±4.9	70.7 ±2.8	61.5 ±9.7	73.7 ±4.1	**90.3** ±3.0	90.2 ±2.4
Average (14 families)		57.5	58.8	64.9	65.8	**71.9**	74.1
HEEDB (external)							
Normal / sinus rhythm	80.6% (6,505,136)	51.6 ±1.2	51.7 ±1.4	62.3 ±1.4	**70.0** ±4.4	67.0 ±8.8	69.5 ±4.4
Sinus bradycardia	12.2% (985,473)	69.3 ±8.8	67.5 ±9.4	84.3 ±1.9	**85.1** ±3.4	84.7 ±3.9	84.2 ±1.9
Sinus tachycardia	9.0% (727,181)	64.9 ±8.4	67.5 ±3.3	73.7 ±4.8	74.8 ±5.9	**86.0** ±3.7	87.4 ±1.6
AF / flutter	9.1% (733,698)	53.5 ±8.5	54.6 ±8.2	59.1 ±2.3	56.6 ±4.1	**62.1** ±2.6	71.0 ±3.8
Bundle-branch / IV block	15.1% (1,219,764)	69.0 ±16.2	72.1 ±12.6	79.7 ±0.9	82.0 ±3.4	**89.7** ±1.5	87.3 ±1.2
AV / fascicular / axis	26.9% (2,172,192)	58.4 ±5.6	59.9 ±6.2	70.5 ±1.6	66.0 ±0.2	**73.3** ±3.4	78.7 ±2.2
Atrial / ventricular ectopy	12.1% (977,991)	55.9 ±1.9	57.7 ±3.5	**59.7** ±2.2	50.3 ±1.6	59.0 ±4.8	67.1 ±0.8
Hypertrophy / enlargement	15.8% (1,275,436)	56.7 ±3.4	57.2 ±3.5	60.3 ±6.9	53.9 ±1.6	**66.6** ±6.3	68.9 ±5.1
Myocardial infarction	20.4% (1,645,479)	52.4 ±2.4	53.0 ±3.3	**61.0** ±3.3	54.9 ±4.5	60.8 ±1.3	61.9 ±2.6
Ischemia / ST-T	31.2% (2,522,378)	54.0 ±7.7	53.9 ±6.3	58.4 ±5.1	57.6 ±5.8	**62.9** ±1.4	70.5 ±4.0
Low voltage	9.5% (767,354)	65.2 ±6.8	62.3 ±5.8	**69.0** ±5.0	68.4 ±3.6	65.3 ±5.2	83.5 ±7.4
Paced rhythm	2.5% (205,847)	59.4 ±8.1	58.2 ±4.3	68.8 ±1.9	68.3 ±3.2	**79.0** ±3.7	69.3 ±5.1
QT prolongation	2.6% (208,025)	44.7 ±1.3	44.9 ±1.6	50.9 ±2.0	56.7 ±2.6	**66.6** ±2.2	59.8 ±2.4
Misc. rare conditions							
WPW	0.084% (6,759)	54.8 ±12.1	55.5 ±9.4	41.7 ±7.0	50.6 ±5.5	**77.9** ±4.1	79.3 ±3.0
Posterior MI	0.22% (17,928)	55.2 ±2.2	55.7 ±3.6	43.3 ±1.1	46.1 ±1.3	**63.5** ±0.8	65.5 ±2.0
RVH	0.81% (65,810)	62.9 ±5.8	64.4 ±4.9	53.3 ±9.0	60.1 ±5.8	**77.7** ±3.6	81.2 ±3.5
Pericarditis	0.055% (4,473)	62.9 ±11.9	62.0 ±10.9	71.0 ±3.1	82.6 ±2.1	**85.7** ±2.6	87.7 ±3.0
Dextrocardia	0.002% (190)	59.7 ±14.1	52.2 ±6.3	37.4 ±1.8	42.4 ±2.4	**66.9** ±10.3	73.2 ±10.4
Other rare rhythms	8.6% (693,778)	52.4 ±3.0	**53.1** ±3.0	53.0 ±0.7	48.1 ±3.0	52.4 ±0.2	55.4 ±3.7
Average (14 families)		57.7	58.1	65.1	63.8	**69.7**	72.5

Table 6: **Age and sex results, showing that CARP follows the requested age and sex most closely and is the best augmenter for both.**

(a) Real-to-synthetic. CARP follows the requested age and sex most closely, overall and among records with the same report; the baselines lose age almost entirely when the report is abnormal. We train a sex classifier and an age regressor only on real MIMIC-IV-ECG and test them on ECGs from each generator. We report the sex AUROC (%) and the correlation r between predicted and requested age over all 2,055 MIMIC test records ("Overall"), and among records with identical report text, after subtracting each report group's mean (groups of $\geq$2 records), so the report cannot explain differences in age or sex. We split these by the report's verdict: "Normal ECG" (10 groups, 350 records) and "Abnormal" or "Borderline ECG" (138 groups, 501 records).

	Sex AUROC			Age r		
Method	Overall	Normal	Abnormal	Overall	Normal	Abnormal
Real	94.1	92.8	91.5	0.76	0.80	0.67
SE-Diff	69.3	75.7	68.0	0.48	0.46	0.16
DiffuSETS	75.1	79.2	72.1	0.47	0.43	0.10
CARP	**85.9**	**88.8**	**81.1**	**0.67**	**0.72**	**0.49**

(b) Synthetic-to-real. CARP is the best augmenter for both sex and age, in-domain on MIMIC-IV and out-of-domain on PTB-XL and HEEDB. Both tasks are binary: sex (male vs. female) and age (18–44 vs. $\geq$70).

	Sex			Age		
Method	MIMIC-IV	PTB-XL	HEEDB	MIMIC-IV	PTB-XL	HEEDB
Real	53.0 ±5.5	53.6 ±2.1	50.7 ±2.9	70.0 ±10.5	63.2 ±6.4	64.7 ±8.0
Real-Aug	54.0 ±6.6	53.7 ±4.1	52.4 ±5.1	66.8 ±10.2	62.8 ±6.8	62.7 ±9.0
SE-Diff	62.3 ±1.9	62.4 ±2.8	56.6 ±1.4	75.3 ±5.4	71.4 ±6.3	68.5 ±6.1
DiffuSETS	66.6 ±3.2	66.2 ±0.9	59.1 ±0.9	81.4 ±3.1	75.1 ±4.3	74.7 ±3.1
CARP	**72.0** ±1.2	**67.9** ±0.8	**64.2** ±1.1	**85.8** ±2.2	**77.9** ±1.6	**79.1** ±0.9

Table 7: **Ablation results for choice of encoder**, finding that CARP's outperformance is robust to this decision. We repeat the rare-condition experiment of §3.2, with the input being the raw waveform. As before, *Real (200 pos.)* is a reference that trains the same classifier on 200 real positives instead of 190 synthetic ones. We report the mean AUPRC (%) over the five rare conditions.

Encoder	Test set	SE-Diff	DiffuSETS	CARP	Real (200 pos.)
DiffuSETS VAE	MIMIC-IV	62.5 ±2.4	70.4 ±0.7	**85.5** ±0.4	88.9 ±0.2
	PTB-XL	49.8 ±5.2	56.5 ±2.3	**78.7** ±2.4	77.8 ±1.5
	HEEDB	49.3 ±2.8	56.4 ±1.1	**74.4** ±2.9	77.4 ±3.5
None	MIMIC-IV	66.2 ±0.7	76.5 ±1.7	**86.6** ±0.8	92.0 ±0.8
	PTB-XL	45.3 ±2.6	60.9 ±5.4	**79.4** ±0.5	83.8 ±0.8
	HEEDB	52.6 ±0.5	59.6 ±2.4	**74.2** ±0.5	77.7 ±0.8

Table 8: **Additional results comparing classic augmentation strategies (AUPRC %), demonstrating that CARP outperforms classic augmentation strategies on the rare conditions.** We vary the augmentation used for producing the synthetic positives, with Gaussian noise and scaling (Real-Aug), SMOTE interpolation between nearest positives (Chawla et al., 2002), Mixup of two positives (Zhang et al., 2018), and focal loss (γ=2) without augmentation.

Label	Real	Real-Aug	SMOTE	Mixup	Focal loss	CARP
MIMIC-IV						
WPW	59.8 ±8.5	60.3 ±8.8	52.8 ±4.2	62.8 ±3.2	53.2 ±11.3	**71.6** ±3.8
Posterior MI	82.3 ±2.6	82.8 ±3.3	78.9 ±7.6	81.0 ±4.2	82.6 ±4.2	**91.4** ±1.7
RVH	71.5 ±3.8	72.8 ±3.5	70.2 ±4.1	72.8 ±2.9	68.7 ±7.0	**84.5** ±1.2
Pericarditis	74.2 ±12.2	73.2 ±14.9	71.2 ±5.0	70.2 ±8.6	74.4 ±9.5	**88.4** ±1.7
Dextrocardia	86.2 ±3.5	86.4 ±6.5	83.4 ±9.6	84.3 ±7.9	85.1 ±4.6	**91.4** ±1.7
Average	74.8 ±1.3	75.1 ±0.7	71.3 ±1.8	74.2 ±0.1	72.8 ±3.7	**85.5** ±0.4
PTB-XL (external)						
WPW	50.4 ±3.8	50.7 ±3.8	49.1 ±2.4	50.4 ±3.7	48.1 ±6.0	**73.7** ±4.2
Posterior MI	57.3 ±8.2	58.0 ±10.3	57.9 ±15.7	57.3 ±11.2	56.5 ±12.6	**72.2** ±1.3
RVH	68.1 ±4.9	70.7 ±2.8	67.4 ±7.6	71.3 ±4.2	68.5 ±6.8	**90.3** ±3.0
Average	58.6 ±3.4	59.8 ±4.0	58.1 ±6.6	59.7 ±3.6	57.7 ±6.4	**78.7** ±2.4
HEEDB (external)						
WPW	54.8 ±12.1	55.5 ±9.4	45.1 ±6.7	59.7 ±7.2	50.7 ±15.6	**77.9** ±4.1
Posterior MI	55.2 ±2.2	55.7 ±3.6	54.2 ±5.9	55.4 ±3.9	56.3 ±5.3	**63.5** ±0.8
RVH	62.9 ±5.8	64.4 ±4.9	63.0 ±6.5	64.7 ±4.3	61.5 ±8.8	**77.7** ±3.6
Pericarditis	62.9 ±11.9	62.0 ±10.9	54.5 ±1.9	56.0 ±4.1	60.6 ±11.8	**85.7** ±2.6
Dextrocardia	59.7 ±14.1	52.2 ±6.3	52.6 ±14.6	52.8 ±12.0	55.2 ±16.9	**66.9** ±10.3
Average	59.1 ±3.9	58.0 ±0.8	53.9 ±1.7	57.7 ±2.3	56.9 ±5.2	**74.4** ±2.9

Table 9: **Real-to-synthetic performance on MIMIC-IV-ECG, showing that CARP ECGs are more clinically aligned to their conditioning reports**. We evaluate a multilabel classifier trained only on real MIMIC-IV-ECG on ECGs from each generator; AUPRC (%) is reported per label and macro-averaged over the 12 labels on which the classifier is validated. Higher is better. Real (reference) applies the same classifier to the real ECGs of the same 2,055 test records. ± is the bootstrap standard error (1,000 paired resamples of test records). Bold marks the best generator in each row.

Label	Real (reference)	SE-Diff	DiffuSETS	CARP
Normal	92.5 ±1.3	61.2 ±3.4	44.8 ±3.1	**65.3 ±3.0**
Sinus bradycardia	97.9 ±1.0	**92.7 ±1.8**	90.3 ±2.0	91.9 ±2.1
Sinus tachycardia	97.9 ±0.7	92.2 ±1.8	93.3 ±1.3	**96.6 ±0.8**
Atrial fibrillation	94.7 ±1.2	67.7 ±3.2	64.6 ±3.1	**78.7 ±2.7**
LBBB	96.1 ±2.1	7.0 ±2.6	68.8 ±7.7	**91.2 ±3.9**
RBBB	96.9 ±0.8	15.2 ±2.0	86.1 ±2.2	**97.7 ±0.8**
Left axis deviation	94.6 ±1.3	38.1 ±2.9	47.6 ±3.3	**75.9 ±3.0**
1st-degree AV block	83.1 ±3.8	7.9 ±1.5	16.3 ±3.9	**54.1 ±5.7**
PVC	88.5 ±3.1	13.0 ±1.6	42.1 ±4.9	**90.7 ±2.1**
LVH	81.9 ±3.1	30.2 ±3.4	42.0 ±4.0	**76.7 ±3.5**
MI	95.4 ±1.0	31.8 ±2.7	36.7 ±3.2	**77.6 ±2.7**
Ischemia	80.1 ±2.3	35.5 ±3.0	43.0 ±3.3	**66.9 ±3.3**
Macro-AUPRC	91.6 ±0.6	41.0 ±0.8	56.3 ±1.1	**80.3 ±0.9**

Metric evaluation. For the metric evaluation in §3.3, we only include ECGs with patient-level age and sex information and physiologically plausible heart rate ($20 \leq HR \leq 300$). For MIMIC, we use a randomly-drawn subset of the test set, with 2,055 ECGs from 412 patients. For PTB-XL, we filter for English-language as well as cardiologist-written reports that agree with diagnostic labels and then randomly draw 2,941 records from 2,868 patients. For HEEDB we draw a sample of 1,000 eligible I0001 recordings, with no outcome balancing. Table 10 extends Table 1 to HEEDB.

Cardiac axis computation. We find, via comparison to richly annotated datasets such as PTB-XL+ (Strodthoff et al., 2023), that the common two-lead method of computing frontal QRS axis is often sensitive to error due to lags and noise, so we instead include a redundant lead, and fit the frontal vector to with three limb leads (I, II, III) by least squares. We find that this method

Table 10: Clinical measurements and distribution metrics on all three datasets. Top: MAE of ECGDeli measurements between each generated ECG and its paired real ECG, mean ± 95% CI (patient-level bootstrap); HR error is relative to the conditioning heart rate. Bottom: FID and k-NN precision and recall (k=3, %) in the DiffuSETS VAE embedding. Bold marks the best generator per dataset. The MIMIC-IV and PTB-XL columns repeat Table 1.

	MIMIC-IV			PTB-XL (external)			HEEDB (external)		
Metric	SE-Diff	DiffuSETS	CARP	SE-Diff	DiffuSETS	CARP	SE-Diff	DiffuSETS	CARP
HR (bpm) ↓	1.7 ±0.2	1.8 ±0.3	**1.1** ±0.2	1.4 ±0.1	7.1 ±0.1	**1.0** ±0.1	1.5 ±0.2	2.1 ±0.3	**0.8** ±0.1
PR (ms) ↓	27.7 ±1.7	26.2 ±1.8	**23.5** ±1.7	26.7 ±0.8	23.8 ±0.8	**23.1** ±0.7	24.1 ±1.2	22.7 ±1.2	**20.4** ±1.1
QRS (ms) ↓	14.0 ±0.7	11.0 ±0.7	**10.9** ±0.6	12.4 ±0.4	**10.8** ±0.3	11.0 ±0.3	12.3 ±0.6	10.3 ±0.6	**10.2** ±0.5
QT (ms) ↓	27.0 ±1.8	25.8 ±2.5	**24.3** ±2.1	24.0 ±0.7	24.5 ±0.7	**23.5** ±0.7	23.5 ±1.3	**22.2** ±1.2	22.9 ±1.1
P duration (ms) ↓	13.7 ±0.6	12.4 ±0.5	**11.8** ±0.6	14.4 ±0.4	12.9 ±0.4	**12.3** ±0.3	13.1 ±0.7	12.3 ±0.6	**11.6** ±0.5
T duration (ms) ↓	14.8 ±1.0	14.3 ±0.8	**13.4** ±0.9	12.6 ±0.4	12.8 ±0.4	**11.9** ±0.4	13.9 ±0.7	13.3 ±0.7	**13.1** ±0.6
ST J+60 (μV) ↓	56.3 ±3.9	54.0 ±4.3	**51.7** ±4.4	52.4 ±1.2	59.6 ±1.5	**52.0** ±1.4	52.2 ±2.3	**51.4** ±2.3	54.2 ±2.5
Cardiac axis (°) ↓	39.8 ±2.6	38.1 ±2.1	**30.0** ±1.9	36.6 ±1.1	44.2 ±1.3	**35.4** ±1.1	38.3 ±2.1	37.9 ±2.0	**33.1** ±1.8
FID ↓	17.54	6.68	**2.72**	28.47	22.92	**15.93**	22.52	**20.88**	25.47
Precision ↑	88.0	**90.4**	89.3	45.2	54.4	**57.0**	75.6	**81.5**	78.7
Recall ↑	80.4	83.9	**90.5**	75.7	85.5	**89.8**	75.0	85.7	**90.5**

computes axis at-par with commercial, proprietary software such as GE-12SL (GE , 2019) and Uni-G (Macfarlane et al., 2005).

For each lead, we first baseline-correct the Q, R, and S peaks against the median of the PR segment. We then compute the per-beat amplitude as $a = r + \min(q, s)$. We define the lead's *net amplitude* (i.e. $n_{\mathrm{I}}, n_{\mathrm{II}}, n_{\mathrm{III}}$) as the median of a across all valid beats. We fit a 2D vector to these net amplitudes based on the standard hexaxial directions $(0°, 60°, 120°)$ via least-squares fit, which has the closed form:

$$h = 2n_{\mathrm{I}} + n_{\mathrm{II}} - n_{\mathrm{III}} \tag{9}$$

$$f = \sqrt{3}(n_{\mathrm{II}} + n_{\mathrm{III}}) \tag{10}$$

The final cardiac axis is then $\hat{\phi} = \mathrm{atan2}(f, h)$.

Axis alignment with the report. We assign each report to at most one of three groups by parsing it statement by statement. A report is *Stated* if it names a left, right or extreme axis deviation, and the expected category is the named one. A report is *Implied* if it names no axis but a definite left anterior or left posterior fascicular block, which imply a left and a right axis respectively. All remaining reports are *Unstated*, and we expect a normal axis for them. We drop reports whose axis cannot be assigned reliably from the text alone. These include hedged findings ("possible left anterior fascicular block", "borderline left axis deviation"), an indeterminate axis, and bi- or trifascicular block, whose axis direction is not fixed.

We measure the axis of each generated ECG and the corresponding real ECG with the method above. Alignment is the share of records whose generated axis falls in the expected category: normal ($-30°$ to $90°$), left ($-90°$ to $-30°$), right ($90°$ to $180°$) or extreme ($-180°$ to $-90°$). We confirm our experiment design, we computed the same measurement on the real ECGs; the real ECGs align with 93.1% of the *Implied* reports. On *Unstated* reports, CARP aligns more often than the real ECGs (92.3% vs. 90.1%), which is likely due to fewer mild deviations in the generated ECGs. Table 11 provides the full results for MIMIC-IV and PTB-XL.

C.6 RQ3 Details

Table 12 compares CARP with two eight-epoch ablations on MIMIC-IV-ECG: one trained without the auxiliary loss, and one whose generator is trained from random initialization while the text reasoner stays pretrained.

C.7 Inadequacy of using MSE to Assess Clinical Fidelity

We find that metrics such as the mean squared error (MSE) and median absolute error are unsuitable to measure clinical fidelity, i.e., to compare the real and generated ECGs. Such metrics are sensitive to phase shift (Wang et al., 2026), amplitude scale if not normalized, or hardware specific lag

Table 11: **Axis error and report alignment by how the report conveys the axis.** *Stated* reports explicitly state the axis deviation; *Implied* reports contain diagnoses that imply the axis; *Unstated* reports have neither and so imply a normal axis. Axis categories are normal ($-30°$ to $90°$), left deviation ($-90°$ to $-30°$), right deviation ($90°$ to $180°$) and extreme ($-180°$ to $-90°$). Alignment is the share of records whose measured axis falls in the category the report implies; *Real ECGs* applies the same measurement to the reference recordings.

			Median angular error (°, ↓)			Report alignment (%, ↑)			
Dataset	Reports	n	CARP	SE-Diff	DiffuSETS	Real ECGs	CARP	SE-Diff	DiffuSETS
MIMIC-IV	Stated	17749	**12.2** [12, 13]	26.0 [25, 27]	24.1 [23, 25]	89.5	**89.2** [88, 90]	59.3 [58, 61]	59.8 [59, 61]
	Implied	3151	**9.2** [9, 10]	36.0 [33, 39]	24.9 [23, 27]	93.1	**92.2** [91, 93]	48.6 [46, 51]	59.5 [58, 61]
	Unstated	4992	**26.5** [25, 27]	31.2 [30, 32]	30.1 [29, 31]	90.1	**92.3** [92, 93]	81.6 [80, 83]	86.3 [85, 87]
PTB-XL (ext.)	Stated	193	**12.3** [11, 16]	18.3 [16, 21]	19.1 [16, 22]	97.4	**88.5** [84, 93]	74.1 [68, 80]	72.0 [66, 79]
	Implied	15	**9.3** [7, 26]	18.9 [6, 68]	20.3 [5, 40]	93.3	**93.3** [80, 100]	53.3 [27, 80]	60.0 [33, 87]
	Unstated	2676	**29.3** [28, 31]	30.6 [29, 32]	37.2 [36, 39]	94.1	**90.4** [89, 91]	88.3 [87, 90]	78.9 [77, 80]

Table 12: **Ablation of the L_{metric} loss and generator pretraining**, where all models were trained for 8 epochs. *w/o pretrained generator* keeps the pretrained text reasoner and randomly initializes the generator. We use the MIMIC-IV-ECG test split.

Metric	Full	w/o L_{metric}	w/o pretrained generator
Measurement error ↓			
PR (ms)	**18.2** ±1.7	19.0 ±1.9	21.3 ±1.6
QRS (ms)	**9.7** ±0.7	10.5 ±1.0	12.9 ±1.8
QT (ms)	**20.9** ±1.2	21.7 ±1.5	38.4 ±2.4
P duration (ms)	**17.5** ±2.8	17.7 ±2.7	21.6 ±2.9
T duration (ms)	**17.7** ±1.7	18.4 ±1.7	23.4 ±2.0
ST J+60 (μV)	**44.7** ±4.6	46.3 ±5.6	45.7 ±4.4
Cardiac axis (°)	**35.5** ±2.3	36.0 ±2.4	42.9 ±2.9
Mean RR (ms)	**12.6** ±1.7	13.3 ±1.8	157.4 ±8.8
SDNN (ms)	**26.0** ±3.6	26.6 ±3.5	194.4 ±7.6
Real-to-synthetic AUPRC (%) ↑			
All 12 labels	**73.0** ±2.3	71.7 ±2.4	17.8 ±1.4
Morphology (7)	**67.0** ±3.7	65.3 ±3.7	19.2 ±2.1
1st-degree AV block	**43.8** ±11.3	38.5 ±11.5	3.3 ±0.9
LBBB	**75.2** ±15.7	66.2 ±16.0	15.4 ±10.4
RBBB	97.1 ±1.6	**98.2** ±1.3	11.3 ±2.5
Left axis deviation	66.5 ±6.4	**68.1** ±6.6	27.3 ±4.8
LVH	**61.9** ±8.1	59.2 ±8.1	28.2 ±6.5
MI	62.2 ±6.4	**63.2** ±6.6	23.8 ±4.8
Ischemia	62.3 ±6.7	**63.8** ±6.4	24.9 ±4.6
Rate/rhythm (4)	**86.2** ±2.4	85.4 ±2.5	17.1 ±1.8
Sinus bradycardia	90.2 ±4.5	**90.3** ±4.3	19.6 ±3.2
Sinus tachycardia	**95.8** ±1.8	94.9 ±2.2	21.6 ±4.6
Atrial fibrillation	**76.5** ±5.0	76.1 ±5.3	18.5 ±2.9
PVC	**82.2** ±6.1	80.2 ±6.1	8.7 ±3.5
Normal	**62.2** ±6.2	62.2 ±6.0	11.2 ±1.5

between leads (e.g., if limb leads lag behind chest leads). Thus, even spurious generators such as a flatline or a fixed normal ECG can demonstrate the lowest MSE and MAE, as shown in Table 13.

C.8 Inference Efficiency

Table 14 compares the generation speed of CARP, SE-Diff and DiffuSETS. Note that text conditioning for both baselines is excluded from their timings: SE-Diff additionally requires an LLM call to generate its experience-retrieval report, and DiffuSETS requires an external text-embedding model (`ada-002`). CARP encodes the report with its own backbone and the associated latency is included in the timing. We use one NVIDIA A100 80GB GPU for running the latency experiment. SE-Diff and DiffuSETS run 1,000 sampling steps, whereas CARP runs 35 steps. At each batch size, we record the wall-clock time, and report the mean over three batches.

Table 13: **Pointwise error metrics such as MAE can reward spurious generators that ignore the text.** Pointwise error (μV) with the real ECG, mean $\pm$ 95% CI. Flatline (all-zero) and Fixed NSR (one median normal ECG) ignore the text.

	MIMIC-IV		PTB-XL (ext.)	
Method	MAE	RMSE	MAE	RMSE
Flatline	$83.9_{\pm 5.6}$	$160.3_{\pm 9.6}$	$120.5_{\pm 1.6}$	$191.9_{\pm 2.3}$
Fixed NSR	$115.6_{\pm 5.2}$	$216.1_{\pm 9.9}$	$147.8_{\pm 1.5}$	$250.7_{\pm 2.7}$
SE-Diff	$109.4_{\pm 6.1}$	$213.9_{\pm 10.5}$	$147.9_{\pm 1.9}$	$252.2_{\pm 3.1}$
DiffuSETS	$113.4_{\pm 6.0}$	$223.2_{\pm 10.2}$	$148.8_{\pm 1.8}$	$258.0_{\pm 3.0}$
CARP	$121.3_{\pm 7.4}$	$246.9_{\pm 12.5}$	$166.5_{\pm 2.1}$	$286.5_{\pm 3.1}$

Table 14: **Inference throughput of the model for generating one batch of 10-second, 12-lead ECG**, under different batch sizes. SE-Diff and DiffuSETS are run for 1000 diffusion steps, and CARP for 35 steps. Baseline numbers do not include time taken for LLM calls (SE-Diff) or embedding (SE-Diff, DiffuSETS).

	SE-Diff		DiffuSETS		CARP	
Batch size	s / batch	s / sample	s / batch	s / sample	s / batch	s / sample
32	27.79	0.869	29.62	0.926	4.84	0.151
128	29.77	0.233	31.17	0.244	18.43	0.144
256	32.75	0.128	35.23	0.138	38.15	0.149
512	37.07	0.072	44.20	0.086	73.89	0.144

C.9 Inference Parameters

We defer to Cosmos 3 (NVIDIA et al., 2026) for the choice of inference techniques, and use its native rectified-flow sampler and noise schedule without modification. Starting from Gaussian noise, we integrate the flow with a second-order UniPC solver and apply classifier-free guidance, using the null text as the unconditional branch (the text is dropped with probability 0.1 during training). Table 15 lists the settings.

Table 15: Inference settings for CARP.

Setting	Value
Solver	UniPC (B(h) = 2 variant), order 2
Sampling steps	35
Classifier-free guidance scale	1.5
Noise schedule	Karras, $\sigma \in [0.147, 200]$
Flow shift	1.0
Initial noise	$\epsilon \sim \mathcal{N}(0, I)$