arXiv:2508.02720v2 [cs.LG] 2 Feb 2026

ECGTwin: Personalized ECG Generation Using Controllable Diffusion Model

Yongfan Lai [1 2 3] **Bo, Liu** [1 2] **Xinyan Guan** [3] **Qinghao Zhao** [4] **Hongyan Li** [1 2] **Shenda Hong** [3]

Abstract

Personalized electrocardiogram (ECG) generation is to simulate a patient's ECG digital twins tailored to specific conditions. It has the potential to transform traditional healthcare into a more accurate individualized paradigm, while preserving the key benefits of conventional population-level ECG synthesis. However, this promising task presents two fundamental challenges: extracting individual features without ground truth and injecting various types of conditions without confusing generative model. In this paper, we present **ECGTwin**, a two-stage framework designed to address these challenges. In the first stage, an *Individual Base Extractor* trained via contrastive learning robustly captures personal features from a reference ECG. In the second stage, the extracted individual features, along with a target cardiac condition, are integrated into the diffusion-based generation process through our novel *AdaX Condition Injector*, which injects these signals via two dedicated and specialized pathways. Both qualitative and quantitative experiments have demonstrated that our model can not only generate ECG signals of high fidelity and diversity by offering a fine-grained generation controllability, but also preserving individual-specific features. Furthermore, ECGTwin shows the potential to enhance ECG auto-diagnosis in downstream application, confirming the possibility of precise personalized healthcare solutions.

1. Introduction

Personalized ECG generation enables the creation of a patient's ECG digital twins across a range of cardiac conditions. It not only retains the key benefits of traditional population-level ECG generation, such as providing data for rare diseases and supporting cardiology education (Lai et al., 2025), but also introduces a transformative opportunity for ECG auto-diagnosis by enabling personalized models that are finetuned on a patient's own ECG data, allowing the model to focus on individual-specific features (Luo et al., 2017; Ding et al., 2025). Since ECG signals can vary significantly among patients—even under the same cardiac condition—this personalized approach offers improved diagnostic accuracy (Shusterman & London, 2024) compared to conventional models trained on population-level datasets. Together, these applications highlight personalized ECG generation as a promising direction for future research.

ECG generative model have been discussed by many work (Chen et al., 2022; Chung et al., 2023; Lai et al., 2025), but generating personalized ECG signals presents two new challenges: **(1) Extract individual features from a reference ECG without ground truth.** While deep learning models have demonstrated the ability to capture individual patient characteristics from medical data (Yang et al., 2024), it remains unclear how to effectively leverage these methods for extracting personal features specifically from reference ECG signals. Unlike demographic statistics, individual features extracted from ECG signals lack ground truth for supervision, making their extraction and verification more challenging. Prior work on personalized ECG generation (Hu et al., 2024) has adopted Vector Quantization (van den Oord et al., 2017) based approach to encode individual features. However, this approach complicates training, as the feature extractor and generative model are tightly coupled and must be optimized jointly.

(2) Inject different types of conditions into generative process. The generation of personalized ECGs requires incorporating various types of conditional information, such as individual patient features, target cardiac diagnosis, and demographic attributes. While the abundance of conditions offers the potential for fine-grained control, it also poses a significant challenge: if not properly integrated, these conditions may confuse the generative model, leading to degraded performance or even model collapse (Fetaya et al., 2020; Thanh-Tung & Tran, 2020). Previous work on ECG generation has been largely unconditioned or conditioned only on high-level labels, which is simple but limits both the diversity and controllability of the generated results.

[1]State Key Laboratory of General Artificial Intelligence, Beijing, China [2]School of Intelligence Science and Technology, Peking University, Beijing, China [3]National Institute of Health Data Science, Peking University, Beijing, China [4]Department of Cardiology, Peking University People's Hospital, Beijing, China. Correspondence to: Shenda Hong <hongshenda@pku.edu.cn>.

Preprint. February 3, 2026.

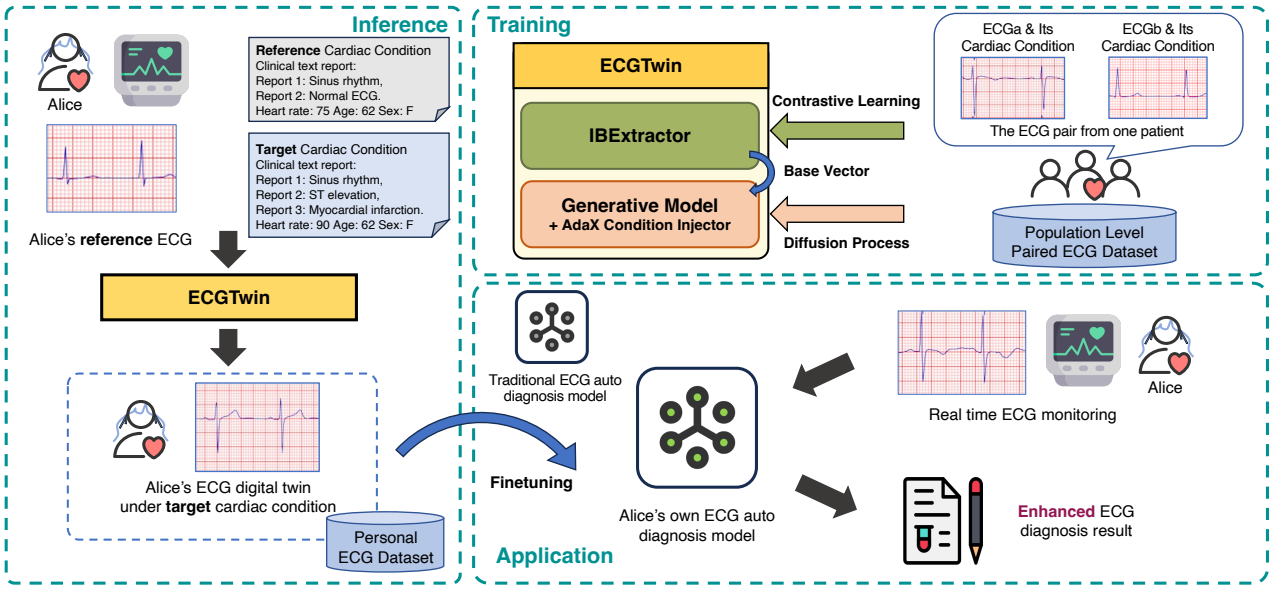


Figure 1. **Training, inference and application of ECGTwin.** Trained via contrastive learning and diffusion objectives respectively, and acting in a two-stage manner, our model is capable of simulating the plausible ECG signal (i.e. ECG digital twins) under a specified cardiac condition with input of a reference ECG and associated cardiac condition. The generated ECG digital twins carries more individual-relevant patterns and can be used to enhance the ECG diagnosis driven by deep learning methods.

To address aforementioned challenges, in this paper, we introduce **ECGTwin**, a diffusion-based model designed for personalized ECG generation. Given a reference ECG signal along with the reference and target cardiac condition, our model synthesizes high-quality ECG digital twins tailored to the target conditions (Fig. 1). To effectively extract the base vector shared across a patient's ECGs without ground truth, we build an *Individual Base Extractor* and train it separately using contrastive learning in a self-supervised manner. For the generation process, we adopt the Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020). To better incorporate various types of conditional information into the diffusion process, we equip the noise prediction model with our proposed *AdaX Condition Injector*, which enables controllable and effective conditioning through two dedicated pathways. We build ECGTwin on a curated ECG pair dataset derived from the MIMIC-IV-ECG dataset (Gow et al., 2023), which is approximately 20 times larger than the dataset used in prior personalized ECG generation study.

In summary, our contributions are:

(1) We formally define the task of personalized ECG generation and propose **ECGTwin**, a two-stage diffusion-based model capable of generating high-fidelity ECG digital twins.

(2) We demonstrate that models trained via contrastive learning can effectively extract personalized features from raw ECG signals. Building on this insight, we develop the **Individual Base Extractor** as the first stage of ECGTwin.

(3) we introduce the **AdaX Condition Injector**, which incorporates conditions through two dedicated pathways, enabling precise and effective modulation of the ECG generation process.

(4) We conduct extensive experiments to evaluate the fidelity, personal consistency and explainability of the generated ECG digital twins. We also simulate a downstream ECG auto diagnosis application and demonstrate that ECGTwin significantly improves the performance of ECG auto diagnosis, highlighting its potential in personalized healthcare.

2. Problem Definition

Definition 2.1. Cardiac Condition
We use the term cardiac condition $\mathbf{c}$ to represent the state of the heart. It encompasses variables from multiple dimensions to provide a comprehensive characterization of cardiological and physiological situation. In our work, cardiac condition $\mathbf{c}$ consists of a set of textual clinical reports describing cardiological findings, heart rate reflecting cardiac features, age as an indicator of physiological status, and sex for hormonal factors. Notably, when data availability allows, the cardiac condition $\mathbf{c}$ can be flexibly extended to include additional variables, enabling a more complete physio-cardiac descriptor.

Definition 2.2. Personalized ECG Generation
Given a reference ECG $\mathbf{x}_{\text{ref}}$ of a patient and its associated

cardiac condition $\mathbf{c}_{\text{ref}}$, the goal of personalized ECG generation is to simulate the plausible ECG signal $\hat{\mathbf{x}}$ that can reflect the patient's cardiac state under the target cardiac condition $\mathbf{c}_{\text{tar}}$. This is achieved by learning and sampling from the conditional distribution of $P(\hat{\mathbf{x}}|\mathbf{x}_{\text{ref}}, \mathbf{c}_{\text{ref}}, \mathbf{c}_{\text{tar}})$.

3. Method

ECGTwin performs personalized ECG generation in a two-stage manner. In first stage, the Individual Base Extractor extracts the base vector of the patient from reference ECG and its associated cardiac condition. Then, in second stage, the latent diffusion model with dedicatedly proposed AdaX Condition Injector integrates the base vector and the input target cardiac condition to generate ECG digital twins.

3.1. Individual Base Extractor

Given a reference ECG $\mathbf{x}_{\text{ref}}$ and its associated cardiac condition $\mathbf{c}_{\text{ref}}$, we introduce an *Individual Base Extractor* to extract the base vector $\mathbf{b}$ of the patient. The base vector $\mathbf{b}$ serves as a compact representation of individual-specific traits and is designed to remain invariant across ECGs recorded under different cardiac conditions for the same patient. As the substitution of reference ECG $\mathbf{x}_{\text{ref}}$ and cardiac condition $\mathbf{c}_{\text{ref}}$, we only forward extracted base vector $\mathbf{b}$ to the diffusion process for personalized ECG generation. This can be interpreted as a decomposition of the target conditional distribution: $P(\hat{\mathbf{x}}|\mathbf{x}_{\text{ref}}, \mathbf{c}_{\text{ref}}, \mathbf{c}_{\text{tar}}) \rightarrow P(\hat{\mathbf{x}}|\mathbf{b}, \mathbf{c}_{\text{tar}})$, where $\mathbf{b} = f_{\text{IBE},\theta}(\mathbf{x}_{\text{ref}}, \mathbf{c}_{\text{ref}})$ and θ stands for learnable parameters. This two-stage computation framework offers several advantages. First, it explicitly designates personalized features in the generation process, enabling the diffusion model to more effectively capture individual-specific information by reducing redundant conditioning signals. Second, it decouples the extraction of personalized information from the generative model, resulting in more stable training and simplified optimization. Additionally, we choose to extract base vector from the VAE encoded ECG latent $\mathbf{z}_{\text{ref}}$ rather than directly from raw ECG signal, i.e. $\mathbf{b} = f_{\text{IBE},\theta}(\mathbf{z}_{\text{ref}}, \mathbf{c}_{\text{ref}})$, so as to better align the base vector with the latent space used by the generation model.

The remaining challenge is how to train the Individual Base Extractor to accurately learn the base vector, especially in the absence of ground-truth labels for supervision. To address this, we adopt a self-supervised learning approach. The pretext training task is to maximize the similarity of base vectors extracted from ECGs of the same patient, while minimizing the similarity of base vectors from different patients. Inspired by the CLIP Loss (Radford et al., 2021), we construct a dataset of ECG pairs, where each pair consists of two ECGs recorded from the same patient. The loss function for the Individual Base Extractor is defined as:

$$\mathcal{L}_{IBE} = -\frac{1}{NB}\sum_{k=1}^{N}\sum_{i=1}^{B}\frac{1}{2}\left(\log\frac{\exp\left(\text{Sim}(\mathbf{b}_i, \mathbf{b'}_i)/\tau\right)}{\sum_{j\neq i}\exp\left(\text{Sim}(\mathbf{b}_i, \mathbf{b'}_j)/\tau\right)} + \log\frac{\exp\left(\text{Sim}(\mathbf{b'}_i, \mathbf{b}_i)/\tau\right)}{\sum_{j\neq i}\exp\left(\text{Sim}(\mathbf{b'}_i, \mathbf{b}_j)/\tau\right)}\right) \tag{1}$$

where the $\mathbf{b}_i$ and $\mathbf{b'}_i$ are base vectors extracted from the i-th ECG pair. $\text{Sim}(\cdot,\cdot)$ is the similarity function and we use the cosine similarity. B is the batch size, N is number of batches, and τ is the learnable temperature parameter. By explicitly pairing two ECGs from the same patient and viewing as positive samples, we can effectively leverage the hidden personalized features from a population level dataset as supervising signals. It is worth noting that cases where multiple pairs, e.g. $(\mathbf{b'}_i, \mathbf{b}_i)$ and $(\mathbf{b'}_j, \mathbf{b}_j)$, originate from the same patient could potentially introduce ambiguity into the model. However, since the number of ECG pairs from any patient is far less than the total number of ECG pairs, we can consider the positive matches outside designated pairings are extremely sparse, even within a large batch. As a result, the probability of such collisions is close to zero, and their impact on training can be safely ignored. An additional benefit of ECG pairing is that the same dataset can be directly used for training ECG generation model in the next stage, ensuring intrinsic consistency even when the components are optimized separately.

We implement the Individual Base Extractor function $f_{\text{IBE},\theta}$ based on Transformer encoder(Vaswani et al., 2017). As shown in Fig. 2, an cross-attention layer is added to receive the reference cardiac condition $\mathbf{c}_{\mathbf{ref}}$. The conditional mechanism of $\mathbf{c}_{\mathbf{ref}}$ aligns with the AdaX Cardiac Condition Pathway, which will be detailed in section 3.2. Nevertheless, to enhance robustness, we randomly mask $\mathbf{c}_{\text{ref}}$ at a fixed ratio during training and replace it with a special learnable embedding. This prepares the model to handle scenarios where $\mathbf{c}_{\text{ref}}$ is unavailable at inference time.

3.2. AdaX Condition Injector

Given a base vector $\mathbf{b}$ encapsulating personalized features and a target cardiac condition $\mathbf{c}_{\text{tar}}$ specifying desired morphological features of results, the ECG generation process should handle different kinds of conditional information meticulously so as to synthesize the patient's ECG digital twins of high quality and fidelity. In ECGTwin, we adopt latent diffusion process to model the target conditional distribution $P(\hat{\mathbf{z}}_0|\mathbf{b}, \mathbf{c}_{\text{tar}})$ by iteratively denoising the ECG latent $\hat{\mathbf{z}}_t$. During the process, a noise predictor must take the timestep t, the base vector $\mathbf{b}$, and the target cardiac condition $\mathbf{c}_{\text{tar}}$ as conditional inputs to compute the current noise $\epsilon_\theta(\hat{\mathbf{z}}_t, t, \mathbf{b}, \mathbf{c}_{\text{tar}})$ required by the DDPM reverse process.

These conditional inputs are heterogeneous in both semantic

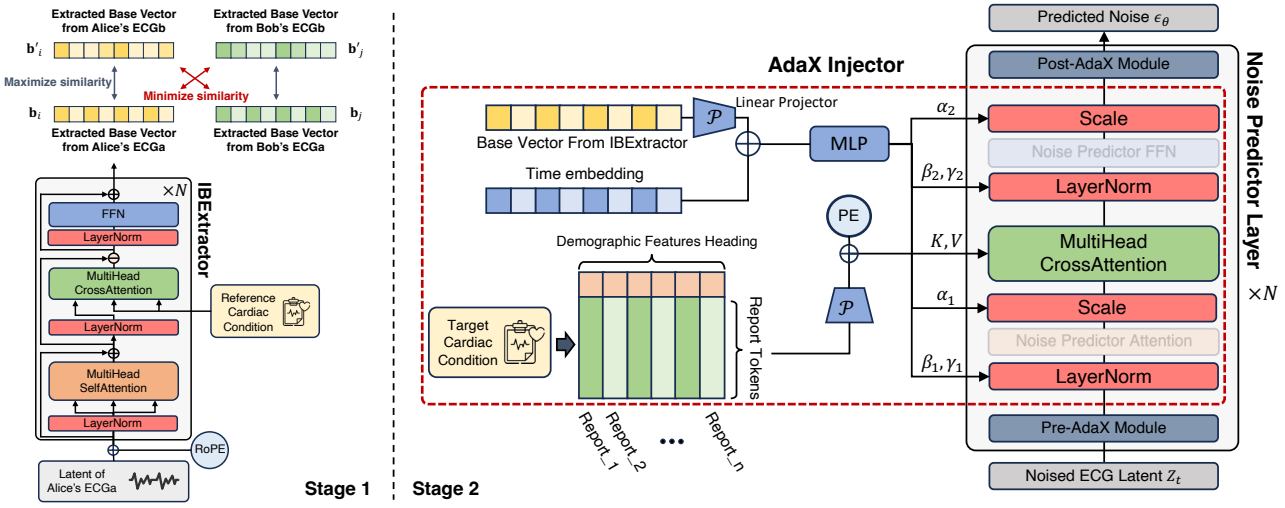


Figure 2. **Architecture of modules in ECGTwin's two stages.** The complete flow chart can be find in App. B.

meaning and numerical dimension, and thus should be handled through separate, specialized ways. Notably, the target cardiac condition $\mathbf{c}_{\text{tar}}$ carries the most detailed information and directly governs the waveform of the generated ECG signal. Therefore, an effective and expressive mechanism for encoding cardiac condition is critical for successful controllable ECG generation. To achieve this, we design the *AdaX Condition Injector* module as the conditioning interface of the noise predictor, enabling effectively integration all relevant conditions. As illustrated in Fig. 2, the module processes conditions through two distinct pathways:

3.2.1. Cardiac Condition Pathway

We utilize each component of the cardiac condition — the clinical reports, sex, age, and heart rate — to construct a cardiac condition sequence that serves as the key and value in a cross-attention mechanism. Note that an ECG signal is often associated with multiple clinical reports, each describing different aspects such as rhythm, morphology, or diagnosis. We treat each report as a token and employ the nomic-embed-text-v1.5 model (Nussbaum et al., 2024) to obtain embedding $\mathbf{e}(\in \mathbb{R}^{768})$ for each report token. Compared with traditional methods like byte-pair encoding (BPE) (Sennrich et al., 2015) or concatenating all reports into a single prompt (Lai et al., 2025), our report-level tokenization allow the model to selectively attend to different reports and adaptively emphasize the most relevant information.

For remaining features, we compile sex, age, and heart rate into a vector $\mathbf{p}(\in \mathbb{R}^3)$, where age and heart rate are zero-score normalized, and sex is binary encoded as 1 (male) and 0 (female). Then, the vector is duplicated and appended to each report embedding as headings, ensuring that these features are always accessible regardless of which report is being attended to. Next, the resulting augmented embeddings are stacked together to form the cardiac condition sequence. This sequence is then linearly projected into the embedding space of the noise predictor to ensure dimensional compatibility and representation consistency. Finally, positional encoding (PE) is applied to preserve the hierarchical information among the report tokens:

$$K, V = \text{Stack}(\{\text{Concat}(\mathbf{e}_i, \mathbf{p})\}_{i=1}^{n}) \cdot W + \text{PE} \quad (2)$$

3.2.2. Base Vector and Time Pathway

We design the conditioning pathway for the base vector and time embedding using adaptive normalization, as both represent global information and should be injected into the model holistically. Specifically, we obtain the time embedding $\mathbf{t}$ following the sinusoidal encoding method in DiT original implementation(Peebles & Xie, 2023). Meanwhile, the base vector $\mathbf{b}$, extracted from the Individual Base Extractor, is linearly projected to produce a dimension-aligned embedding $\bar{\mathbf{b}}$. The two embeddings are added element-wise and passed through a Multi-Layer Perceptron (MLP) to predict the scaling factor α, the LayerNorm shift and scale parameters β and γ, respectively:

$$\alpha, \beta, \gamma = \text{MLP}(\mathbf{t} + \bar{\mathbf{b}}) \quad (3)$$

These parameters are then used to adaptively normalize and scale the latent feature $\mathbf{z}$ as follows:

$$\mathbf{z}_{\text{layernorm}} = \frac{\mathbf{z} - \text{E}[\mathbf{z}]}{\sqrt{\text{Var}[\mathbf{z}] + \epsilon}} \cdot \gamma + \beta, \ \mathbf{z}_{\text{scale}} = \alpha \cdot \mathbf{z} \quad (4)$$

where ϵ is a small constant added for numerical stability.

Another advantage of AdaX Condition Injector is its compatibility with *Prompt-to-Prompt* editing(Hertz et al., 2022),

which enables a further fine-grained and flexible control over the results, crediting to the cross-attention mechanism integrated. For more details of this post-generation ECG editing technique, please refer to App. F.

3.3. Personalized ECG Generation

We unify aforementioned Individual Base Extractor and AdaX Injector within the framework of a latent diffusion process (Rombach et al., 2022) for personalized ECG generation. In the diffusion forward process, the clean ECG latent is progressively corrupted by Gaussian noise according to:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t \,, \epsilon_t \sim \mathcal{N}(0, \mathbf{I}) \tag{5}$$

where $\bar{\alpha}_t$ are noise scheduling hyperparameters with formulations: $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. In the diffusion reverse process, the model iteratively denoises $\mathbf{z}_t$ by sampling $\mathbf{z}_{t-1}$ from $\mathcal{N}(\mu_t, \sigma_t^2\mathbf{I})$, where:

$$\mu_t := [\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\mathbf{z}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\mathbf{z}}_0]/(1-\bar{\alpha}_t),$$

$$\hat{\mathbf{z}}_0 := [\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{b}, \mathbf{c}_{\text{tar}})]/\sqrt{\bar{\alpha}_t}; \tag{6}$$

$$\sigma_t^2 := (1-\alpha_t)(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t) \tag{7}$$

Here, the noise predictor $\hat{\epsilon}_\theta$ takes the noisy latent $\mathbf{z}_t$, diffusion timestep t, the base vector $\mathbf{b}$, and the target cardiac condition $\mathbf{c}_{\text{tar}}$ as inputs. It is trained using the simplified denoising score matching loss as in DDPM:

$$\mathcal{L}_{ECGTwin} = \|\epsilon_t - \hat{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{b}, \mathbf{c}_{\text{tar}})\|^2 \tag{8}$$

Finally, a pre-trained VAE Decoder (Kingma & Welling, 2014) is employed to reconstruct the denoised latent $\mathbf{z}_0$ back to signal space, yielding personalized ECG digital twin $\hat{\mathbf{x}}$. Notably, the training objective can be readily extended to more recent formulations such as flow matching losses (Lipman et al., 2023), and we leave it for future exploration.

4. Experiment

4.1. Experiment Settings

4.1.1. Datasets

We use the publicly available **MIMIC-IV-ECG** (Gow et al., 2023) dataset for the training and evaluation of our method, and link it to MIMIC-IV-Clinical dataset (Johnson et al., 2023) via "subject_id" to retrieve necessary data items. To construct paired samples for the sake of effective training, we group the ECG records by "subject_id", and exclude patients with only a single record. For each remaining patient, we generate all possible pairs of ECG records, which results in $\binom{n_i}{2}$ samples for a patient with n_i records. Following this procedure, we obtain a training set of 6,408,782 ECG record pairs and a testing set of 399,499 ECG pairs. More details related to MIMIC-IV-ECG dataset and the introduction of external validation dataset PTB-XL (Wagner et al., 2020) can be found in App. C.

4.1.2. Implementations

All of our trainings and tests are based on PyTorch 2.1.1, on GeForce RTX 3090, DiT based model iterates approximately 140 time steps per second when generating batch is set to 10 while Unet based model iterates approximately 60 time steps per second on the same environment. Our code is available at `https://github.com/Raiiyf/ECGTwin`. For details of model implementation, please refer to App. C.

4.1.3. Baselines

We compare our method with LAVQ (Hu et al., 2024), which to the best of our knowledge is the only existing approach designed for personalized ECG generation. Although the task setting in LAVQ differs from ours and requires more information for generation, we adapt it for evaluation on the MIMIC-IV-ECG dataset to the extent possible. Additionally, we construct a personalized version of the population-level state-of-the-art method DiffuSETS (Lai et al., 2025), denoted as DiffuSETS-p, by simply concatenating the reference ECG latent with the diffusion noise. This adaptation serves as a baseline to evaluate the effectiveness of our proposed condition injection strategy.

4.2. Generation Quality Evaluation

In this section, we demonstrate that ECGTwin is capable of generating high-quality ECG digital twins. We use the three-level evaluation protocol (Lai et al., 2025) to comprehensively assess the generation results from the scope of data distribution (using FID, Precision, Recall and F1) to semantic alignment (using HR-MAE and CLIP Score). We build and test ECGTwin on two types of noise predictor backbone architectures: DiT and Unet. To highlight the effectiveness and adaptivity of our proposed AdaX Condition Injector, we perform ablation studies on both backbones by limiting the conditioning mechanism to a single pathway. For clarity of the core mechanism, models incorporating the Cardiac Condition Path are suffixed by CA (cross-attention) and models with the Base Vector and Time Path are suffixed by adaLN (adaptive LayerNorm). Detailed descriptions of the evaluation metrics and the ablated model variants are provided in App. D. The test result are summarized in Tab. 1.

Compared with other ECG digital twin generation methods, the two variants of ECGTwin present the most outstanding overall performance. This superiority can be attributed to our two-stage framework, which effectively reduces confounding information and simplifies the training process. Furthermore, both ECGTwin models incorporating AdaX Condition Injector outperform their respective ablated counterparts in each branches, highlighting the its effectiveness in integrating conditions through two dedicated paths. In particular, models equipped with the Cardiac Condition

Table 1. **Three-level evaluation result.** Best values among each architectures are marked in blue, and global-best value is bolded in **red**.

Branch	Model Name	Signal Level FID (↓)	Signal Level Precision (↑)	Signal Level Recall (↑)	Signal Level F1 (↑)	Feature Level HR-MAE (↓)	Diagnostic Level Clip (↑)
Baseline	DiffuSETS-p	245±28.0	0.848±0.003	0.849±0.004	0.849±0.002	10.76±0.58	0.709±0.002
	LAVQ	234±0.3	0.829±0.004	0.519±0.008	0.639±0.006	38.96±0.04	0.674±0.001
UNet	UNet-*CA*	175±8.7	**0.925**±0.003	0.755±0.008	0.831±0.004	**4.63**±0.30	0.780±0.001
	UNet-*adaLN*	34±1.1	0.884±0.004	0.847±0.011	0.865±0.007	8.22±0.66	0.763±0.001
	ECGTwin (*Ours*)	**18**±3.1	0.883±0.003	0.860±0.004	0.871±0.002	7.31±0.56	0.784±0.001
DiT	DiT-*CA*	99±4.9	0.855±0.003	**0.903**±0.006	0.879±0.002	9.04±0.26	0.783±0.001
	DiT-*adaLN*	51±8.1	0.873±0.003	0.893±0.004	0.883±0.003	7.73±0.37	0.752±0.001
	ECGTwin (*Ours*)	26±2.8	0.872±0.002	0.896±0.002	**0.884**±0.002	7.03±0.50	**0.789**±0.000

Path consistently achieve higher scores in clinical text report alignment (measured by CLIP Score). This observation directly reflects the synergy between our report-level tokenization strategy and the cross-attention mechanism — an appropriate match for the structured nature of ECG clinical text reports. In summary, this three-level evaluation confirms that ECGTwin can generate ECG of high fidelity and diversity, thus is also compatible for applications of conventional (i.e. population-level) ECG synthesis models. Regarding the possible distribution bias brought by single training set, we also perform an **external validation on PTB-XL dataset for the generalization ability evaluation**, please refer to App. D for details.

4.3. Personal Consistency Assessment

In this section, we first test the efficacy of our Individual Base Extractor (IBExtractor) trained in a self-supervised manner. Then we show that the ECG digital twins generated by ECGTwin encapsulate personalized features of the reference ECG within the scope defined by the base vector.

4.3.1. Efficacy of Individual Base Extractor

We choose ten patients from the test set, each with more than 85 ECG records. From these records, we obtain a set of base vectors B, where each element is derived using the Individual Base Extractor. To show the original data distribution as a reference, we also test the identity mapping, i.e. directly flattening the ECG latent as output. Fig. 3 presents the t-SNE visualization of extracted base vectors. To numerically assess the intra- and inter-individual similarity, we propose a **similarity score** s:

$$s = \frac{1}{|B|}\sum_{\mathrm{b}_i \in B}\left(\frac{1}{|B_I|}\sum_{\mathrm{b}_j \in B_I}\frac{\mathrm{b}_i \cdot \mathrm{b}_j}{\|\mathrm{b}_i\|\|\mathrm{b}_j\|} - \frac{1}{|B_I^c|}\sum_{\mathrm{b}_j \notin B_I}\frac{\mathrm{b}_i \cdot \mathrm{b}_j}{\|\mathrm{b}_i\|\|\mathrm{b}_j\|}\right) \tag{9}$$

where B_I is the i-th patient's base vector set, and B_I^c is the complement of B_I, i.e. $B \setminus B_I$. This score can reflect the overall base vectors quality from the cognateness for the same individual and the distinction among different individuals. We also use the traditional silhouette coefficient (Rousseeuw, 1987) with euclidean distance metric to quantify clustering quality in the embedded space. The quantitative results are presented in Tab. 2.

Table 2. **Similarity score and silhouette coefficient result.** Best values are bolded, while the second-best are underlined.

Methods	Similarity Score	Silhouette Coefficient
Efficacy of Individual Base Extractor (Fig. 3(a)–(d))		
Identity Mapping	0.0022	-0.1447
IBExtractor w/o training	0.0894	0.2298
IBExtractor	<u>0.2808</u>	<u>0.6439</u>
IBExtractor w/o c_ref	**0.3079**	**0.6738**
Generation Consistency (Fig. 3(e)–(h))		
ECGTwin	<u>0.3334</u>	**0.7401**
ECGTwin w/o c_ref	**0.3440**	<u>0.7380</u>
ECGTwin w/o IBExtractor	0.1689	0.1715
DiffuSETp	0.2412	0.6577

From the comparison of Fig. 3(a) and (c), we observe that **the base vectors extracted by Individual Base Extractor form distinct clusters for each patient**, while the original latents exhibit a stochastic distribution. This entropy reduction illustrates the model's ability to capture personalized ECG characteristics. Fig. 3(b) surprisingly shows immature clustering, which may be affected by reference cardiac condition $\mathbf{c}_{\text{ref}}$. Nevertheless, compared with Fig. 3(c), the substantial difference brought by model training highlights the value of our self-supervised training strategy. We also test the performance of our method in the absence of $\mathbf{c}_{\text{ref}}$, and it even show superior result on these ten patients according to the numerical assessment in Tab. 2, which confirms the model's robustness originating from our masked training. We conduct a further analysis of Individual Base Extractor regarding to $\mathbf{c}_{\text{ref}}$, which is provided in App. E. **Furthermore, to demonstrate the robustness of Individual Base Extractor, we test its resistant ability in App. E.2 when the input ECG is corrupted by noise at different scales.**

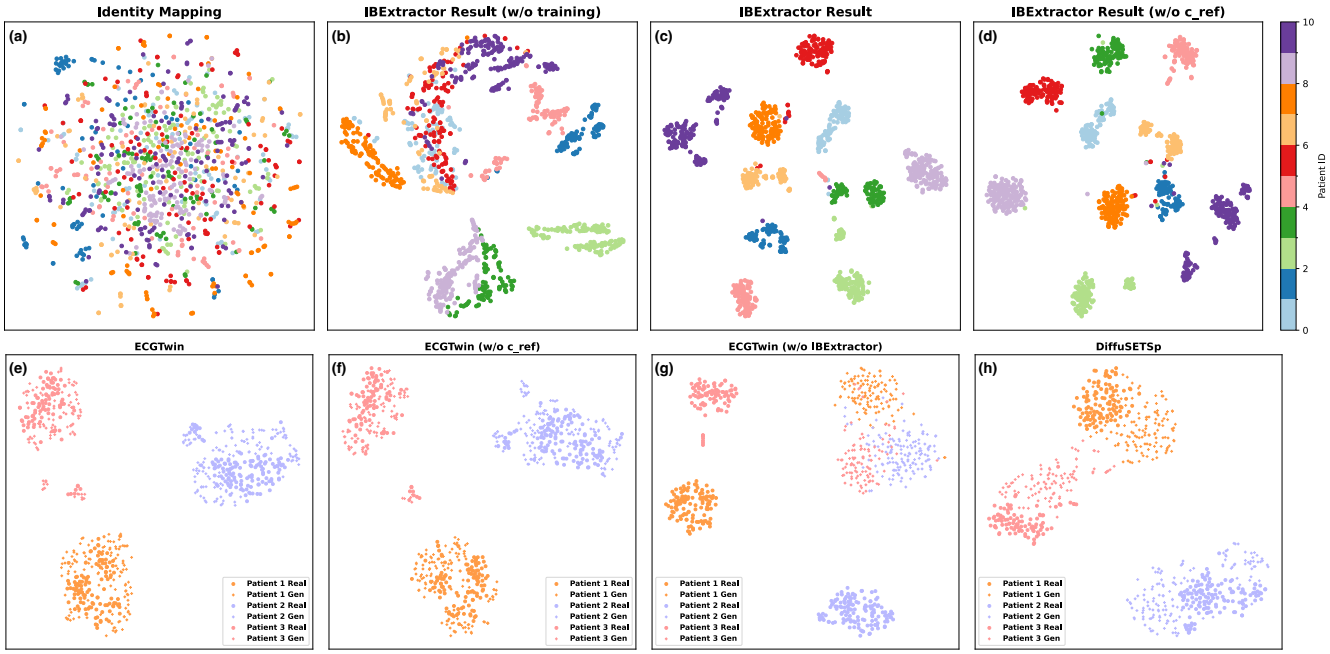


Figure 3. **The visualization result of base vector t-SNE embeddings.** (a)–(d): Base vectors of real ECGs from ten patients; (e)–(h): Base vectors of real ECGs and generated digital twins from three patients

4.3.2. Generation Consistency

Having validated the pre-trained Individual Base Extractor, we now use it to assess personalization consistency in ECG generation. Specifically, we select three patients from test set and generates their ECG digital twins. We then analysis the base vector extracted from both the real and generated ECGs together by t-SNE visualization and numerical evaluation (Fig. 3(e)–(h) and Tab. 2). We include two baselines for comparison: the result when the base vector is zero out and the result of DiffuSET-p.

Our findings show that whether with the auxiliary of $\mathbf{c}_{\text{ref}}$, **ECGTwin can generate ECG accurately preserving patient-specific features** (Fig. 3(e) and (f)). In contrast, when the base vector from stage 1 is omitted (Fig. 3(g)), the generated ECGs lose individualized characteristics, resulting in base vectors that no longer cluster meaningfully. This confirms the necessity of the base vector and the efficacy of the Individual Base Extractor in Stage 1. Finally, in both visual and quantitative terms, the personal patterns retained by DiffuSETSp (Fig. 3(h)) are less distinct than those captured by our proposed method. This supports the effectiveness of our two-stage framework, and again underscores the critical role of the AdaX Condition Injector in conditional generation.

4.4. Personalized Healthcare Simulation

In this section, we simulate a downstream application of personalized ECG diagnosis and show that the ECG digital twins generated by ECGTwin can further enhance the performance of ECG auto diagnosis model through data augmentation. To prevent data leakage, we select 293 patients from the ECGTwin test dataset, each of whom has more than 10 ECG records, for this diagnostic evaluation. A baseline population-level ECG diagnosis model based on ResNet (He et al., 2016) is first trained on the remaining ECG records in test set. This model is tasked with binary classification to distinguish between normal and abnormal ECG signals. For the personalized setting, we designate the earliest ECG of each patient as the reference and generate multiple ECG digital twins using predefined, patient-agnostic cardiac conditions as targets. These target conditions are chosen to include a broad range of abnormalities and ensure that the generated dataset is balanced between normal and abnormal classes. The baseline model is then fine-tuned individually for each patient using their own set of ECG digital twins, resulting in personalized diagnosis models. To further validate the effectiveness of ECGTwin, we compare it against models augmented by ECG synthesized in population-level scope. We report the evaluation results for these 293 patients in Tab. 3 from both patient wise (computing metrics for each patient then averaging) and population wise (directly averaging among all test data).

The result of simulation experiments **validate the potential of personalized healthcare and show that ECGTwin is well-suited for this promising application.** Compared with population-level diagnosis model, the personalized diagnosis model augmented with ECGTwin generated ECG digital

Table 3. **ECG auto diagnosis test.** ECGTwin use DiT architecture as noise predictor backbone. Best values are bolded.

Method Scope	Augmented By	Patient-Wise		Population-Wise	
		Acc.	Macro-F1	Acc.	Macro-F1
Population	N/A (Base Model)	0.755	0.672	0.759	0.753
	DiffuSETS	0.797	0.696	0.800	0.785
Personalized	DiffuSETS-p	0.806	0.694	0.809	0.783
	LAVQ	0.810	0.727	0.808	0.795
	ECGTwin (*Ours*)	**0.816**	**0.731**	**0.819**	**0.804**

twins demonstrates substantial performance improvements. This can be attributed to the personal features-preserving nature of our method, which enables the finetuned model to focus on individual-specific patterns, leading to more targeted and accurate diagnosis. Furthermore, among all personalized augmentation methods, ECGTwin yields the best performance. This result highlights its superior ability to capture the most relevant and distinct personal features via contrastive learning, while generating high-quality ECG signals through effective condition injection.

4.5. Case Study and Interpretability Analysis

In this section, we take the common cardiac event *"Ventricular Premature Complex (PVC)"* as a representative case to illustrate the process of personalized ECG generation and the intrinsic explainability of ECGTwin. *PVC* has a specific appearance of the QRS complexes and T waves on ECG, which are different from normal readings. To simulate this, we generate a ECG digital twin with *PVC* features of one patient, using her normal ECG as reference. Notably, to establish a direct correlation between attention map and raw ECG waveform, we perform the generation on a special version of ECGTwin that operates directly in the signal space without a VAE. Fig. 4(a) shows the complete input setup of personalized ECG generation, including the reference ECG signal, its associated cardiac condition (providing auxiliary description of current state), and the target cardiac condition (specifying the desired cardiac state). The generation result is presented in Fig. 4(b). For clarity, we only display ECG lead V1 here. For holistic view of all 12-leads and **additional case studies**, please refer to App. G.

Next, we visualize the average cross-attention map of report token *PVC* in the *AdaX Condition Injector* at diffusion timestep $T = 100$. As shown in Fig. 4(b), malformed QRS complexes, which is exactly the characteristic pattern of *PVC*, appear at the localizations where the report token *PVC* places most of its attention. This alignment demonstrates that ECGTwin is capable of correctly identifying and localizing clinically meaningful patterns during personalized ECG generation. Such behavior not only enhances the interpretability of the model, which is an essential requirement for clinical adoption (Tonekaboni et al., 2019), but also lays the foundation for a fine-grained post-generation editing, as we further explore and showcase in App. F.

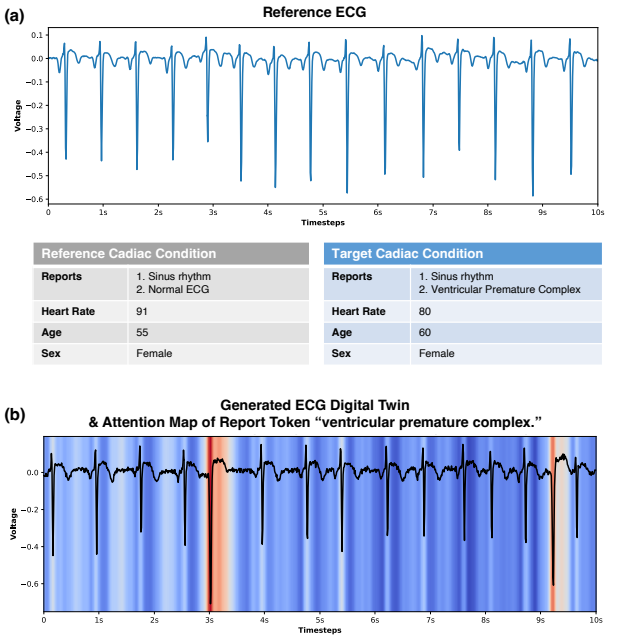


Figure 4. **Case study of personalized ECG generation.** (a): Input data, including the reference ECG, associated cardiac condition, and the target cardiac condition. (b): The ECG digital twin generated by ECGTwin, along with the average cross-attention map of the report token PVC. Redder regions indicate higher amount of attention.

5. Conclusion

In this paper, we present ECGTwin, a novel two-stage diffusion-based model for personalized ECG generation that addresses the challenges of individual feature extraction and multi-condition integration. By combining a contrastively trained *Individual Base Extractor* with the diffusion process equipped by *AdaX Condition Injector*, our method enables controllable and interpretable generation of ECG digital twins that faithfully retain personal characteristics. Through comprehensive experiments, we demonstrate that ECGTwin achieves superior performance in generation quality, while also exhibiting strong downstream utility. These results suggest that ECGTwin is well-positioned to advance the development of personalized healthcare and foster broader adoption of generative models in clinical applications.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alcaraz, J. M. L. and Strodthoff, N. Diffusion-based conditional ecg generation with structured state space models. *Computers in Biology and Medicine*, 163:107115, 2023.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *preprint at arXiv*, 2016.

Chen, J., Liao, K., Wei, K., Ying, H., Chen, D. Z., and Wu, J. Me-gan: Learning panoptic electrocardio representations for multi-view ecg synthesis conditioned on heart diseases. In *International Conference on Machine Learning*, pp. 3360–3370. PMLR, 2022.

Chung, H., Kim, J., Kwon, J.-m., Jeon, K.-H., Lee, M. S., and Choi, E. Text-to-ecg: 12-lead electrocardiogram synthesis conditioned on clinical text reports. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Ding, C., Yao, T., Wu, C., and Ni, J. Advances in deep learning for personalized ecg diagnostics: A systematic review addressing inter-patient variability and generalization constraints. *Biosensors and Bioelectronics*, 271:117073, 2025. ISSN 0956-5663. doi: https://doi.org/10.1016/j.bios.2024.117073. URL https://www.sciencedirect.com/science/article/pii/S0956566324010807.

Fetaya, E., Jacobsen, J., Grathwohl, W., and Zemel, R. S. Understanding the limitations of conditional generative models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1lPleBFvH.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. 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.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Gow, B., Pollard, T., Nathanson, L. A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Berkowitz, S., Moukheiber, D., Eslami, P., et al. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset (version 1.0). *PhysioNet*, 2023. doi: 10.13026/4nqg-sb35.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *preprint at arXiv*, 2022.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y. CLIPScore: A reference-free evaluation metric for image captioning. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hu, Y., Chen, J., Hu, L., Li, D., Yan, J., Ying, H., Liang, H., and Wu, J. Personalized heart disease detection via ecg digital twin generation. In Larson, K. (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 5872–5881. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/649. URL https://doi.org/10.24963/ijcai.2024/649. Main Track.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-w. H., Celi, L. A., and Mark, R. G. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1), 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. URL http://dx.doi.org/10.1038/s41597-022-01899-x.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. *Improved precision and recall metric for assessing generative models*, pp. 3927–3936. Curran Associates Inc., Red Hook, NY, USA, 2019.

Lai, Y., Chen, J., Zhao, Q., Zhang, D., Wang, Y., Geng, S., Li, H., and Hong, S. DiffuSETS: 12-lead ECG generation

conditioned on clinical text reports and patient-specific information. *Patterns*, 2025. URL https://doi.org/10.1016/j.patter.2025.101291.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *preprint at arXiv*, 2023.

Luo, K., Li, J., Wang, Z., and Cuschieri, A. Patient-Specific deep architectural model for ECG classification. *J Healthc Eng*, 2017:4108720, May 2017.

Nussbaum, Z., Morris, J. X., Duderstadt, B., and Mulyar, A. Nomic embed: Training a reproducible long context text embedder. *preprint at arXiv*, 2024.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *Preprint at arXiv*, abs/1508.07909, 2015. URL http://arxiv.org/abs/1508.07909.

Sharma, D. and Kohli, N. Wfdb software for python: A toolkit for physiological signals. In *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*, pp. 86–92. IEEE, May 2023. doi: 10.1109/icsccc58608.2023.10176714. URL http://dx.doi.org/10.1109/icsccc58608.2023.10176714.

Shusterman, V. and London, B. Personalized ecg monitoring and adaptive machine learning. *Journal of Electrocardiology*, 82:131–135, 2024. ISSN 0022-0736. doi: https://doi.org/10.1016/j.jelectrocard.2023.12.006. URL https://www.sciencedirect.com/science/article/pii/S0022073623002662.

Thanh-Tung, H. and Tran, T. Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, pp. 1–10. IEEE, 2020.

Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pp. 359–380. PMLR, 2019.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., and Schaeffter, T. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1), 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0495-6. URL http://dx.doi.org/10.1038/s41597-020-0495-6.

Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D., and Ghassemi, M. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, 30(10):2838–2848, October 2024.

A. Related Work

A.1. ECG Generative Methods

ECG generation methods have evolved significantly over the years, progressing from early unconditional approaches or those conditioned on high-level labels (Chen et al., 2022; Alcaraz & Strodthoff, 2023) to more refined control using textual clinical reports (Chung et al., 2023; Lai et al., 2025). Some researchers (Hu et al., 2024) have explored personalized ECG generation using Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) combined with Vector Quantization (VQ) based methods. In their framework, a VQ-variational autoencoder (VAE) (van den Oord et al., 2017) is used for feature disentanglement and should be optimized jointly with GAN components, which would complicate the already delicate training process typical of GANs. Furthermore, their approach requires an additional input ECG reflecting the target cardiac condition as part of the conditioning signal, which may be impractical when the condition is complex or rare. Beyond GANs, researchers have applied more expressive architectures, such as autoregressive models (Chung et al., 2023) and diffusion models (Alcaraz & Strodthoff, 2023; Lai et al., 2025), to conditional ECG generation. However, their works do not incorporate a reference ECG as input and therefore cannot achieve personalized generation. In our work, we use the powerful diffusion model for ECG generation and build an Individual Base Extractor to obtain the personalized ECG base vector, which is trained separately using contrastive learning, thus will not add extra burden to the optimizing of diffusion process.

A.2. Conditional Generation

Methods for incorporating conditional information into generative models can generally be categorized into three types (Peebles & Xie, 2023): (1) In-context conditioning. The condition signals are directly appended to the input, either along the sequence length axis or the embedding (channel) dimension. (2) Adaptive normalization. The learnable parameters in normalization layers—such as the scale and shift parameters γ and β in LayerNorm (Ba et al., 2016)—are dynamically modulated by the condition signals. (3) Cross-attention. The model architecture is augmented with additional multi-head cross-attention layers (Vaswani et al., 2017), where the condition signals are used as the Key and Value inputs, allowing the model to attend selectively to the conditioning information during generation.

Prior work (Peebles & Xie, 2023) has conducted extensive experiments demonstrating the superiority of zero-initialized adaptive LayerNorm in terms of both generated image fidelity and computational efficiency. However, other researchers (Hertz et al., 2022) have also shown that cross-attention layers offer more nuanced control, particularly due to the interpretability of the attention maps, which capture meaningful relationships between the input and the conditional signals. Moreover, a common defect for these methods is that they could not handle complex, heterogeneous conditions. Therefore, in our work, we develop the AdaX Condition Injector, combining the benefit of both adaptive LayerNorm and cross-attention and processing different types of conditions in separate pathways, to effectively integrate all the conditions concerning with the personalized generation process.

B. ECGTwin Framework

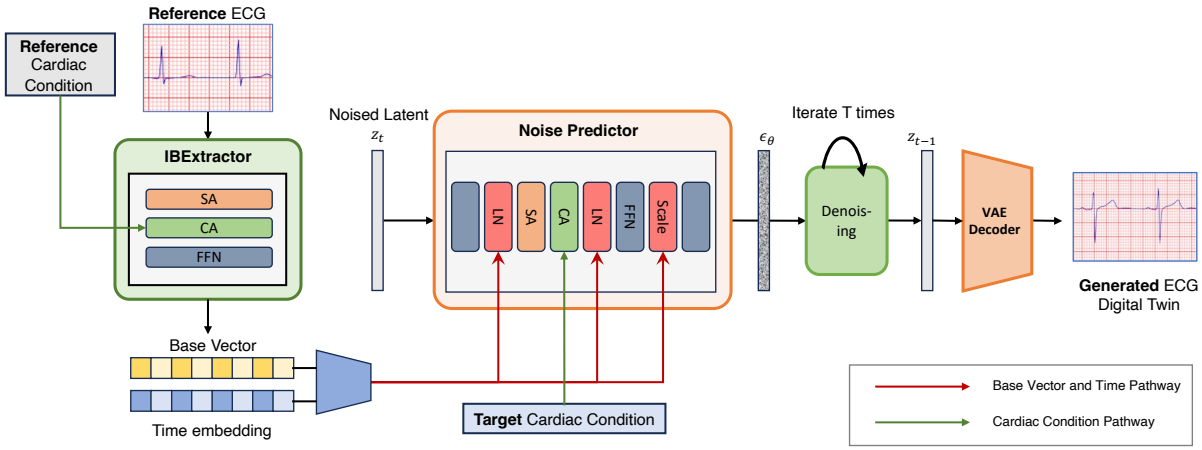


Figure 5. **The flow chart of ECGTwin.** The Individual Base Extractor first extracts the base vector of the patient from reference ECG and reference cardiac condition, then the latent diffusion model with AdaX Condition Injector integrates the base vector, current diffusion timestep and target cardiac condition to generate ECG digital twins by repeatedly denoising a latent sampled from Gaussian distribution.

C. Experiment details

C.1. Dataset

C.1.1. MIMIC-IV-ECG

We use the publicly available **MIMIC-IV-ECG** dataset (Gow et al., 2023) for the training and evaluation of our method. MIMIC-IV-ECG contains 800,035 records from 159,538 unique patients, where each record comprising a 10-second 500Hz 12-lead ECG waveform and 1 to 17 associated clinical text reports written in English. First we downsample the ECG waveform to 102.4Hz, resulting a ECG signal matrix $\in \mathbb{R}^{1024\times12}$. We then link the MIMIC-IV-ECG dataset to MIMIC-IV-Clinical dataset (Johnson et al., 2023; Goldberger et al., 2000) via subject_id to retrieve the ECG owner's age and sex, and compute the heart rate using recorded RR interval. When RR interval shows anomaly (i.e. 0 or 65536 ms), we directly parse the ECG waveform by utilizing the XQRS detector from WFDB toolkit(Sharma & Kohli, 2023) to manually obtain QRS interval then compute the heart rate. After preprocessing, we retain 794,372 ECG records with complete and valid information. These are split into 744,372 records for preparation of training dataset and 50,000 records for preparation of testing dataset.

Dataset	MIMIC-IV-ECG		PTB-XL
	Train split	Test split	
# of ECG Records	744,372	50,000	21799
# of ECG Pairs	6,408,782	399,499	4269
Avg. # of ECG Records Per Patient	4.98	4.89	1.16
Max # of ECG Records Per Patient	260	168	10

Table 4. **Statistics of MIMIC IV ECG and PTB-XL Dataset**

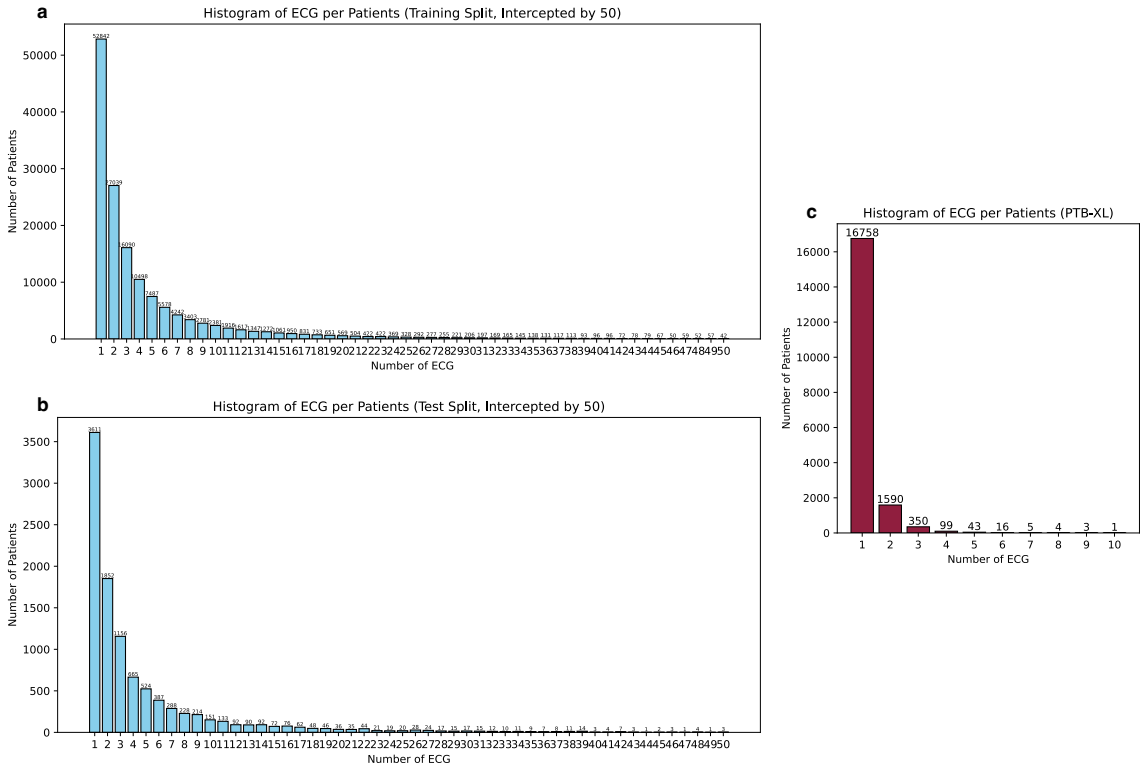


Figure 6. **The histogram of per patient ECG records number.** (a) and (b): MIMIC-IV-ECG dataset; (c): PTB-XL dataset.

To construct paired samples for training the Individual Base Extractor and DDPM noise predictor, we group the ECG records by subject_id, and exclude patients with only a single record. The distribution and statistics of ECG counts per patient

are shown in Table 4 and Figure 6. For each remaining patient, we generate all possible pairs of ECG records, ordered chronologically to reflect potential temporal causality in the conditional distribution $P(\hat{\mathbf{x}}|\mathbf{x}_{\text{ref}}, \mathbf{c}_{\text{ref}}, \mathbf{c}_{\text{tar}})$. This results in $\binom{n_i}{2} = \frac{1}{2}n(n-1)$ samples for a patient with n_i records. Following this procedure, we obtain a training set of 6,408,782 ECG record pairs and a testing set of 399,499 ECG pairs based on the aforementioned split.

C.1.2. PTB-XL

We use PTB-XL dataset (Wagner et al., 2020) for external validation. It contains 21,799 12-lead ECG records from 18,885 individuals. We use the waveform data in 10-second 500Hz format and downsample them into 102.4Hz. Each record is associated with a single clinical text report, written either in English or German, along with 1–3 labels that describe the record from diagnostic, form and rhythm perspectives. To enhance the textual condition, we expand each label into a descriptive phrase and treat these as additional text reports. Consequently, after preprocessing, each record is accompanied by 2–4 clinical reports. Notably, since the PTB-XL labels do not include heart rate related values, we compute the heart rate directly from the waveform data. The pairing strategy for constructing ECG pairs in PTB-XL follows the same procedure as used for the MIMIC-IV-ECG dataset. All the ECG pairs are used for test without further splitting. The distribution and statistics of ECG record counts per patient are also shown in Table 4 and Figure 6.

C.2. Implementation details

For Individual Base Extractor at stage 1, the layer of Transformer encoder is set to 3 and the model dimension, i.e. the resulting base vector dimension, is set to 256. We use a large batch size of 65,536 and use the gradient accumulation trick with mini batch of 512 for compatibility to our CUDA memory. Moreover, as stated previously, we randomly substitute the $\mathbf{c}_{ref}$ with a learnable embedding at probability of 0.15 to enhance the robustness of model. We use adamW with learning rate of 1e-3 for optimization and set number of training epochs to 40.

For the latent diffusion model at stage 2, we test both prevailing style for noise predictor, Unet (Ronneberger et al., 2015) and DiT (Peebles & Xie, 2023). The training details of these two architectures are listed in Table 5. We adopt the pre-trained VAE model with latent space of $\mathbb{R}^{4\times128}$ from DiffuSETS (Lai et al., 2025) without further parameter-tuning. The number of time step T in training phase is set to 1000, and the noise β_t of diffusion forward process is assigned to linear intervals of $[8.5 \times 10^{-4}, 1.2 \times 10^{-2}]$. Also, we randomly mask the base vector from stage 1 with zero vector at a ratio of 0.15 to retain the model's ability for non-personalized generation.

Among all hyperparameters, we observe that the learning rate has the most significant impact on generation quality. The noise predictor model fails to converge when the learning rate is set either too high or too low, and we perform a grid search to identify the best setting for stable training and good performance. For other hyperparameters, such as diffusion noise schedules, we follow configurations from previous works (Lai et al., 2025; Peebles & Xie, 2023). We acknowledge that more intensive hyperparameter tuning could further improve ECGTwin's performance. However, in this paper, our primary goal is to introduce the method and demonstrate its efficacy. Hence this additional optimization is left to future work for the research community interested in ECG digital twin generation.

All of our trainings and tests are implemented with PyTorch 2.1.1, on GeForce RTX 3090. During inference phase, DiT based model iterates approximately 140 time steps per second when generating batch is set to 10 while Unet based model iterates approximately 60 time steps per second on the same environment. Therefore, without explicitly specifying, the ECGTwin refers to the DiT version model.

Methods	**DiT**		**UNet**	
# Epochs	30		30	
Batch Size	512		1024	
Learning Rate	1e-4		5e-4	
Model Size	∼9M		∼8M	
Model-Specific	Model Dim.	256	kernel size	7
	# Layers	7	# Levels	6

Table 5. **Hyperparameters of different noise predictor backbones.**

D. Supplement of Quality Evaluation

D.1. Three-level Evaluation

We employ the three-level evaluation (Lai et al., 2025) for the generated ECG twin. The signal level focuses on the fidelity and stability of the generated signals by evaluating the distribution similarity (Fréchet Inception Distance Score, FID) and the structural resemblance between real and generated ECG signals (The improved Precision, Recall and F1 (Kynkäänniemi et al., 2019)). The feature level examines whether the ECG signals generated by the model align with the input descriptions of patient-specific information. In this case, we test the Mean Absolute Error (MAE) between the Heart Rate extracted from the generated ECG and the heart rate specified in target cardiac condition (HR-MAE). Last but not least, the diagnostic level use CLIP scores (Hessel et al., 2021) to assess the semantic alignment between the generated ECG signals and the clinical text reports in target cardiac condition. All the tests are repeated 5 times with the mean and standard deviation of results computed.

D.2. Ablated Models

To provide a penetrating analysis of our proposed AdaX Condition Injector, we design two type of condition injectors, each inherits one of its dedicated condition pathways. The model with suffix *CA* preserves the report-level tokenization trick and the cross attention mechanism for cardiac condition injection while simply adding the time embedding and base vector to the noisy latent representation after necessary linear projection for dimension adaptation. The model with suffix *adaLN* uses adaptive LayerNorm to integrate all the conditions. In this case, we concatenate all the reports together and obtain one text embedding. Successive process is the same as the Equation 2 except that the condition sequence reduces into a single vector. The resulting cardiac condition vector alongside the time embedding and base vector are finally summed together and forwarded into an MLP for modulating the factors of layernorm and scaling gate.

D.3. External Validation

To evaluate the generalization capability of ECGTwin, we perform the three-level evaluation on an external dataset, PTB-XL, without any additional fine-tuning. A detailed introduction to PTB-XL is provided in Appendix C. Notably, the most significant difference between the MIMIC-IV-ECG and PTB-XL datasets lies in the distribution and language of clinical text reports associated with each ECG record. As shown in Table 6, ECGTwin maintains strong performance on this out-of-distribution dataset, whereas the ablated models exhibit substantial performance degradation. This performance gap underscores the challenge of effectively injecting diverse condition types and further validates the necessity and synergy of the two-pathway design in the AdaX Condition Injector. It is also worth noting that CLIP Score metrics on the external validation set are slightly lower than those in Table 1 across models trained on MIMIC-IV-ECG dataset. This may be attributed to the fact that many clinical text reports in PTB-XL are written in German, a language for which the employed text embedding model may not be fully optimized. Nevertheless, ECGTwin achieves the best semantic alignment performance among all baselines and retains competitive absolute scores, further confirming its robustness and adaptability.

Model Name	Signal Level				Feature Level	Diagnostic Level
	FID (↓)	Precision (↑)	Recall (↑)	F1 (↑)	HR-MAE (↓)	Clip (↑)
DiffuSETSp	335±69	0.690±0.009	0.917±0.012	0.784±0.006	13.76±0.66	0.689±0.003
LAVQ	588±02	**0.844**±0.004	0.310±0.023	0.455±0.022	39.82±0.03	0.718±0.002
DiT-*CA*	416±34	0.706±0.008	**0.950**±0.008	0.810±0.007	7.35±0.34	0.723±0.002
DiT-*adaLN*	199±26	0.686±0.013	0.948±0.005	0.796±0.009	**6.88**±0.46	0.708±0.002
ECGTwin-*DiT*(*Ours*)	**41**±13	0.743±0.009	0.949±0.008	**0.833**±0.007	10.23±0.59	**0.729**±0.002

Table 6. **Three-level evaluation result on PTB-XL dataset.** Best values are bolded, while the second-best are underlined.

D.4. Generated Distribution Visualization

We randomly select 1,000 ECG pairs from the test dataset and generate one ECG digital twin for each pair, conditioned on the reference ECG latent, reference cardiac condition, and target cardiac condition. We then visualize the real target ECG latents and the latents of the generated ECG digital twins using t-SNE. As shown in Figure 7, the distribution of the generated latents closely aligns with that of the real ECG latents, with generated samples evenly dispersed among the real

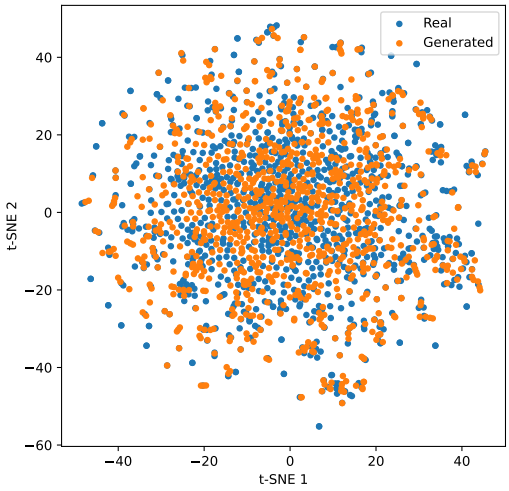


Figure 7. **The scatters of generated ECG digital twin latents and real target ECG latents.**

ones. This result provides strong evidence that ECGTwin is capable of producing realistic ECG signals that resemble real data from a latent space perspective.

D.5. Heart Rate Scatters

To evaluate whether ECGTwin can generate ECG digital twins that accurately reflect the input heart rate, we visualize scatter plots of generated versus target heart rates in Figure 8. Across both datasets, the data points closely align with the identity line ($y = x$), indicating that ECGTwin effectively interprets the heart rate specified in the input target cardiac condition and generates ECG digital twins that adhere to this rhythm constraint. While ECGTwin may not achieve the lowest HR-MAE among all methods, the deviation is minor and does not impair its ability to model the intrinsic relationship between input heart rate value and waveform periodicity. This confirms the model's competence in generating temporally accurate and physiologically plausible ECG signals.

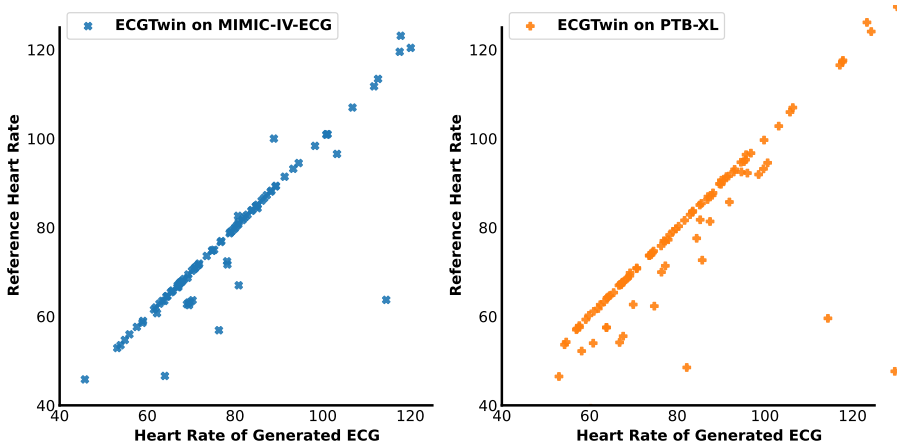


Figure 8. **The scatters of heart rate in generated ECG and input target cardiac condition.**

E. Supplementary Analysis of Individual Base Extractor

E.1. Model Input Analysis

In this part, we focus on the interesting observation in Table 2, that is, the Individual Base Extractor shows a better result when reference cardiac condition $\mathbf{c}_{\text{ref}}$ is absent. We try to figure out whether it is because possible bias from $\mathbf{c}_{\text{ref}}$ or it is just a coincidence related to specific test patient group. Note that we have verified in the main text that the model can capture personalized ECG characteristics in both cases, hence this analysis is for deeper exploration and is never intended to provide an arena for performance duelling. After all, they are the same model with the same weights but in different working status.

E.1.1. Discussion on Reference Cardiac Condition

We first justify the reference cardiac condition $\mathbf{c}_{\text{ref}}$ used in Individual Base Extractor. The intention of $\mathbf{c}_{\text{ref}}$ is to explicitly designate the spurious factor of which the base vector should be invariant. However, concern may be raised based on the potential data bias: in the real dataset, the ECG records of the same patient may tend to have similar cardiac conditions because the sampling is unevenly conducted on certain in-hospital period, during which the cardiac condition change is not substantial regarding to the person's lifelong time. This data bias may lead model to learn a wrongly shortcut by relying on $\mathbf{c}_{\text{ref}}$, which is entirely opposite to the original purpose.

In spite of the potential bias in the training data, here we show that our model is not affected by it from test time observations. We examine the reference cardiac conditions of the ten patients used in Section 4.3 efficacy experiment. We find that they not only differ intra-patient to some extent, but also show overlapping inter-patients. This implies that the distributions of $\mathbf{c}_{\text{ref}}$ from patient trajectories is more dispersed and intersected than expectation, which may hugely ease the intrinsic bias. Figure 3(b) further confirms our assumption, since the entangled clustering result somehow reflects the distribution of $\mathbf{c}_{\text{ref}}$. However, in Figure 3(c), points form clusters with distinct edge after training. This disentanglement indeed demonstrate that model can avoid the data bias, and thus we can tell that the performance gap may not come from data bias introduced by $\mathbf{c}_{\text{ref}}$.

E.1.2. Scaling Up Test

We then provide a comprehensive evaluation of the performance under two input cases by selecting different numbers of patient and repeating the efficacy experiment in Section 4.3. As illustrated in Figure 9(a) and (b), when the test group scales up, both model can capture individual-specific patterns with respect to the intra-individual similarity and inter-individual dis-similarity. Specifically, Figure 9(c) and (d) show the t-SNE visualization of the 20 patient result, which are both impressive considering the task difficulty. From this extended assessment, we can observe that the model **with** $\mathbf{c}_{\text{ref}}$ shows preferable according to similarity score while the model **without** $\mathbf{c}_{\text{ref}}$ performs better on the scope of cluster silhouette. From this observation, we can **conclude that each method has its own merits, and considering the design of personalized ECG generation task, we choose to include $\mathbf{c}_{\text{ref}}$ as long as it is applicable.**

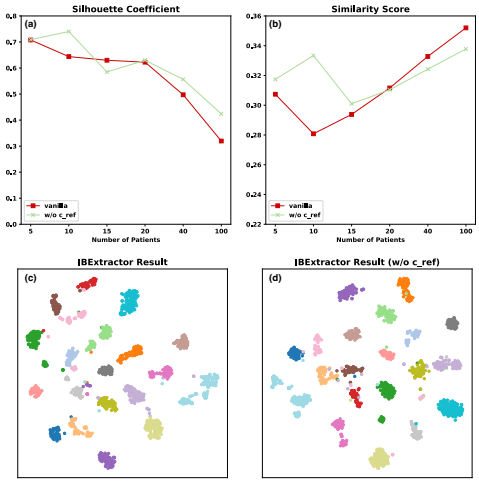


Figure 9. **The scaling up test on Individual Base Extractor with and without reference cardiac condition.** (a) and (b): Quantitative result; (c) and (d): t-SNE visualization of ECG base vector from twenty patients.

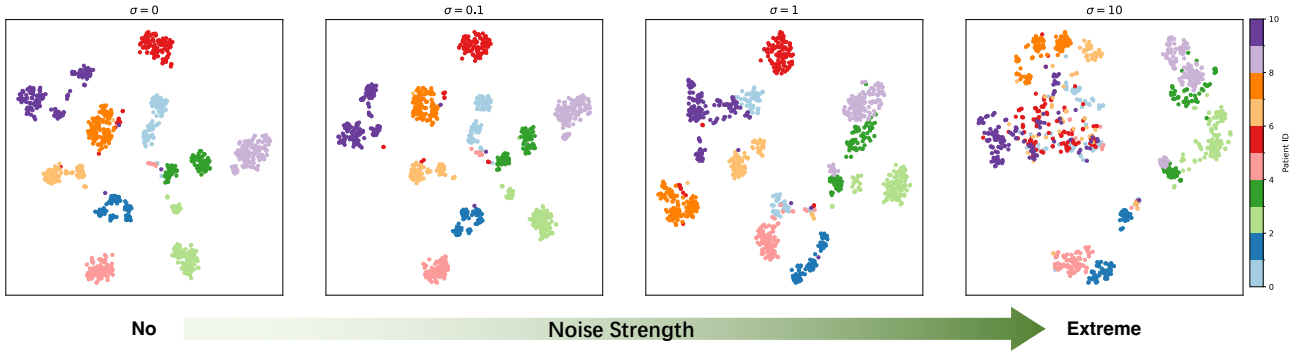


Figure 10. **Visualization of base vector clustering under different noise level.** $\sigma = 0$ (no noise), $\sigma = 0.1$ (mild noise), $\sigma = 1$ (strong noise), and $\sigma = 10$ (extreme noise).

E.2. Model Robustness Evaluation

We evaluate the robustness of the Individual Base Extractor by assessing its resistance to noise when the reference ECG $\mathbf{x}\text{ref}$ is corrupted at different noise levels. Specifically, we simulate common signal degradations encountered in clinical practice, such as electrical interference and device noise, by adding Gaussian noise with varying standard deviation σ directly to $\mathbf{x}\text{ref}$ in the signal space. We then visualize the clustering behavior of the extracted base vectors $\mathbf{b}$, following the setup in Section 4.3. Considering the typical voltage amplitude of ECG signals, we examine four noise regimes: $\sigma = 0$ (no noise), $\sigma = 0.1$ (mild noise), $\sigma = 1$ (strong noise), and $\sigma = 10$ (extreme noise).

As shown in Fig. 10, the Individual Base Extractor remains largely unaffected under mild noise ($\sigma = 0.1$) and continues to capture distinct patient-specific features even under strong noise conditions ($\sigma = 1$). Under extreme noise ($\sigma = 10$), although some feature overlap across patients emerges, the relative similarity inter patient groups are preserved. Overall, these results demonstrate that **the Individual Base Extractor is robust to substantial input corruption**. Moreover, because the feature extraction stage is decoupled from the generation stage, **this robustness further contributes to the stability and quality of the ECG digital twins produced by ECGTwin.**

F. Prompt-to-Prompt ECG Editing

In this section, we provide technical details and experimental results regarding prompt-to-prompt editing (Hertz et al., 2022) in ECGTwin. This editing mechanism offers enhanced controllability over the generation process by allowing modifications to the generated results through natural language descriptions. Importantly, this functionality can be seamlessly integrated into ECGTwin, thanks to the Cardiac Condition Pathway in the AdaX Injector and the use of report-level tokenization. These design choices enable precise and interpretable conditioning based on textual prompts (report tokens in our case), facilitating more intuitive and flexible ECG manipulation after the initial generation.

F.1. Theory of Prompt-to-Prompt ECG Editing

The theoretical basis of prompt-to-prompt editing lies in the directional behavior and interpretability of cross-attention mechanisms. Typically, the attention map M is defined as:

$$M = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \tag{10}$$

where d is the model dimension. According to this definition, the entry M_{ij} represents the attention weight of the j-th report token on the i-th latent representation. As the cross-attention operation in AdaX Injector Cardiac Condition Path determines how report tokens (i.e., prompts) influence the generated ECG, we can manipulate the generation result by injecting the attention maps M that were obtained during the generation with the original prompt $\mathcal{P}$, into a second generation with the modified prompt $\mathcal{P}^*$. This allows the synthesis of an edited ECG signal $\mathbf{x}^*$ that both reflects the changes introduced by the new prompt and preserves the structural characteristics of the originally generated ECG $\mathbf{x}$.

The algorithm of prompt-to-prompt ECG editing is detailed in Algorithm 1. We denote by $DM(z_t, P, t, s)$ the computation of a single step t of the diffusion process, which outputs the noisy latent z_{t-1}, and the attention map M_t (omitted if not used). We also define by $M(z_t, \mathcal{P}, t, s)\{M \leftarrow \widehat{M}\}$ the diffusion step where we override the attention map $\widehat{M}$ with an additional given map M, but keep the values V from the supplied prompt. Furthermore, we define $Edit(M_t, M^*t, t)$ to be a general edit function, receiving as input the t-th attention maps of the original and edited images during their generation. The specific edit function used in our work are modeled in Equation 11. Specifically, to add a new report token, the attention map of new report token M_t^* is appended to the source attention map M_t along the token axis after a predefined diffusion timestep τ:

$$Edit(M_t, M_t^*, t) := \begin{cases} \text{Concat}(M_t, M_t^*) & \text{if } t < \tau \\ M_t & \text{if } t \geq \tau \end{cases} \tag{11}$$

Algorithm 1 Prompt-to-Prompt ECG editing

Require: A source prompt $\mathcal{P}$, a target prompt $\mathcal{P}^*$, and a random seed s.
Ensure: A source ECG latent z_0 and an edited image z_0^*.
1: $z_T \sim \mathcal{N}(0, I)$ a unit Gaussian random variable with random seed s;
2: $z_T^* \leftarrow z_T$;
3: **for** $t = T, T-1, \ldots, 1$ **do**
4: $\quad z_{t-1}, M_t \leftarrow DM(z_t, \mathcal{P}, t, s)$;
5: $\quad M_t^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s)$;
6: $\quad \widehat{M}_t \leftarrow Edit(M_t, M_t^*, t)$;
7: $\quad z_{t-1}^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s_t)\{M \leftarrow \widehat{M}_t\}$;
8: **end for**
9: **Return** (z_0, z_0^*)

This technique allows a modulation to the ECGTwin-generated ECG to reflect the changes introduced by the new report token while preserving the overall structural and personal characteristics. Notably, we employ this method to edit the generated ECG, enabling more nuanced control over the ECG generation process. However, this approach can also be applied to edit real ECG signals with the help of DDIM reverse. This opens up the possibility for an alternative method for personalized ECG generation, which we plan to explore in future work.

F.2. Post-Generation Editing: Holistic View

In this section, we showcase the usage of prompt-to-prompt editing in ECGTwin by a typical cardiac disease *Right Bundle Branch Block (RBBB)*. The source ECG digital twin is generated under clinical tokens of *sinus rhythm, Low QRS voltages in precordial leads, abnormal ecg* in target cardiac condition, and we add a new report token of *Right bundle branch block* to it by performing post-generation ECG editing described aforementioned. As illustrated in Fig. 11, the resulting signal exhibits clear RBBB patterns while maintaining consistency with the original signal in terms of phase and morphology for regions unrelated to the new condition. Notably, all 12 leads exhibit coherent and semantically meaningful alterations, demonstrating the holistic nature of the modification. This capability for post-generation editing arises naturally from the Cardiac Condition Pathway and offers an additional layer of control over the generation process. It opens up new opportunities for applications such as causal analysis or being an interactive tools for cardiology education, both of which are exciting directions for future exploration.

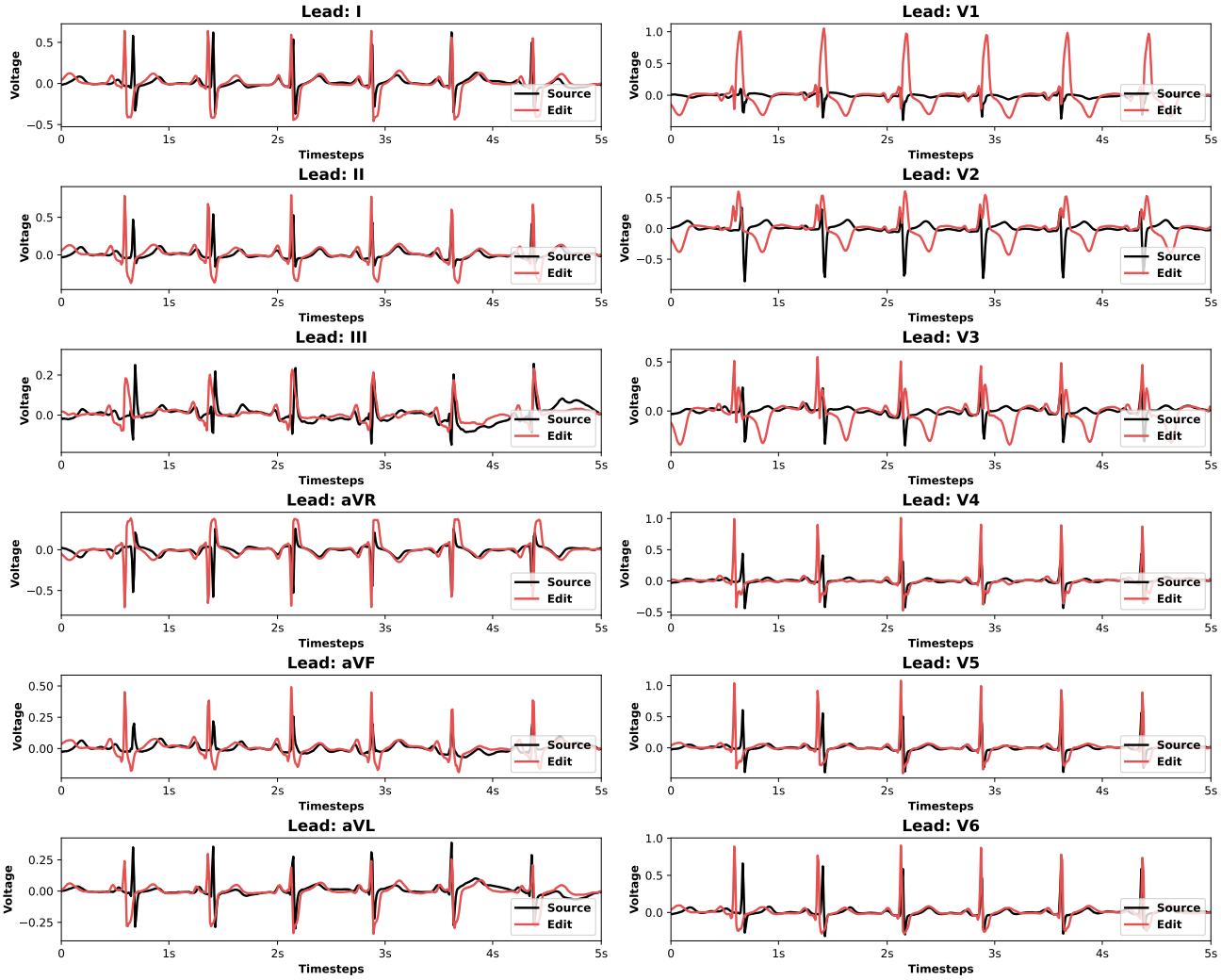


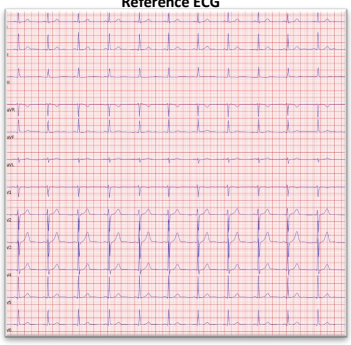
Figure 11. **Post-generation editing example.**

G. Personalized ECG Generation Case Studies

Input

Patient_ID: 10093425

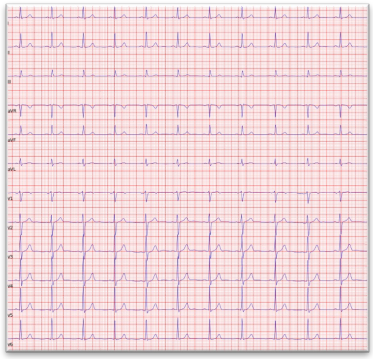
Reference ECG



Reference Cadiac Condition	Value
Reports	1. Sinus rhythm 2. Short PR interval 3. Borderline ECG
Heart Rate	68.99
Age	56
Sex	Male

Target Cadiac Condition	Value
Reports	1. Sinus rhythm 2. Normal ECG
Heart Rate	68.55
Age	60
Sex	Male

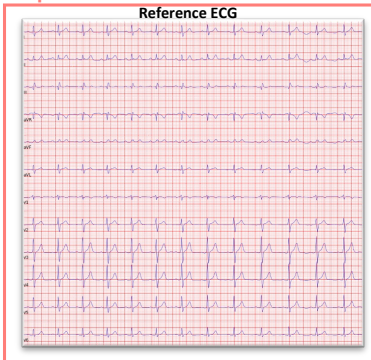
ECG Digital Twin Generated by ECGTwin



Input

Patient_ID: 10070581

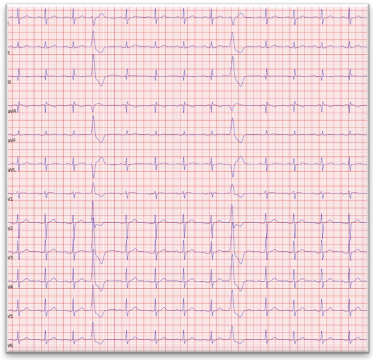
Reference ECG



Reference Cadiac Condition	Value
Reports	1. Sinus rhythm 2. Normal ECG
Heart Rate	79.82
Age	59
Sex	Male

Target Cadiac Condition	Value
Reports	1. Sinus rhythm with frequent PVCs 2. Abnormal ECG
Heart Rate	79.38
Age	59
Sex	Male

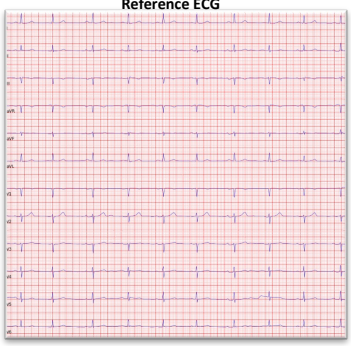
ECG Digital Twin Generated by ECGTwin



Input

Patient_ID: 10034469

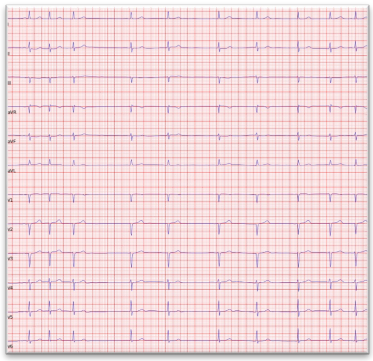
Reference ECG



Reference Cadiac Condition	Value
Reports	1. Sinus bradycardia 2. Possible anterior infarct - age undetermined 3. Low QRS voltages in precordial leads 4. Abnormal ECG
Heart Rate	58.20
Age	76
Sex	Female

Target Cadiac Condition	Value
Reports	1. Atrial fibrillation 2. Possible anterior infarct - age undetermined 3. Abnormal ECG
Heart Rate	66.56
Age	77
Sex	Female

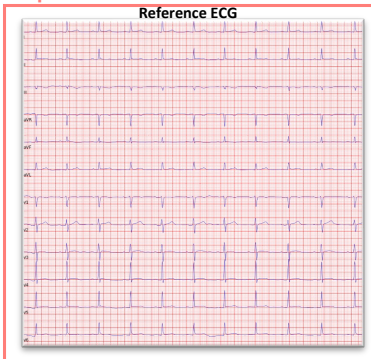
ECG Digital Twin Generated by ECGTwin



Input

Patient_ID: 10578325

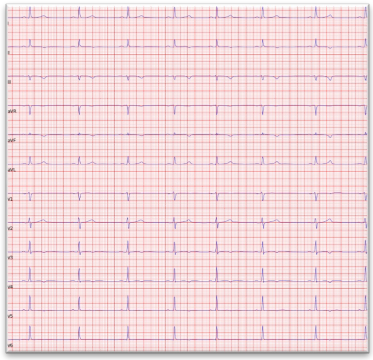
Reference ECG



Reference Cadiac Condition	Value
Reports	1. Sinus rhythm 2. Inferior T wave changes are nonspecific 3. Borderline ECG
Heart Rate	64.83
Age	41
Sex	Male

Target Cadiac Condition	Value
Reports	1. Sinus bradycardia 2. Prolonged QT interval 3. Extensive T wave changes are nonspecific 4. Low QRS voltages in precordial leads 5. Borderline ECG
Heart Rate	45.78
Age	43
Sex	Male

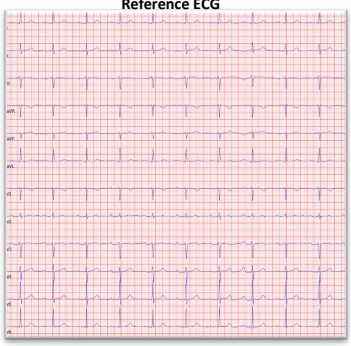
ECG Digital Twin Generated by ECGTwin



Patient_ID: 10511944

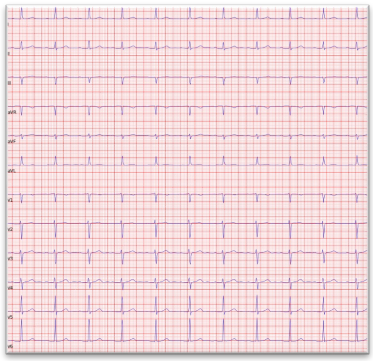
Input

Reference ECG



Reference Cadiac Condition	Value
Reports	1. Possible ectopic atrial rhythm 2. Possible anterior infarct - age undetermined 3. Abnormal ECG
Heart Rate	62.24
Age	74
Sex	Male
Target Cadiac Condition	**Value**
Reports	1. Sinus rhythm with 1st degree A-V block 2. Possible right atrial abnormality 3. Poor R wave progression - probable normal variant 4. Septal T wave changes are nonspecific 5. Abnormal ECG
Heart Rate	65.34
Age	74
Sex	Male

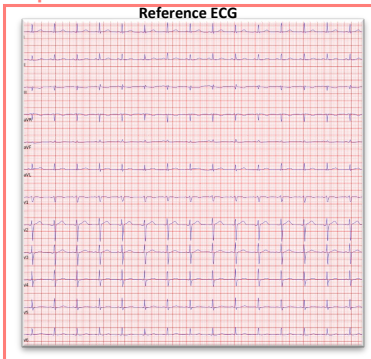
ECG Digital Twin Generated by ECGTwin



Patient_ID: 10048001

Input

Reference ECG



Reference Cadiac Condition	Value
Reports	1. Warning: Data quality may affect interpretation 2. Sinus rhythm 3. Short PR interval 4. Inferior T wave changes are nonspecific 5. Borderline ECG
Heart Rate	90.96
Age	65
Sex	Male
Target Cadiac Condition	**Value**
Reports	1. Probable supraventricular tachycardia 2. Extensive ST-T changes are nonspecific 3. Abnormal ECG
Heart Rate	158.27
Age	65
Sex	Male

ECG Digital Twin Generated by ECGTwin

