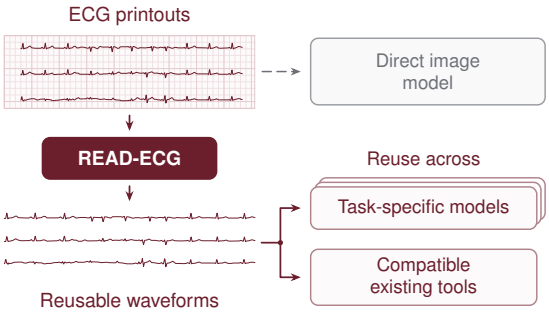
READ-ECG: ECG PRINTOUT DIGITISATION VIA STRUCTURED VISUAL SEQUENCE REGRESSION FOR CARDIOVASCULAR ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electrocardiograms (ECGs) are a primary cardiovascular diagnostic modality, but recordings available only as scanned or photographed printouts cannot be directly processed by signal-based AI tools. Direct image analysis entangles signal extraction with clinical interpretation, while multi-stage digitisation pipelines rely on brittle segmentation steps. We formulate ECG printout digitisation as a structured visual sequence regression problem within a decoupled *Image* $\rightarrow$ *Signal* $\rightarrow$ *Analysis* pipeline. We introduce *READ-ECG* (Reconstruction and Extraction via Attention-based Decoding of ECGs), an end-to-end framework that bypasses intermediate trace segmentation, employing a topology-aware attention-based decoder to directly regress spatial features into continuous multi-lead waveforms. To support future quality-control workflows, we incorporate aleatoric uncertainty modelling to identify potentially unreliable reconstructions. Evaluated across diverse benchmarks, READ-ECG achieves high-fidelity reconstruction on the combined PhysioNet evaluation (PCC: 0.998, SNR: 28.83 dB) and can adapt to severe real-world visual artefacts using only 81 real training images. Across three PTB-XL classification task families, a residual 1D CNN trained on original signals retains a mean AUROC of 0.927 on READ-ECG reconstructions. Evaluation with two pre-trained ECG foundation models further suggests that the digitised signals retain diagnostic information. By decoupling signal "reading" from "interpretation", READ-ECG provides reusable waveforms to the signal-based cardiac AI ecosystem.



Model		AUROC / AP↑
Image models		*Printout*
ResNet-18		.9106 / .5872
ConvNeXt-Tiny		.9129 / .6182
Signal models	*Original*	**READ-ECG**
Residual 1D CNN	.9269 / .6075	**.9270 / .6082**
ECG-JEPA	.9036 / .6142	**.9037 / .6126**
ST-MEM	.9099 / .5843	**.9090 / .5827**

Figure 1: Left: Two approaches to analysing ECG printouts. Direct image models learn downstream tasks from the image. READ-ECG instead reconstructs waveforms that can be reused by downstream signal models and compatible pretrained tools. **Right:** Classification on 1,689 clean PTB-XL ECGs held out from READ-ECG's training and validation sets: AUROC and AP are each averaged equally across diagnosis, rhythm and form, then over three seeds. Original and READ-ECG denote matched scan-visible reference and reconstructed waveforms, evaluated with the same fitted signal model. ECG-JEPA and ST-MEM use frozen encoders with MLP probes. Task-wise scores, paired confidence intervals and model details are in Section A.6.

Project page and demo: https://anonymous.4open.science/r/read-ecg-review

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.

1 INTRODUCTION

The 12-lead electrocardiogram (ECG) remains a standard non-invasive diagnostic modality in both emergency and inpatient care (Surawicz et al., 2009; Rautaharju et al., 2009; Hancock et al., 2009; Wagner et al., 2009). Recent advances in artificial intelligence have expanded the information that can be extracted from ECG signals, supporting automated disease detection and prediction of future health risks (Attia et al., 2019; Raghunath et al., 2020). Nevertheless, accurate interpretation of atypical ECG patterns remains challenging because diagnostically relevant changes may be subtle, localised and difficult to distinguish on paper tracings. The format in which ECGs are available creates a further barrier to automated analysis. In pre-hospital settings such as ambulances, recordings are frequently shared as printouts or photographs of them rather than raw digital signals, limiting compatibility with signal-native AI models. Together, these challenges motivate robust digitisation methods that reconstruct high-fidelity signals from ECG printout images, enabling automated analysis to support clinical interpretation.

The common approach has been to develop end-to-end models that perform analysis directly on the image, a paradigm we call *Image* $\rightarrow$ *Analysis* (Barros et al., 2024; Bodagh et al., 2025). While convenient, it forces a single model to learn two vastly different tasks: 1) ignoring complex 2D image artefacts such as gridlines, scanner noise, and physical degradation, and 2) identifying subtle 1D diagnostic patterns embedded within the waveform's morphology. This entanglement can lead to models learning irrelevant visual correlations and limit generalisation, while producing task-specific predictions rather than a reusable waveform.

This challenge is further highlighted by the recent progress in automated ECG analysis, driven by foundation models such as ECG-JEPA (Kim, 2026) and ST-MEM (Na et al., 2024) that operate on ECG waveforms. These models, pre-trained on large collections of digital ECG records, have demonstrated strong performance across a wide range of interpretive tasks. However, their reliance on digital signals creates a paradigm mismatch. Many recent analytical tools in cardiology require a digital signal input, so applying them to images of ECG printouts first requires reconstructing the underlying signals.

In this work, we bridge this domain mismatch by formalising a decoupled *Image* $\rightarrow$ *Signal* $\rightarrow$ *Analysis* pipeline (Fig. 1, diagram). We propose that the initial step should be the high-fidelity reconstruction of the underlying digital signal. By decoupling the problem of *reading* the ECG from the problem of *interpreting* it, we can employ the strengths of specialised models for each task. The digitised signal can then serve as a reusable, artefact-free, standardised input for the ecosystem of signal analysis algorithms (Singh & Krishnan, 2023), from clinical parameter extraction (Makowski et al., 2021; Pan & Tompkins, 1985; Bazett, 1920) to advanced classification (Hannun et al., 2019).

Unlike existing digitisation methods that rely on brittle, heuristic-driven multi-stage pipelines, such as grid removal, mask segmentation, and skeletonisation, we reframe ECG printout digitisation as a structured visual sequence regression task. We introduce READ-ECG (Reconstruction and Extraction via Attention-based Decoding of ECGs), an end-to-end framework that treats an ECG scan as a structured visual document containing continuous multi-channel sequences. By integrating a topology-aware attention decoder, READ-ECG uses explicit temporal and lead queries to directly read spatial visual evidence into waveform samples. This reduces reliance on hand-engineered geometry, directly optimising the multi-lead clinical signal.

Our contributions are as follows:

1. We formulate ECG printout digitisation as direct, structured visual sequence regression. We introduce READ-ECG, an end-to-end framework that directly reconstructs digital signals from scanned images, integrating aleatoric uncertainty modelling for uncertainty-based rejection.
2. We propose the decoupled *Image* $\rightarrow$ *Signal* $\rightarrow$ *Analysis* paradigm. By comparing our framework against direct image classifiers, we examine the trade-offs between coupling visual processing with interpretation and producing a reusable waveform for downstream analysis.
3. We provide a broad evaluation across synthetic, out-of-distribution, and real-world artefact-heavy domains. We show our reconstructed signals achieve high fidelity and preserve diagnostic information used by independent tools and signal foundation models.

2 Related Work

Classical digitisation. Early ECG digitisation systems treated printouts as geometric image-processing problems. Typical pipelines removed background grids and extracted the waveform trace with operations, such as thresholding, Hough transforms, or morphological skeletonisation, then converted the trace pixels into voltage-time samples (Wang et al., 2024; Wu et al., 2022; Garg et al., 2012; Ravichandran et al., 2013; Baydoun et al., 2019). These methods are interpretable and can work well for clean, standardised images, but they depend strongly on assumptions about alignment, grid colour, paper quality, calibration marks, and trace connectivity. Consequently, they are vulnerable to common degradations such as low contrast, stains, broken traces, overlaid text, folded paper, and camera variability.

Segmentation pipelines. To improve visual robustness, recent deep learning methods replace hand-crafted trace extraction with neural segmentation, often using U-Net style architectures to simultaneously remove grids and perform pixel-wise waveform segmentation from noisy ECG images (Li et al., 2020; Mishra et al., 2021; Shivashankara et al., 2024; Demolder et al., 2025; Krones et al., 2024). These methods substantially reduce the need for manually designed low-level image filters, but they still preserve the same multi-stage pipeline structure, first predicting an intermediate visual representation, such as trace mask, skeleton, or lead segment, and then recover numerical samples through heuristic post-processing, such as column-wise pixel averaging or graph-based tracing. This separation is brittle and introduces errors when masks are fragmented, traces overlap, text intersects the waveform, lead boundaries are ambiguous, or calibration assumptions fail. In contrast, our approach directly optimises the final multi-lead signal, bypassing the need for intermediate segmentation and hand-engineered rules.

Direct image analysis. An alternative paradigm bypasses digitisation and trains image classifiers to predict clinical labels directly from ECG printouts (Barros et al., 2024; Bodagh et al., 2025). This strategy is attractive when the target task is fixed, because the model can be trained end-to-end from image to diagnosis. While this end-to-end approach bypasses reconstruction, it answers a fundamentally different, narrower question than digitisation and couples visual processing with clinical interpretation. The model must simultaneously suppress 2D visual artefacts while learning subtle 1D clinical morphology, which can encourage reliance on spurious visual correlations. Its predictions remain tied to the tasks chosen during training, even when one classifier learns several tasks. A reconstructed waveform can instead be reused for interval measurement, physician overread, signal processing algorithms (Pan & Tompkins, 1985), and signal-native foundation models (Kim, 2026; Na et al., 2024) without changing the image reader. This distinction motivates our comparison between direct *Image* $\rightarrow$ *Analysis* baselines and our decoupled *Image* $\rightarrow$ *Signal* $\rightarrow$ *Analysis* pipeline.

Structured visual decoding. The value of high-fidelity digitisation has increased as digital ECG analysis has advanced. Classical signal-processing algorithms remain widely popular for clinical parameter estimation and interval measurement (Pan & Tompkins, 1985), while deep residual networks have reached cardiologist-level performance (Hannun et al., 2019; Liang et al., 2025). Recent signal foundation models include ECGFounder (Li et al., 2025) and the ECG-text models D-BETA (Pham Hung et al., 2025) and MELP (Wang et al., 2025). All require a digital waveform as input. To bridge the gap between ECG images and these modern tools, we reframe ECG digitisation away from pixel segmentation and towards structured visual sequence decoding. ECG printouts form a structured visual document: the image contains a known arrangement of continuous multi-channel sequences, and the desired output is not a class label or binary mask but calibrated waveform samples. This perspective connects ECG digitisation to a broader class of detection problems, specifically image-to-sequence (Shi et al., 2017). For example, SLPM (Yu et al., 2026) also predicts waveform points from cropped lead images. In related image-to-sequence tasks, including optical and handwritten text recognition, attention-based encoder-decoder networks attend to visual feature maps to generate ordered outputs without explicit bounding boxes, segmentation, or layout heuristics (Singh & Karayev, 2021; Li et al., 2023). READ-ECG adopts this principle for ECGs by using topology-aware temporal and lead queries to read spatial image evidence into continuous waveform predictions. Using this formulation, attention provides an end-to-end bridge between the visual layout of a clinical document and the sequential nature of a time-series signal.

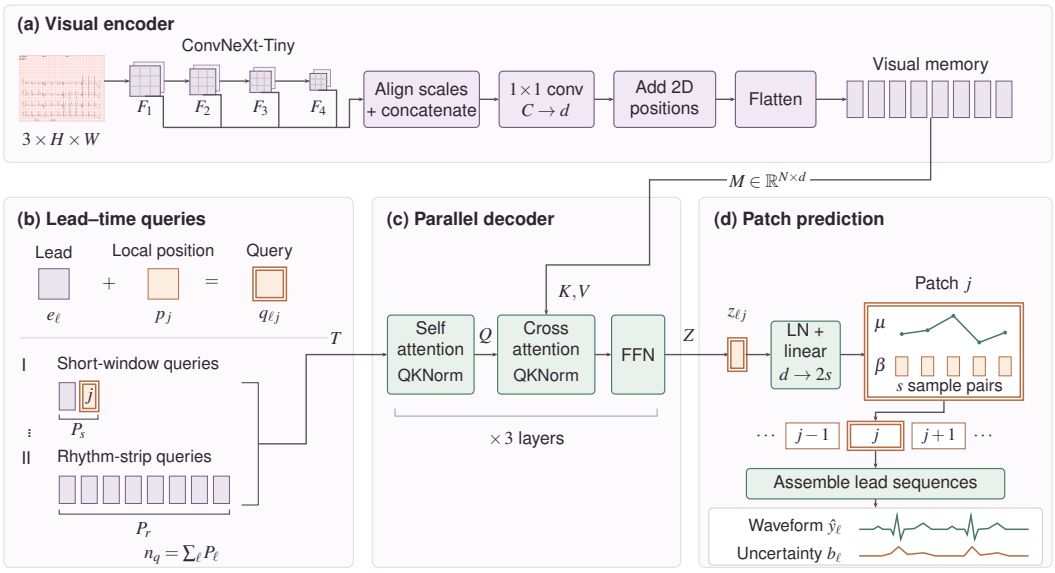


Figure 2: READ-ECG architecture. (a) Four ConvNeXt stages are aligned to stage-3 resolution, concatenated, projected, and combined with 2D positional embeddings to form visual memory M. (b) Query $q_{\ell j}$ combines a learned lead embedding and local patch position; the lead-II rhythm strip uses four times as many queries as each short window. (c) Three decoder layers update queries through self-attention and read M through cross-attention, both with QKNorm. (d) Each decoded query $z_{\ell j}$ predicts s waveform samples and s raw scales; ordered patches form lead sequences and sample-wise uncertainties. Batch axes and internal residual paths are omitted; waveforms are schematic.

3 Methodology

3.1 Model Architecture

The proposed READ-ECG architecture (Fig. 2) is an end-to-end framework designed to perform structured visual sequence regression directly from an ECG image. Rather than treating the image as a grid of pixels to be segmented, we treat it as a clinical document containing a known arrangement of signals. The complete configuration is given in Table A1.

Visual encoding. The encoder uses a ConvNeXt-Tiny vision backbone (Liu et al., 2022) (Fig. 2a). Accurately reconstructing an ECG requires both extracting highly localised, high-frequency trace details while maintaining global spatial context required to locate the 12 distinct leads across the ultra-wide input image. This motivates the use of multi-scale feature fusion, which combines high-resolution features from early stages with coarser features from later stages. The fused spatial features are then projected to a latent dimension (d_{model}) and injected with 2D sinusoidal positional embeddings so that subsequent attention layers can utilise both the content and location of the image patches.

Topology-aware decoding. To read the visual evidence into continuous waveforms, we use a parallel, non-autoregressive Transformer decoder (Vaswani et al., 2017) (Fig. 2b–c). Because standard ECG printouts follow a strict spatial-temporal layout, we can provide the decoder with topology-aware target queries that combine learnable lead embeddings with 1D temporal positional embeddings, similar in spirit to the lead and position embeddings used for digital ECG patches in ST-MEM (Na et al., 2024). These queries correspond to specific leads and time regions in the 3×4 plus Rhythm Strip ECG layout, and they direct the cross-attention mechanism to attend to the relevant image patches, as illustrated by the learned maps in Fig. A4.

Patch-wise regression. Predicting 1D continuous signals point-by-point is computationally expensive and can lead to unstable training. To improve both efficiency and training stability, READ-ECG performs patch-wise decoding. The output head (Fig. 2d) predicts a small sequence of signal values per query vector, rather than a single point, which reduces the effective sequence length and encourages local coherence in the reconstructed waveform. Among the tested settings, a patch size of

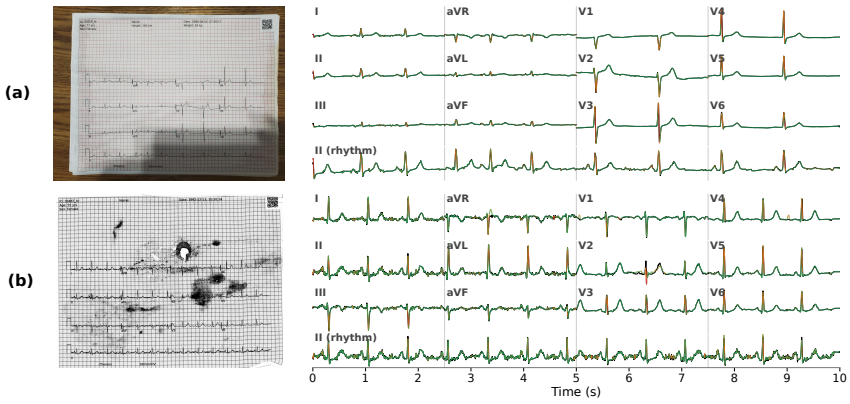


Figure 3: Kaggle-tuned reconstructions on held-out printouts: ground truth in black; predictions coloured by uncertainty (green: low; red: high). (a) Shadows and wrinkles (20.45 dB SNR), with minor uncertainty at sharp QRS peaks and aliased noise. (b) Mould stains and holes (13.22 dB SNR), with high uncertainty in unreadable traces, including the hallucinated second QRS complex and subsequent T wave obscured by a stain in V2.

5 achieved the highest SNR (Table A3). Single-point prediction was less stable, while larger patches required each 256-dimensional query to represent more output values. The output head predicts both the mean waveform and, when uncertainty modelling is enabled, an aleatoric scale parameter for each sample.

Attention stabilisation. As the model learns to reconstruct the signal, the cross-attention weights naturally sharpen towards specific image regions. In standard Transformers, this can lead to "attention collapse," where weights degrade into delta functions, causing vanishing gradients and diverging training. To prevent this, we integrate widely-used Query-Key Normalisation (QKNorm) (Henry et al., 2020) into both the self- and cross-attention layers of the decoder. In our experiments, QKNorm stabilised training at high reconstruction fidelity. Empirically, we found that omitting QKNorm causes eventual attention collapse and training divergence after approximately 25 epochs (Table A4 and Figs. A2 and A3). Furthermore, the stability provided by combining QKNorm with patch-wise decoding eliminated the need for traditional Transformer warm-up schedules (Vaswani et al., 2017).

3.2 Datasets and Preprocessing

We evaluate READ-ECG on in-distribution signals, unseen signal morphologies, and real-world visual degradation. To increase training-set size and cross-cohort variability, our primary paired image-signal dataset combines data from six independent clinical archives (PTB-XL, Georgia, CPSC2018, CPSC-Extra, PTB, and Incart (Perez Alday et al., 2020)) hosted on PhysioNet (Goldberger et al., 2000). Following Coppola et al. (2024), signals are resampled, standardised using fixed per-lead means and standard deviations from the combined PhysioNet training set, and segmented for the common 3×4 layout (2.5s per lead) with a 10s lead-II rhythm strip. Conversion back to physical units is described in Section A.1. We synthetically generate artefact-free image-signal pairs using ECG-Image-Kit (Shivashankara et al., 2024) at a fixed paper speed and gain (25 mm/s and 10 mm/mV). At inference, we assume the same layout, paper speed and gain rather than estimating them from the image. We render images at an ultra-wide 512×2560 input resolution to preserve morphological details and prevent time-axis aliasing, as the amplitude axis is more robust to downsampling. Standard visual augmentations (colour jitter, Gaussian noise) are applied to prevent overfitting. Representative augmented images and record-level split counts are included in the supplementary material (Fig. A1 and Table A2). The PTB-XL test set serves as our primary in-distribution benchmark because it is large, well curated, and provides a common signal source for comparison with prior digitisation models.

To verify that our attention decoder learns to read spatial image features rather than memorising training-distribution waveforms, we evaluate it zero-shot on 7,422 synthetic ECG images generated from the MIMIC-IV-ECG matched subset (Gow et al., 2023), specifically the studies contained in the first ten source RECORDS manifests. This tests the model's capacity to reconstruct unseen signal morphologies and pathologies under a controlled rendering process without further training.

Finally, to assess robustness to real-world visual degradation, we utilise the recent Kaggle ECG Digitisation dataset (Reyna et al., 2025). This dataset consists of 8,793 images from 977 patient records captured under highly unconstrained conditions, including mobile phone photos, varied lighting, stains, monitor artefacts, and severe paper damage. To standardise the inputs, we apply a standard perspective correction to these images, and no images are discarded by this step. We create record-level training, validation, and held-out test partitions, keeping all images from each record in one split. The model is fine-tuned by mixing real and synthetic images 1:1 across each epoch to retain learned signal morphology and limit overfitting. This evaluation tests adaptation to new visual artefacts and imaging conditions, whereas the zero-shot MIMIC-IV evaluation tests generalisation to unseen signals under controlled rendering.

3.3 Training Protocols and Loss Functions

Formulating digitisation as sequence regression allows us to optimise the encoder, decoder, and output head jointly in a single training stage. To optimise reconstruction fidelity while simultaneously quantifying the reliability of the digitisation, READ-ECG can model aleatoric uncertainty (Kendall & Gal, 2017). The output head predicts the signal mean (μ) and a scale parameter $b \geq 0$ per sample, optimising a Laplace negative log-likelihood (NLL) loss:

$$L_{\text{total}} = \frac{1}{N} \sum_{i} \left(\log(2b_i + \epsilon) + \frac{|\mu_i - y_i|}{b_i + \epsilon} \right), \tag{1}$$

where y_i denotes the ground-truth signal value, N is the number of output samples, and ϵ is a small constant for numerical stability. This outputs an estimate of heteroscedastic input-dependent uncertainty, which can help prioritise reconstructions for human review. When uncertainty modelling is disabled, the model optimises a standard mean absolute error (MAE) loss.

We initialise both the ConvNeXt-Tiny encoder and Transformer decoder, along with learnable lead embeddings, from scratch to allow for maximum flexibility in learning the unique visual and temporal structure of ECG printouts. Both 2D and 1D positional embeddings are initialised with sinusoidal functions to provide a stable temporal reference for the decoder. The reconstruction head and uncertainty head share the same visual and decoder features, with the uncertainty branch trained only through the Laplace likelihood above. Thus, uncertainty is not fitted as a post-hoc calibration model, but is learned as an input-dependent estimate of local reconstruction reliability. READ-ECG was implemented in PyTorch (Paszke et al., 2019) and trained across a 4× NVIDIA H200 GPU cluster using automatic mixed precision (BFloat16).

READ-ECG was trained with the AdamW optimiser (Loshchilov & Hutter, 2019) ($lr = 2 \times 10^{-4}$, batch size 16) and cosine annealing over 50 epochs. For visual robustness evaluations, we fine-tune with the same reconstruction objective on full-data and few-shot Kaggle subsets. For the downstream comparison, the image classifiers and residual 1D CNN are fitted end-to-end with AdamW and binary cross-entropy. For ECG-JEPA and ST-MEM, we freeze the encoders and train only their task heads. Checkpoints are selected on validation data; downstream results average three independently fitted seeds; full training schedules and probe details are in Section A.6.

3.4 Evaluation Protocol

We evaluate four questions: (1) how accurately READ-ECG reconstructs waveforms and clinical intervals, (2) whether reconstructions preserve downstream performance, (3) whether uncertainty identifies unreliable reconstructions, and (4) how well the model adapts to real degraded ECG images.

Signal and clinical fidelity. To test continuous sequence regression against intermediate mask prediction, we evaluate the raw signal recovery quality. PTB-XL provides the common signal source for comparison with ECG Digitiser (Krones et al., 2024), Demolder et al. (2025), and SLPM (Yu et al., 2026). To complement the global reconstruction metrics, we also assess preservation of clinically relevant parameters. We independently extract heart rate, QRS duration, and Bazett QTc from original and reconstructed lead-II signals using NeuroKit2 with prominence-based delineation (Makowski et al., 2021; Emrich et al., 2024; Bazett, 1920). We assess agreement using Bland–Altman analysis and single-measure absolute-agreement ICC with extraction details given in Section A.5.3. Finally, we perform a diagnostic superclass breakdown on the PTB-XL test set to ensure the digitisation does not disproportionately degrade specific pathologies, like ST and T wave abnormalities.

Table 1: Reconstruction across models and test sets. SNR is in dB; protocols are specified below.

Model / test set	PCC↑	SNR↑
PTB-XL comparison		
ECG Digitiser[†]	0.965	14.94
Demolder et al.[‡]	0.981	17.64
SLPM[‡]	0.970	14.66
READ-ECG (Ours)	**0.999**	**30.50**
READ-ECG: other test sets		
Combined PhysioNet	**0.998**	**28.83**
MIMIC-IV	**0.998**	**26.49**
Kaggle	**0.990**	**19.40**

Table 2: PTB-XL fidelity by diagnostic superclass (record counts in parentheses). MAE is normalised; fidelity remains high across pathologies.

Class (records)	MAE↓	SNR↑
Normal (963)	0.0091	31.24
Myocardial infarction (550)	0.0117	30.40
ST/T change (521)	0.0130	30.04
Conduction disturbance (496)	0.0151	30.49
Hypertrophy (262)	0.0165	29.20

For the PTB-XL comparison, each method is evaluated on clean images in a visual style it was developed to handle. [†]Our ECG Digitiser reproduction on clean renderings. [‡]Published clean-image results. See Section A.5.1 for training and evaluation details.

Downstream utility. On clean PTB-XL, we evaluate multilabel diagnosis (5 labels), rhythm (12), and form (19). The 1,689 test ECGs come from patients absent from READ-ECG's PTB-XL training and validation sets. We report macro AUROC and average precision (AP) for each task family, then average equally across families and three independently fitted seeds (Fig. 1, table). The image baselines are shared three-task ResNet-18 (He et al., 2016) and ConvNeXt-Tiny (Liu et al., 2022) classifiers trained on $512 \times 2{,}560$ printouts. ConvNeXt-Tiny uses the same visual backbone architecture as READ-ECG.

For waveform analysis, we train a shared three-task residual 1D CNN with four temporal residual stages. Its input contains only the waveform visible in the printout: at 250 Hz, short leads occupy their printed 2.5-second windows in a $12 \times 2{,}500$ tensor, unobserved intervals are zero-filled, and lead II retains its full 10-second rhythm strip. We evaluate the same fitted CNN on these reference signals and without retraining on READ-ECG reconstructions. To test compatibility with pretrained signal models, we freeze ECG-JEPA (Kim, 2026) and ST-MEM (Na et al., 2024). For each task family, we train an MLP head with one hidden layer (width 256) on original-signal features, then evaluate it on READ-ECG reconstructions.

Uncertainty and rejection. To evaluate uncertainty calibration, we compute the Area Under the Calibration Error curve (AUCE) by integrating the absolute difference between nominal and empirical Laplace coverage intervals across 50 evenly spaced confidence levels from 0.01 to 0.99. For future quality-control workflows, we average predicted scales over each image's unpadded samples and evaluate MAE and SNR on the Kaggle test set as the highest-uncertainty images are rejected. This tests whether uncertainty can identify severely degraded, unreliable scans for human review.

Few-shot adaptation. To quantify the model's ability to adapt to severe real-world visual degradation, we perform a few-shot fine-tuning analysis on the Kaggle dataset. Starting from the synthetic-trained model, we fine-tune on varying fractions of the Kaggle training set using three random seeds per fraction, and evaluate on the held-out Kaggle test set. This synthetic-only versus few-shot curve tests how efficiently the model transfers image-to-signal reconstruction learned from synthetic data to real-world visual artefacts.

4 Experiments and Results

4.1 Reconstruction Fidelity and Generalisation

On clean PTB-XL images, READ-ECG reaches 0.999 PCC and 30.50 dB SNR (Table 1), with consistently strong performance across individual limb and precordial leads (Table A6). High fidelity is maintained across the combined PhysioNet evaluation (PCC: 0.998, SNR: 28.83 dB) and on renderings of unseen MIMIC-IV signals, where the model reaches 26.49 dB SNR without further training. Real Kaggle images present a large visual domain shift and require adaptation: full-data fine-tuning yields 19.40 dB SNR on the record-separated held-out test set.

To verify that READ-ECG does not strongly bias towards "healthy" waveforms, we evaluated reconstruction fidelity across the five diagnostic superclasses of the PTB-XL test set (Table 2). The model maintains high fidelity across all pathologies (SNR $\sim$ 30 dB), including ST/T changes (STTC) and Myocardial Infarctions (MI), where subtle waveform morphology is critical for localisation and interpretation. This suggests that READ-ECG does not disproportionately degrade abnormal waveforms and performs well across diverse pathologies.

Bland–Altman analysis of core diagnostic intervals (Fig. 4) on the PTB-XL test set confirms near-zero mean bias across parameters (HR: approximately 0.00 bpm; QRS: $+0.23$ ms; Bazett QTc: $+0.42$ ms), with Intraclass Correlation Coefficient values indicating excellent agreement (HR: 0.999; QRS: 0.992; QTc: 0.983) according to Koo & Li (2016). Together, these results support the preservation of ECG timing measurements through digitisation.

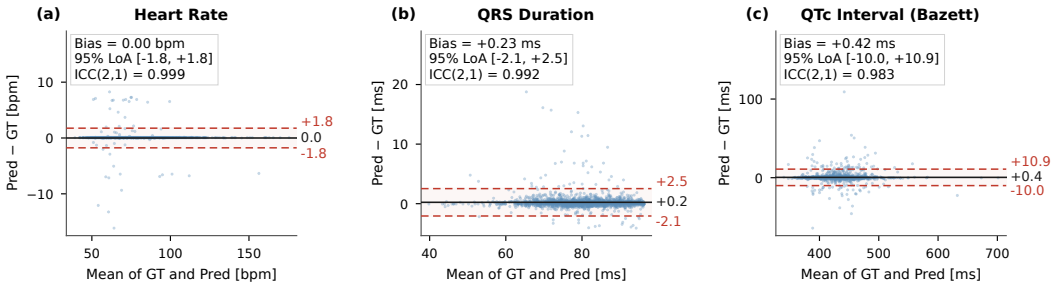


Figure 4: Bland–Altman agreement for (a) heart rate, (b) QRS duration, and (c) QTc on 2,194 PTB-XL lead-II rhythm strips (paired counts in Section A.5.3). Solid/dashed lines: mean (Pred − GT) / 95% limits of agreement ($\pm$1.96 SD). Near-zero biases and ICC(2,1) of 0.999/0.992/0.983 support preserved clinical timing.

4.2 Downstream Utility and Paradigm Comparison

Across the three PTB-XL tasks, READ-ECG followed by the residual 1D CNN reaches a mean AUROC of 0.9270, compared with 0.9106 for image ResNet-18 and 0.9129 for image ConvNeXt-Tiny (Fig. 1, table). The paired mean AUROC differences are $+0.0164$ $[+0.0086, +0.0241]$ relative to ResNet-18 and $+0.0141$ $[+0.0064, +0.0216]$ relative to ConvNeXt-Tiny (95% intervals). Both paired AUROC comparisons favour the READ-ECG/CNN pipeline on this test set.

We also evaluated each signal model on original and READ-ECG waveforms without refitting to assess the effect of reconstruction. The paired changes in mean AUROC (READ-ECG minus original) are $+0.00013$ $[-0.00014, +0.00044]$ for the CNN, $+0.00013$ $[-0.00006, +0.00033]$ for ECG-JEPA, and -0.00090 $[-0.00214, +0.00027]$ for ST-MEM. All three changes are small, and each 95% interval includes zero. Task-wise scores, AP, and full paired analyses are in Section A.6.

4.3 Uncertainty Estimation and Quality Control

Prioritising potentially unreliable reconstructions could support future clinical quality-control workflows. To assess the practical utility of our uncertainty modelling, we evaluated how its estimates relate to local errors and image-level reconstruction quality.

Across unpadded time points in the 6,100 combined PhysioNet test segments, normalised scales and absolute errors correlate strongly (Pearson $r = 0.81$), with AUCE = 0.033. As visualised in Fig. 3, the model assigns higher predicted scale to several unreliable waveform regions.

To assess potential quality-control use, we averaged predicted scales over each image's unpadded samples and performed an uncertainty-based rejection analysis on the held-out Kaggle test set (Fig. 5). By rejecting the 25% most uncertain images, retained-set MAE decreased from 8.50 to 6.37 μV and mean image-level SNR improved from 18.86 to 19.71 dB. The uncertainty-based MAE rejection curve closely follows the oracle obtained by rejecting images according to their true MAE (6.33 μV at 25% rejection), while both metrics improve relative to random rejection. This suggests that READ-ECG uncertainty could help identify unreliable reconstructions for review. Additional PhysioNet



Figure 5: Uncertainty-based rejection on 891 held-out Kaggle images (99 records). Retained MAE and SNR improve over random rejection (grey); model lines (blue) follow the oracles (green).

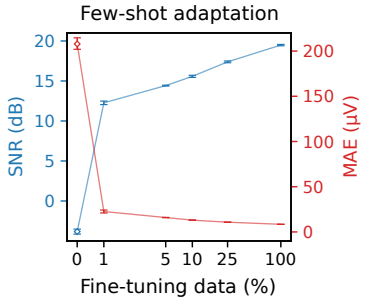


Figure 6: With 1% of Kaggle (nine records; 81 images), SNR improves by 16.1 dB. Bars: SD; 0%: 95% CI.

sample- and lead-level retention curves are reported in Fig. A6, while Fig. A7 compares the predicted uncertainty distributions across domains.

4.4 Few-Shot Real-Image Adaptation

To address the domain gap between synthetic training data and real-world images, we evaluated the sample efficiency of READ-ECG on the Kaggle dataset (Fig. 6). When evaluated zero-shot, the synthetic-trained model fails to generalise to the severe visual artefacts present in the Kaggle images (SNR: -3.83 dB). However, fine-tuning on 1% of the Kaggle training data, with a separate 48-record validation set for checkpoint selection, results in a marked performance improvement, increasing the SNR by nearly 16 dB to 12.3 ± 0.2 dB (MAE: 22.6 μV). Performance continues to improve with diminishing gains as more data are added, reaching 14.4 dB at 5% and 17.4 dB at 25% of the training set, before plateauing at 19.5 dB (MAE: 8.4 μV) with full-data fine-tuning. This sample efficiency suggests that synthetic training establishes the image-to-signal mapping, which can then be adapted to real-world visual conditions through limited fine-tuning. The repeated-capture and failure examples in Figs. A8 and A9 illustrate both the robustness gained through adaptation and the remaining limitations under severe degradation.

5 Discussion and Conclusion

To enable signal-based analysis of ECG printout images, we introduced READ-ECG, a decoupled *Image* $\rightarrow$ *Signal* $\rightarrow$ *Analysis* framework that formulates digitisation as structured visual sequence regression. By digitising these recordings with high fidelity and demonstrating their compatibility with downstream tools and state-of-the-art foundation models, we establish a common waveform input for diverse downstream analysis.

This paradigm allows the visual-processing stage to be shared across diagnostic tasks. READ-ECG extracts a reusable signal once, allowing different downstream tools to analyse it without repeating the digitisation step. The built-in aleatoric uncertainty estimates could help future quality-control systems prioritise potentially unreliable scans for human review.

While reliance on synthetic training data reflects the field's scarcity of public annotated archives, our few-shot adaptation experiments suggest that limited real-image supervision can help address this constraint. The current model assumes the 3×4 plus lead-II layout with fixed lead durations and does not explicitly read printed gain or paper speed. The same architecture can be retrained for alternative layouts by redefining the target queries. Future work will extend this approach to more diverse ECG layouts and recording scales.

AI USE STATEMENT

We used generative AI tools for code implementation and debugging, and occasionally for suggestions on experimental variations. We also used them to improve writing clarity and grammar, refine figures, and review the manuscript for unclear presentation or misunderstandings. We did not use generative AI to formulate the main hypotheses or develop the research methodology. All AI-assisted work was reviewed by the authors, and we take responsibility for the final content of the paper.

REFERENCES

Zachi I. Attia, Suraj Kapa, Francisco Lopez-Jimenez, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nature Medicine*, 25(1): 70–74, 2019. doi: 10.1038/s41591-018-0240-2.

Andrew Barros, Ian German Mesner, N Rich Nguyen, and J Randall Moorman. Age prediction from 12-lead electrocardiograms using deep learning: a comparison of four models. *Physiological Measurement*, 45(8):08NT01, 2024. doi: 10.1088/1361-6579/ad6746.

Mohammed Baydoun, Lise Safatly, Ossama K Abou Hassan, Hassan Ghaziri, Ali El Hajj, and Hussain Isma'eel. High precision digitization of paper-based ecg records: a step toward machine learning. *IEEE journal of translational engineering in health and medicine*, 7:1–8, 2019.

H. C. Bazett. An analysis of the time-relations of electrocardiograms. *Heart*, 7:353–370, 1920.

Neil Bodagh, Kyaw Soe Tun, Adam Barton, et al. GenECG: a synthetic image-based ecg dataset to augment artificial intelligence-enhanced algorithm development. *BMJ Health & Care Informatics*, 32(1):e101335, 2025. doi: 10.1136/bmjhci-2024-101335.

Edoardo Coppola, Mattia Savardi, Mauro Massussi, et al. HuBERT-ECG: a self-supervised foundation model for broad and scalable cardiac applications. *medRxiv*, 2024. doi: 10.1101/2024.11.14.24317328.

Anthony Demolder, Viera Kresnakova, Michal Hojcka, et al. High precision ecg digitization using artificial intelligence. *Journal of Electrocardiology*, 90:153900, 2025. doi: 10.1016/j.jelectrocard.2025.153900.

Jonas Emrich, Andrea Gargano, Taulant Koka, and Michael Muma. Physiology-informed ECG delineation based on peak prominence. In *32nd European Signal Processing Conference (EUSIPCO)*, pp. 1402–1406, 2024.

Deepak Kumar Garg, Diksha Thakur, Seema Sharma, and Shweta Bhardwaj. Ecg paper records digitization through image processing techniques. *International Journal of Computer Applications*, 48(13):35–38, 2012.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, et al. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Brian Gow, Tom Pollard, Larry A Nathanson, et al. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *PhysioNet*, 2023. doi: 10.13026/4nqg-sb35.

E. William Hancock, Barbara J. Deal, David M. Mirvis, et al. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part V. *Circulation*, 119(10): e251–e261, 2009. doi: 10.1161/CIRCULATIONAHA.108.191097.

Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, et al. 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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the ACL: EMNLP*, pp. 4246–4253, 2020. doi: 10.18653/v1/2020.findings-emnlp.379.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, volume 30, 2017.

Sehun Kim. Learning general representation of 12-lead electrocardiogram with a joint-embedding predictive architecture, 2026. URL https://arxiv.org/abs/2410.08559.

Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012.

Felix Krones, Ben Walker, Terry Lyons, et al. Combining hough transform and deep learning approaches to reconstruct ECG signals from printouts. In *CinC*, volume 51, 2024.

Jun Li, Aaron D Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7):AIoa2401033, 2025.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 13094–13102, 2023.

Yao Li, Qixun Qu, Meng Wang, et al. Deep learning for digitizing highly noisy paper-based ecg records. *Computers in Biology and Medicine*, 127:104077, 2020. doi: 10.1016/j.compbiomed.2020.104077.

Yixiu Liang, Arunashis Sau, Boroumand Zeidaabadi, et al. Artificial intelligence-enhanced electrocardiography to predict regurgitant valvular heart diseases: an international study. *European Heart Journal*, 2025. doi: 10.1093/eurheartj/ehaf448.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, et al. A convnet for the 2020s. In *CVPR*, pp. 11966–11976, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, 2021. doi: 10.3758/s13428-020-01516-y. URL https://doi.org/10.3758%2Fs13428-020-01516-y.

Siddharth Mishra, Gaurav Khatwani, Rupali Patil, Darshan Sapariya, Vruddhi Shah, Darsh Parmar, Sharath Dinesh, Prathamesh Daphal, and Ninad Mehendale. Ecg paper record digitization and diagnosis using deep learning. *Journal of medical and biological engineering*, 41(4):422–432, 2021.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WcOohbsF4H.

Jiapu Pan and Willis J. Tompkins. A real-time qrs detection algorithm. *IEEE Transactions on Bio-Medical Engineering*, BME-32(3):230–236, 1985. doi: 10.1109/TBME.1985.325532.

Adam Paszke, Sam Gross, Francisco Massa, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, 2019.

Erick A. Perez Alday, Annie Gu, Amit J. Shah, et al. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement*, 41(12):124003, 2020. doi: 10.1088/1361-6579/abc960.

Manh Pham Hung, Aaqib Saeed, and Dong Ma. Boosting masked ECG-text auto-encoders as discriminative learners. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49277–49291. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/pham-hung25a.html.

Sushravya Raghunath et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature Medicine*, 26(6):886–891, 2020. doi: 10.1038/s41591-020-0870-z.

Pentti M. Rautaharju, Borys Surawicz, and Leonard S. Gettes. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part IV. *Circulation*, 119(10): e241–e250, 2009. doi: 10.1161/CIRCULATIONAHA.108.191096.

Lakshminarayan Ravichandran, Chris Harless, Amit J Shah, Carson A Wick, James H Mcclellan, and Srini Tridandapani. Novel tool for complete digitization of paper electrocardiography data. *IEEE journal of translational engineering in health and medicine*, 1:1800107–1800107, 2013.

Matthew A. Reyna, Deepanshi, James Weigle, et al. PhysioNet - Digitization of ECG Images. https://www.kaggle.com/competitions/physionet-ecg-image-digitization, 2025. Kaggle.

Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11): 2298–2304, 2017. doi: 10.1109/TPAMI.2016.2646371.

Kshama Kodthalu Shivashankara, Deepanshi, Afagh Mehri Shervedani, et al. ECG-Image-Kit: a synthetic image generation toolbox to facilitate deep learning-based electrocardiogram digitization. *Physiological Measurement*, 45(5):055019, 2024. doi: 10.1088/1361-6579/ad4954.

Anupreet Kaur Singh and Sridhar Krishnan. Ecg signal feature extraction trends in methods and applications. *BioMedical Engineering OnLine*, 22(1):22, 2023. doi: 10.1186/s12938-023-01075-1.

Sumeet S. Singh and Sergey Karayev. Full page handwriting recognition via image to sequence extraction. In *ICDAR*, pp. 55–69, 2021. doi: 10.1007/978-3-030-86334-0_4.

Borys Surawicz, Rory Childers, Barbara J. Deal, and Leonard S. Gettes. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part III. *Circulation*, 119(10):e235–e240, 2009. doi: 10.1161/CIRCULATIONAHA.108.191095.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *NeurIPS*, volume 30, 2017.

Galen S. Wagner, Peter Macfarlane, Hein Wellens, et al. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part VI. *Circulation*, 119(10): e262–e270, 2009. doi: 10.1161/CIRCULATIONAHA.108.191098.

Fuying Wang, Jiacheng Xu, and Lequan Yu. From token to rhythm: A multi-scale approach for ECG-language pretraining. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65059–65074. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wang25du.html.

Liang-Hung Wang, Chao-Xin Xie, Tao Yang, et al. Paper-recorded ecg digitization method with automatic reference voltage selection for telemonitoring and diagnosis. *Diagnostics*, 14(17):1910, 2024. doi: 10.3390/diagnostics14171910.

Huiyi Wu, Kiran Haresh Kumar Patel, Xinyang Li, et al. A fully-automated paper ecg digitisation algorithm using deep learning. *Scientific Reports*, 12(1):20963, 2022. doi: 10.1038/s41598-022-25284-1.

Xiankai Yu, Jian Wu, Jiahao Wang, et al. SLPM: a lightweight deep learning model for end-to-end paper ECG digitization. *Physiological Measurement*, 47(2):025005, 2026. doi: 10.1088/1361-6579/ae3fe5.

A SUPPLEMENTARY MATERIAL

This appendix provides architecture and data details (Sections A.1 and A.2), architectural ablations and attention maps (Sections A.3 and A.4), reconstruction protocols and fidelity analyses (Section A.5), downstream comparisons (Section A.6), uncertainty analyses (Section A.7), and real-image adaptation and failure cases (Section A.8).

A.1 ARCHITECTURE AND IMPLEMENTATION DETAILS

Table A1 summarises the configuration used throughout the experiments, including the image resolution, decoder dimensions, output topology, and optimisation settings.

Table A1: Architecture and training configuration of READ-ECG.

Component	**Value**
Input image	$3 \times 512 \times 2560$
Encoder	ConvNeXt-Tiny
Decoder	Topology-aware Transformer
Model dimension (d_{model})	256
Decoder layers	3
Attention heads	8
Feed-forward dimension	1024
QKNorm	Enabled
Prediction patch size	5
Uncertainty output	Laplace NLL
Output signal	12-lead (3×4 + rhythm), 250 Hz
Optimiser and learning rate	AdamW, 2×10^{-4}

A.1.1 SIGNAL SCALING

Fixed per-lead training-set statistics recover amplitudes as $\hat{x}_{\ell,t} = \sigma_\ell \hat{z}_{\ell,t} + m_\ell$ and uncertainty scales as $b^{\mathrm{mV}}_{\ell,t} = \sigma_\ell b_{\ell,t}$. Inference uses no recording-specific reference statistics or explicit printed-gain extraction.

A.2 Datasets and Synthetic Augmentation

Training pairs are produced from digital ECG records using ECG-Image-Kit and light photometric perturbations. Figure A1 illustrates the visual variation introduced during synthetic training, while Table A2 reports the record-level data partition.

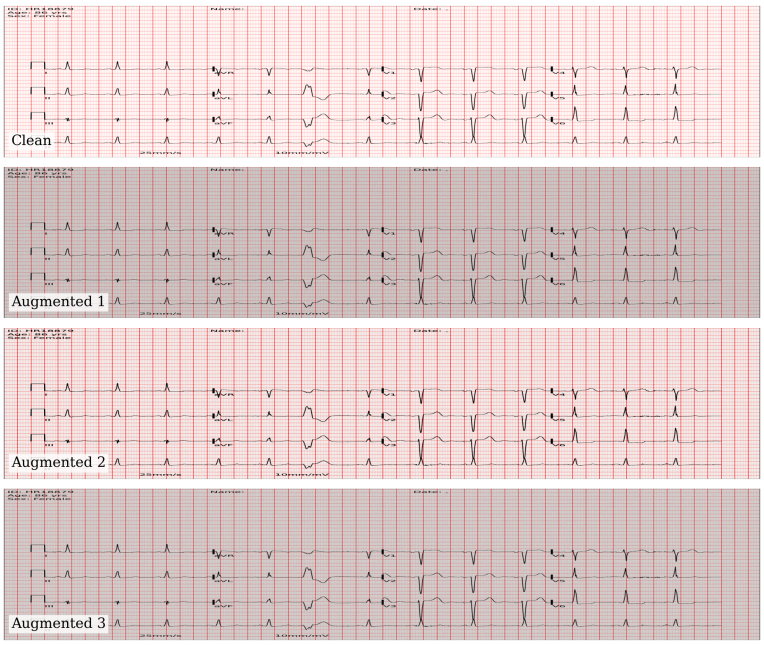


Figure A1: Examples of synthetic training images. The top row shows a clean ECG-Image-Kit rendering; subsequent rows show colour jitter and Gaussian-noise augmentations.

Table A2: Record distribution for the six PhysioNet subdatasets used in training and evaluation and the real-image Kaggle adaptation dataset. ECG-Image-Kit divides PhysioNet records longer than 10 seconds into 10-second segments and generates one image per segment. All images from the same record remain in the same training, validation, or test split.

Subdataset	Records	Images	Train	Val.	Test
PTB-XL	21,837	21,837	18,527	1,116	2,194
Georgia	10,292	10,292	8,771	504	1,017
CPSC-2018	6,867	9,364	5,858	325	684
CPSC-2018E	3,441	4,689	2,919	171	351
PTB	516	5,496	429	34	53
INCART	74	13,320	68	1	5
Kaggle	977	8,793	830	48	99

A.3 Architectural Ablations and Training Stability

Five output samples per query achieved the highest SNR among the tested patch sizes (Table A3). Single-sample prediction was less stable, while a patch size of 25 required the same 256-dimensional query to represent more output values.

QKNorm controls the scale of the attention logits during optimisation. Without it, the distribution remains initially bounded but grows sharply by epoch 30 (Table A4 and Fig. A2), accompanied by eventual loss divergence after approximately epoch 25 (Fig. A3).

Table A3: Patch-size ablation on the combined PhysioNet evaluation set. Higher SNR is better.

Patch size	SNR (dB)	Interpretation
1	24.1	High variance; difficult learning
5	**28.8**	Stable with preserved morphology
25	25.2	More output values per query

Table A4: Cross-attention logit statistics during training, with and without QKNorm. QKNorm keeps the logit distribution bounded through epoch 30.

Condition	Epoch	Mean	Std	Max
Without QKNorm	1	-0.09	1.58	14.3
Without QKNorm	24	-0.67	1.44	13.4
Without QKNorm	30	1,930.00	4,290.00	18,600.0
With QKNorm	1	-1.21	1.08	3.76
With QKNorm	24	1.14	1.16	6.28
With QKNorm	30	1.46	1.21	6.51

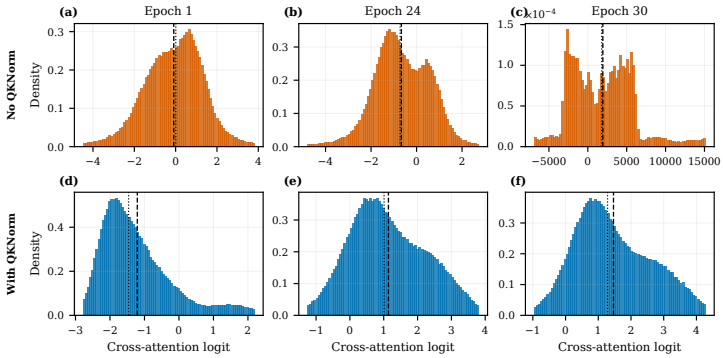


Figure A2: Cross-attention logit histograms across training. Without QKNorm (top), the distribution expands sharply by epoch 30; with QKNorm (bottom), it remains bounded.

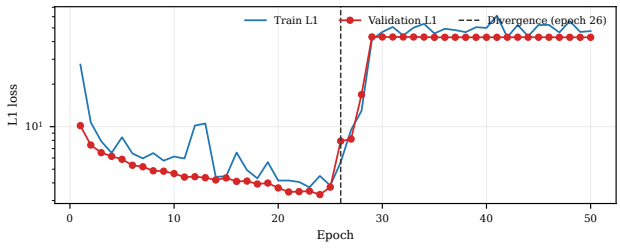


Figure A3: Training and validation L1 loss without QKNorm. Loss divergence emerges after approximately epoch 25, consistent with the concurrent growth of the cross-attention logits.

A.4 Learned Cross-Attention Localisation

The lead-specific maps in Fig. A4 indicate that the decoder learns spatial correspondence between each topology-aware query and its printed trace.

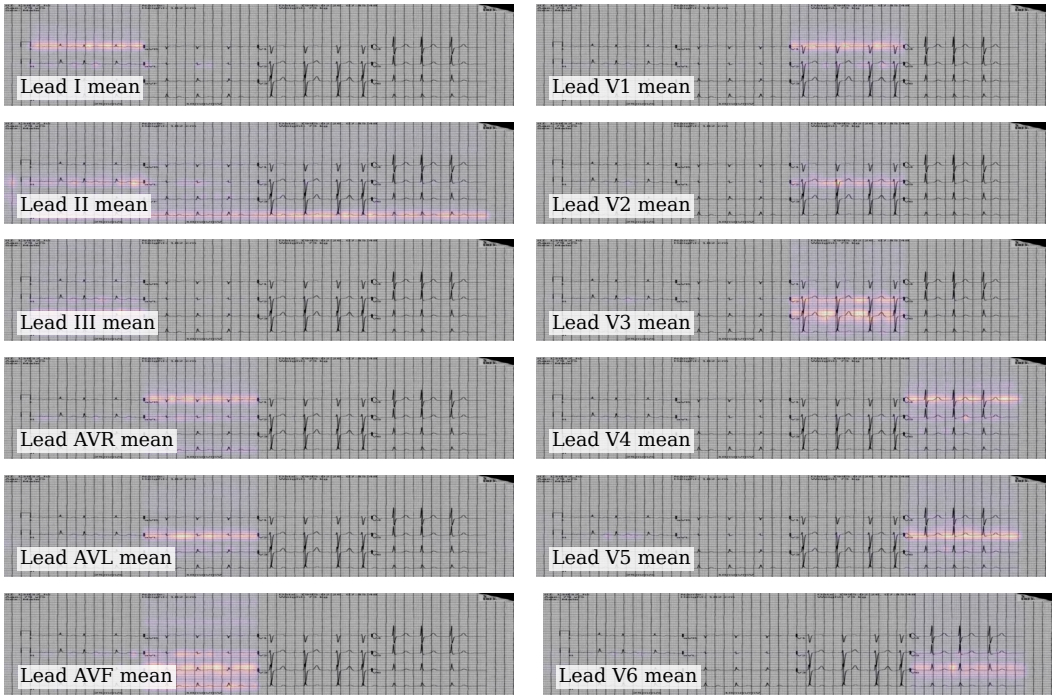


Figure A4: Learned cross-attention maps for all 12 leads from the Kaggle-tuned model evaluated on a Kaggle printout. Each map is concentrated around the image region containing the requested lead.

A.5 Reconstruction Protocols and Fidelity

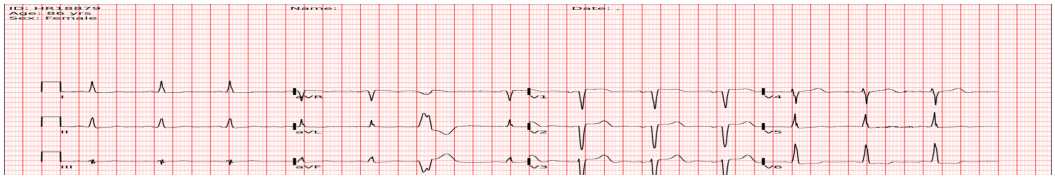
A.5.1 Benchmark Protocols

When a model is tested on an unfamiliar rendering style, a drop in performance may reflect the image shift rather than the digitiser itself. We therefore compare each method on clean images in its own rendering style. READ-ECG and our ECG Digitiser reproduction use our clean renderings; Demolder et al. and SLPM use their authors' reported clean-image results (Table A5 and Fig. A5).

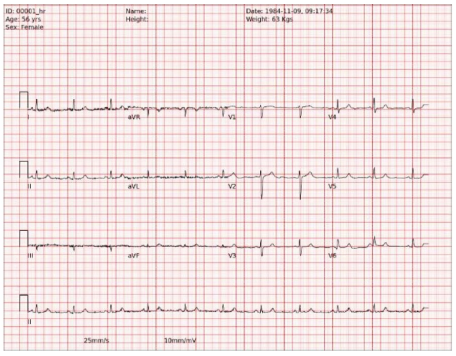
Table A5: Protocols behind the PTB-XL reconstruction results in Table 1.

Method	Training images	Selected test and result source
READ-ECG	ECG-Image-Kit renderings from six PhysioNet archives; clean base images with colour jitter and Gaussian noise.	Our clean PTB-XL test renderings (2,194 records); our evaluation.
ECG Digitiser (Krones et al., 2024)	PTB-XL ECG-Image-Kit images with visual augmentations.	Our clean PTB-XL test renderings; our reproduced result, not the original paper's cross-validation score.
Demolder et al. (2025)	Annotated real ECG images with augmented training patches.	Clean digital renderings in PM-ECG-ID (100 selected PTB-XL ECGs); published result.
SLPM (Yu et al., 2026)	Single-lead images from a modified ECG-Image-Kit mode.	PaperECG_12l: 1,000 PTB-XL 12-lead pages generated with ECG-Image-Kit, cropped by lead for inference; published result.

(a) READ-ECG and reproduced ECG Digitiser



(b) SLPM



(c) PM-ECG-ID, original digital

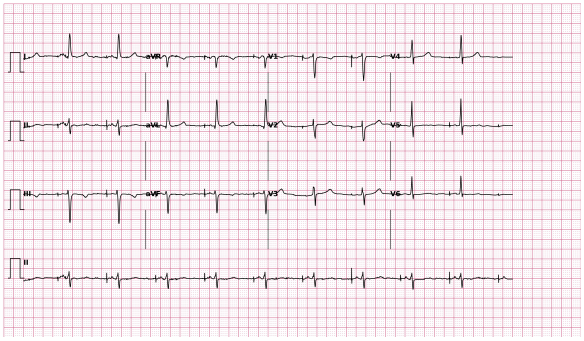


Figure A5: Examples of the clean image styles used in the PTB-XL reconstruction comparison. (a) Our ECG-Image-Kit rendering. (b) A PaperECG_12l image from Yu et al. (2026, Fig. 1(a)), also generated with ECG-Image-Kit. (c) A full-opacity original digital image from PM-ECG-ID, used by Demolder et al. (2025). The examples show different ECG records and are not matched renderings of one signal. Their visual differences do not establish that one style is easier or harder to digitise than another.

A.5.2 PER-LEAD RECONSTRUCTION FIDELITY

Per-lead results in Table A6 show consistently high correlation and SNR across limb and precordial leads on PTB-XL.

Table A6: Per-lead reconstruction metrics on the PTB-XL test set ($N = 2{,}194$ records). Metrics are computed separately for each record and lead and then averaged across records. MAE is measured in μV after inverse normalisation.

Metric	**I**	**II**	**III**	**aVR**	**aVL**	**aVF**
PCC ↑	0.9980	0.9986	0.9985	0.9978	0.9978	0.9986
MAE (μV) ↓	2.6	2.0	2.1	2.3	2.3	1.9
SNR (dB) ↑	29.39	31.65	28.76	28.92	27.01	29.44
Metric	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**
PCC ↑	0.9993	0.9992	0.9989	0.9991	0.9990	0.9991
MAE (μV) ↓	2.1	3.5	3.6	2.8	2.4	1.7
SNR (dB) ↑	31.21	31.51	31.33	31.52	32.06	33.18

A.5.3 INTERVAL EXTRACTION

Original and reconstructed 10-s lead-II signals at 250 Hz are processed independently with NeuroKit2 v0.2.12, using its default cleaning and R-peak detection followed by prominence-based delineation. QRS duration spans QRS onset to offset; QT spans QRS onset to T-wave offset. Eligible QRS and QT durations are 30–300 and 200–700 ms, respectively, with record-level intervals averaged over at least three valid beats. Missing boundaries are not replaced by peaks. Bazett correction uses each beat's preceding RR interval before averaging; eligible RR intervals lie between $60/220$ and 2 s. Heart rate is 60 divided by the mean eligible RR interval, requiring at least one interval. Limits of agreement are the mean difference ± 1.96 sample standard deviations. Single-measure absolute-agreement ICC uses the ICC(2,1) formula for the fixed original/reconstructed conditions.

Among 2,194 test records, 2,191 have paired heart-rate measurements, 2,191 have paired QRS durations, and 2,173 have paired QTc values. For three records, inputs don't yield an eligible heart-rate or QRS measurement. For QTc, two records have only the original measurement, two have only the READ-ECG measurement, and 17 have neither; unavailable values are excluded from the paired analysis.

A.6 Three-Task Downstream Comparison

A.6.1 Evaluation Protocol

This comparison uses 1,689 clean PTB-XL test ECGs from 1,665 patients absent from READ-ECG's training and validation sets, and separately predicts diagnosis (5 labels), rhythm (12 labels), and form (19 labels). Within each family, AUROC and average precision (AP) are macro-averaged across eligible labels; the reported three-task mean weights the families equally. Rows average three independently fitted classifier or probe seeds. The residual 1D CNN is trained on the original visible-waveform representation and evaluated unchanged on original and READ-ECG waveforms. The image classifiers input $512 \times 2{,}560$ renderings. Each original reference and READ-ECG reconstruction contains only the waveform visible on the printout, with unobserved intervals zero-filled. Foundation encoders are frozen; their MLP probes are fitted on original visible-signal features and evaluated unchanged on original and READ-ECG inputs.

The classifier split follows READ-ECG's PTB-XL source partitions after removing patients shared across partitions: 18,183 training, 1,116 validation, and 1,689 test ECGs. A record is eligible for a task family only if it has a recognised annotation in that family (1,671 diagnosis, 1,635 rhythm, and 685 form test records). Within a family, the classifiers minimise unweighted binary cross-entropy over eligible records; family losses are also weighted equally. The image models and residual CNN have shared encoders with separate two-layer task heads (256 hidden units, GELU, and dropout).

A.6.2 Models and Training

ResNet-18 and ConvNeXt-Tiny start from ImageNet-1K-pretrained encoders and are fitted end-to-end with newly initialised task heads; printouts are normalised with ImageNet channel statistics. The residual CNN starts from random weights and uses fixed per-lead training-set scaling. Its 12-channel, 250-Hz input retains the visible 2.5-second segments of the short leads in their printed positions, the complete 10-second lead-II strip, and zeros elsewhere. A 12-to-64-channel convolutional stem (kernel 15) feeds four temporal residual stages with widths 64, 128, 256, and 512 and block counts 2, 3, 4, and 3. The stages use kernel 7 and stride-2 downsampling at the start of each later stage. Global mean pooling gives a 512-dimensional feature vector for the three task heads. The CNN has 14.880 million fitted parameters. We chose this temporal residual design to classify waveforms using both local morphology and longer-range context; the architecture was not tailored separately to each task.

All three end-to-end classifiers use AdamW (weight decay 10^{-4}), batch size 16, BFloat16 autocast, gradient clipping at norm 1, and a 40-epoch cap with patience 10. Their learning rate warms up linearly for three epochs to 3×10^{-4}, then follows cosine decay to 10^{-3} of the base rate. The three fitting seeds are 20260918, 20260919, and 20260920.

ECG-JEPA uses the released multi-block epoch-100 checkpoint and ST-MEM the released ViT-B/75 checkpoint. Both encoders remain frozen and use the same per-family two-layer MLP probe (256 hidden units, GELU, and dropout). ECG-JEPA uses its native 250-Hz eight-lead input (I, II, V1–V6) in physical mV, with unprinted intervals zero-filled; token embeddings are mean-pooled. ST-MEM uses the 12-lead scan-visible input, three equally spaced 2,250-sample crops, a 0.67–40-Hz bandpass filter, and standardisation across leads and time for each crop. We average the three frozen crop embeddings before the MLP, rather than averaging classifier outputs as in the official end-to-end downstream evaluator. Each model's feature means and standard deviations are estimated on its training split and held fixed for validation and both test inputs.

The frozen ECG-JEPA and ST-MEM probes use the same fitting schedule as the image and CNN classifiers: AdamW (weight decay 10^{-4}), batch size 16, BFloat16 autocast, gradient clipping at norm 1, unweighted binary cross-entropy, three warm-up epochs to 3×10^{-4} followed by cosine decay to 10^{-3} of that rate over a 40-epoch horizon, and patience 10. The three task heads are fitted jointly and are then applied without retraining to original and READ-ECG test features. The foundation encoders remain frozen, whereas the image and CNN classifiers are fitted end-to-end.

A.6.3 CLASSIFICATION RESULTS

Table A7: Clean PTB-XL downstream classification. Each cell gives AUROC / AP; the final column is the equally weighted mean of the three task families. Original and READ-ECG rows for each signal model use identical fitted weights. Foundation models use MLP probes.

Model / input	Diagnosis	Rhythm	Form	Three-task mean
ResNet-18 / image	.9318 / .8272	.9469 / .5997	.8531 / .3345	.9106 / .5872
ConvNeXt-Tiny / image	.9288 / .8263	.9486 / .6596	.8612 / .3688	.9129 / .6182
Residual 1D CNN / original	.9408 / .8525	.9577 / .5781	.8821 / .3920	.9269 / .6075
Residual 1D CNN / READ-ECG	.9407 / .8525	.9578 / .5798	.8825 / .3922	.9270 / .6082
ECG-JEPA / original	.9227 / .8099	.9614 / .6864	.8265 / .3463	.9036 / .6142
ECG-JEPA / READ-ECG	.9227 / .8097	.9616 / .6813	.8267 / .3467	.9037 / .6126
ST-MEM / original	.9171 / .7942	.9730 / .6075	.8395 / .3512	.9099 / .5843
ST-MEM / READ-ECG	.9163 / .7926	.9715 / .6050	.8391 / .3505	.9090 / .5827

Table A8: Sample standard deviation across three independently fitted seeds for the equally weighted three-task mean in Table A7. Original and READ-ECG rows share each signal model's fitted seeds.

Model / input	AUROC SD	AP SD
ResNet-18 / image	.0036	.0053
ConvNeXt-Tiny / image	.0042	.0144
Residual 1D CNN / original	.0022	.0061
Residual 1D CNN / READ-ECG	.0021	.0054
ECG-JEPA / original	.0026	.0025
ECG-JEPA / READ-ECG	.0026	.0038
ST-MEM / original	.0042	.0039
ST-MEM / READ-ECG	.0043	.0033

A.6.4 PAIRED COMPARISONS

Table A9: READ-ECG + residual 1D CNN minus image ConvNeXt-Tiny on the same clean PTB-XL test records. Differences and pointwise 95% intervals use 2,000 paired patient-cluster weighted bootstrap replicates with fitted seeds resampled. Intervals crossing zero do not establish a difference.

Task	ΔAUROC [95% interval]	ΔAP [95% interval]
Diagnosis	$+.0119$ $[+.0059, +.0182]$	$+.0262$ $[+.0137, +.0387]$
Rhythm	$+.0092$ $[-.0027, +.0235]$	$-.0798$ $[-.1241, -.0224]$
Form	$+.0213$ $[-.0012, +.0461]$	$+.0234$ $[+.0007, +.0472]$
Three-task mean	$+.0141$ $[+.0064, +.0216]$	$-.0101$ $[-.0275, +.0115]$

Table A10: Additional paired differences with pointwise 95% patient-and-seed bootstrap intervals. The first three rows compare READ-ECG with original waveforms using the same signal model. The final row compares READ-ECG plus the residual CNN with image ResNet-18.

Comparison	ΔAUROC [95% interval]	ΔAP [95% interval]
CNN: READ-ECG − original	$+.00013$ $[-.00014, +.00044]$	$+.00064$ $[-.00056, +.00248]$
ECG-JEPA: READ-ECG − original	$+.00013$ $[-.00006, +.00033]$	$-.00163$ $[-.00649, +.00076]$
ST-MEM: READ-ECG − original	$-.00090$ $[-.00214, +.00027]$	$-.00160$ $[-.00436, +.00151]$
READ-ECG + CNN − ResNet-18	$+.0164$ $[+.0086, +.0241]$	$+.0210$ $[+.0030, +.0379]$

A.6.5 COMPUTATION

Table A11: Computation for three-task classification. Fitting cost is a rounded estimate for one classifier fit, including validation. Total cost adds one inference pass over the 1,689 test ECGs. The last row includes CNN fitting and both READ-ECG and CNN inference; READ-ECG adds 291.278 GFLOPs per scan. Image-model pretraining and READ-ECG training/adaptation are excluded.

Model	Parameters (M)	Fit (PFLOPs)	Inference (GFLOPs/ECG)	Total (PFLOPs)
ResNet-18 / image	11.580	93.025	94.750	93.185
ConvNeXt-Tiny / image	28.420	306.547	232.740	306.940
Residual 1D CNN / waveform	14.880	21.884	12.971	21.906
READ-ECG + residual CNN / scan	—	21.884	304.249	22.398

A.7 Uncertainty and Rejection

Reconstruction tables average SNR across examples within each lead, then equally across leads. Rejection curves instead use SNR from concatenated lead windows per example; combining powers before taking logarithms does not match averaging lead-wise SNRs.

Uncertainty-based retention supports both segment- and lead-level quality control. As the highest-uncertainty samples or leads are rejected, reconstruction metrics improve on the retained subset (Fig. A6).

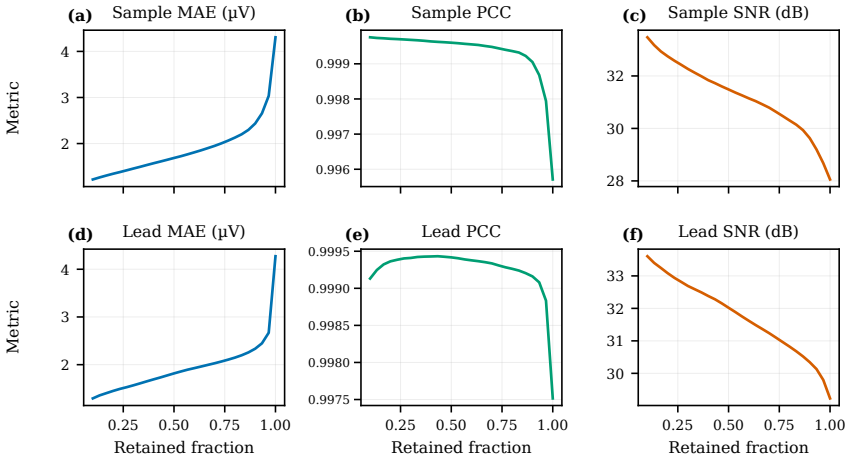


Figure A6: Uncertainty-based retention curves on the combined PhysioNet test set ($N = 6{,}100$ segments). The highest-uncertainty samples (top) or leads (bottom) are progressively rejected, and metrics are recomputed on the retained subset. Lower MAE and higher PCC and SNR are better.

The cross-domain distribution in Fig. A7 is shifted towards larger predicted scale values for Kaggle images, consistent with the greater visual degradation in real photographs than in clean PTB-XL renderings.

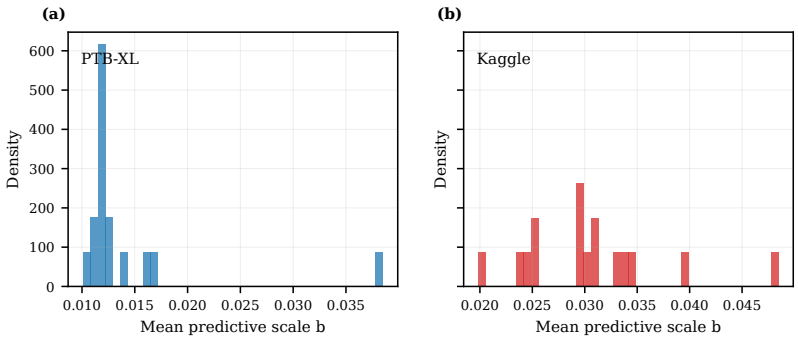


Figure A7: Predicted uncertainty-scale (b) distributions for clean PTB-XL test-set renderings (left) and held-out real images from the Kaggle test set after fine-tuning (right). The Kaggle distribution is shifted towards larger values, indicating lower confidence.

A.8 Real-Image Adaptation, Robustness, and Failure Cases

A.8.1 Few-Shot Adaptation

All Kaggle adaptation runs use the same separate validation split of 48 records (432 images) for checkpoint selection. We repeat each experiment with three random seeds, resampling the training records for each few-shot run while keeping the validation and test sets unchanged. For all Kaggle adaptation runs, we fine-tune all READ-ECG parameters with AdamW (initial learning rate 2×10^{-4}; batch size four per GPU) and cosine decay over 200 epochs, although every run reached its lowest validation MAE by epoch 66. Each epoch mixes equal numbers of real and synthetic training images, shuffled into minibatches.

Table A12: Kaggle few-shot adaptation across three runs. Values at 1–100% are mean ± sample SD across three fits on the fixed 891-image test set; 0% reports the shared checkpoint with 95% record-bootstrap intervals in brackets. Figure 6 uses the same uncertainty measures, with diamonds marking 0%. SNR is averaged over leads. MAE is computed in μV for each image and averaged over the test set.

Training data	SNR (dB) ↑	MAE (μV) ↓
0%	$-3.83\ [-4.14, -3.52]$	$207.88\ [201.93, 214.64]$
1%	12.27 ± 0.22	22.59 ± 1.64
5%	14.42 ± 0.06	15.91 ± 0.15
10%	15.58 ± 0.13	13.02 ± 0.21
25%	17.39 ± 0.11	10.76 ± 0.14
100%	19.48 ± 0.08	8.43 ± 0.06

A.8.2 REPEATED CAPTURES

Repeated captures of the same ECG record provide a controlled view of sensitivity to photographic conditions. After Kaggle fine-tuning, all eight examples in Fig. A8 retain an SNR above 16.7 dB despite changes in viewpoint and degradation.



Figure A8: Reconstructions for eight photographs of one ECG record from the held-out Kaggle test set under varied capture conditions. After Kaggle fine-tuning, READ-ECG achieves SNR above 16.7 dB for each image.

A.8.3 FAILURE CASES

The examples in Fig. A9 show the remaining limitation: when large signal regions are degraded or obscured, reconstruction fidelity falls and predicted uncertainty rises around several affected segments.

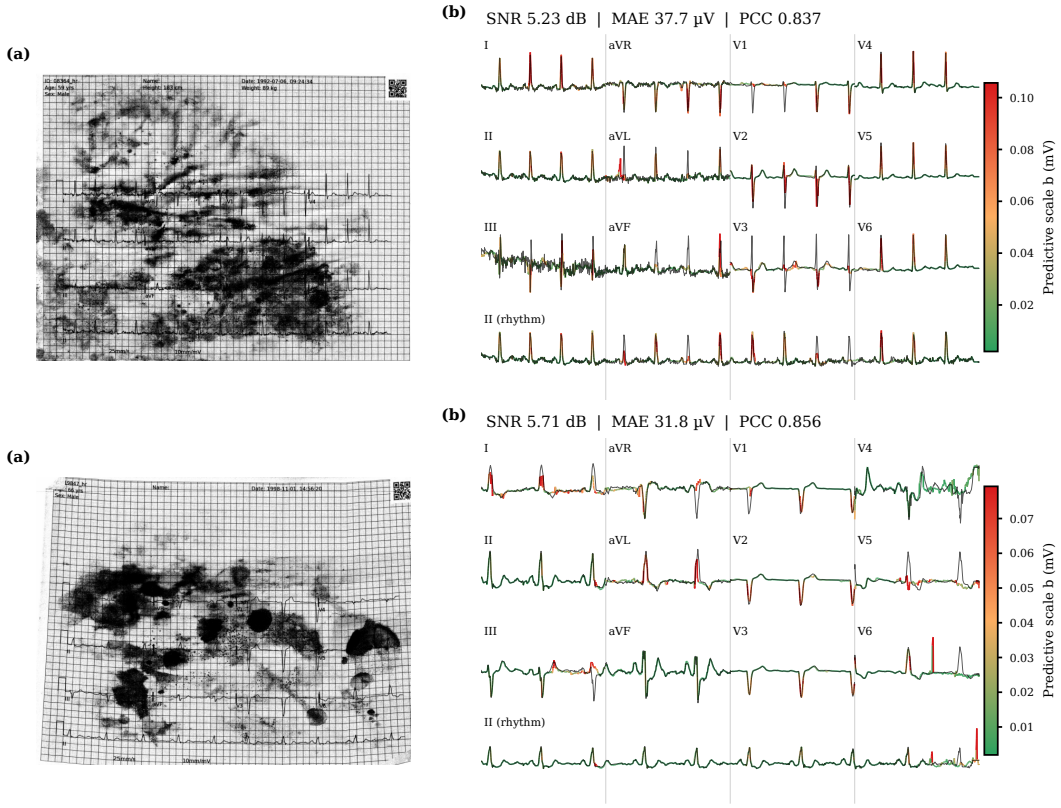


Figure A9: Two representative failure cases from the held-out Kaggle test set after fine-tuning. Large degraded or obscured regions reduce reconstruction fidelity, while predicted uncertainty increases around several affected segments.