

Atoms and Grammar of the Heart: Geometry-Aligned Discrete Tokenization for ECG Language Modeling

Zirui Wang^{*,1,2}, Qiuling Yang^{*,3}, Pang Wu¹, Peng Wang¹
Zhenfeng Li¹, Xianxiang Chen¹, Lidong Du¹, Zhen Fang^{†,1,2}

¹Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China

²School of Electronic, Electrical and Communication Engineering,
University of Chinese Academy of Sciences, Beijing, China

³Institute of Materia Medica, Chinese Academy of Medical Sciences and
Peking Union Medical College, Beijing, China

wangzirui23@mails.ucas.ac.cn; qlyang@imm.ac.cn; zfang@mail.ie.ac.cn

Abstract

Extending discrete tokenization and autoregressive modeling to signals whose latent states follow continuous trajectories introduces phase-morphology ambiguity. Similar local shapes can occur at distinct trajectory positions, while the same position can vary across instances. Standard vector quantization (VQ) may therefore merge contextually different patches and split corresponding ones. We trace this behavior to radial-to-tangential gradient leakage under the straight-through estimator and propose **Geometry Alignment Quantization (GAQ)**. GAQ combines a learned Mahalanobis reshaping of the codebook with a Householder backward map. The map exactly cancels the leakage in the idealized setting, while the numerical stabilization used in practice introduces a bounded deviation. We instantiate GAQ on ECG as **AtomECG**, a tokenizer of sub-waveform atoms composed by a GPT grammar. Atom sequences show long-context scaling from perplexity 488 to 24, split-replicated motif enrichment yields auditable clinical hypotheses, and AtomECG outperforms eight eSSL baselines on more than ten diagnostic tasks. A known-phase limit cycle and PPG results further support GAQ on phase-like signals beyond ECG.

1 Introduction

Discrete tokenization plus autoregressive language modeling is central to modern NLP. A largely unexplored question is how well this recipe extends to physiological signals whose latent state follows a continuous trajectory, such as heartbeats, gait, respiration, and neural oscillations. We identify *phase-morphology ambiguity* as a failure mode of off-the-shelf vector quantization on such signals and develop a geometry-aligned remedy.

The ambiguity is twofold. Sub-waveform patches with similar local shapes can recur at different trajectory positions and carry distinct meanings. Patches at the same position also undergo morphological drift across instances while retaining a shared role. Standard VQ can merge the first case into one token and divide the second across many. We use the electrocardiogram (ECG) (Goldberger et al., 2023; Schlöpfer and Wellens, 2017) as the primary testbed because its phase trajectory is periodic, dense, and clinically annotated (Mehari and Strodthoff, 2022). We also evaluate a known-phase limit cycle and photoplethysmography (PPG). We call the resulting sub-waveform units **atoms**, analogous to subwords, and the autoregressive model over them the cardiac **grammar**.

To realize this view we propose **AtomECG** (Fig. 1), which uniformly patches the raw signal into units significantly shorter than a cardiac cycle and quantizes each patch into an atom from a learned codebook. The technical core is **Geometry Alignment Quantization (GAQ)**, our resolution of phase-morphology ambiguity. GAQ couples a global linear reshaping of the codebook (a learned Mahalanobis metric) with a per-sample, geometry-aware Householder backward map that replaces the straight-through estimator (STE). The resulting atoms are simultaneously morphologically consistent and contextually precise. Following the masked-then-autoregressive recipe of NLP, AtomECG uses masked atom modeling for tokenizer pre-training (akin to Devlin et al. (2019)) and next-atom prediction with a GPT-style grammar (akin to Radford et al. (2018) and Esser et al. (2021)).

Two NLP-native probes test whether AtomECG’s discrete representation supports language-style analysis. Token-level perplexity decreases from 488 to 24 as context lengthens. Split-replicated χ^2 enrichment identifies cohort-

*Equal contribution.

†Correspondence: zfang@mail.ie.ac.cn

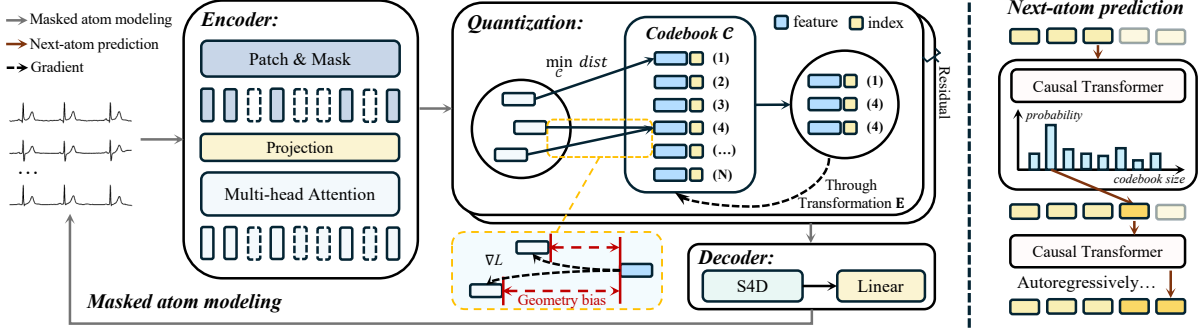


Figure 1: AtomECG uniformly patches the signal and encodes each patch. At every residual level, GAQ assigns a code in the space shaped by $\mathbf{E}$, preserves the assigned code in the forward pass, and transports encoder gradients with $\mathbf{R}_\theta$. The summed codes are decoded for reconstruction. Level-0 indices form the atom sequence used by the GPT grammar.

associated atom motifs whose clinical interpretations remain hypotheses for expert validation. AtomECG also outperforms eight leading eSSL baselines under linear probing on more than ten ECG diagnostic tasks.

Our primary contributions are summarized:

- **A geometric failure mode.** We connect phase–morphology ambiguity on phase-like trajectories to radial-to-tangential gradient leakage under STE.
- **A coupled remedy.** GAQ combines Mahalanobis metric shaping with Householder gradient transport. It cancels the leakage in the idealized case and yields a first-order, one-step phase-drift bound.
- **Evidence across phase-like signals.** AtomECG supports learned tokenization, autoregressive modeling, and linguistic-style probes on ECG. Controlled ECG experiments, a known-phase limit cycle, and PPG evaluate the mechanism and its scope.

2 Related Work

eSSL(ECG Self-Supervised Learning). Existing eSSL frameworks fall into contrastive and generative paradigms. Contrastive methods adapt MoCo (He et al., 2020), SwAV (Caron et al., 2020) and patient-level invariances such as CMSC/CMSMLC (Kiyasseh et al., 2021) to ECG, with extensions exploring adversarial augmentation (Wang et al., 2024) and physiological priors (Lan et al., 2022). Generative methods build on Masked Autoencoders: MaeFE (Zhang et al.,

2023) and RLM (Oh et al., 2022) explore temporal/lead masking, while ST-MEM (Na et al., 2024) models lead–time patches with lead embeddings. A separate line incorporates clinical reports (Liu et al., 2024; Wang et al., 2025) or fine-tunes an LLM on ECG tokens (Han et al., 2025). None of these adopts a learned sub-waveform vocabulary.

ECG Language Processing. This paradigm treats cardiac signals as sequences of tokens, adapting NLP techniques to learn interpretable representations. Early attempts, such as ELP (Mousavi et al., 2021), pioneer this discretization but lack self-supervised objectives, limiting the generalization of the learned vocabularies. Subsequent generative frameworks like ECGBERT (Choi et al., 2023) and HeartLang (Jin et al., 2025) advance the approach by segmenting ECG into waveform tokens and adopting BERT-style masked language modeling. However, these methods are constrained by human-defined waveform categories, which limits the model’s ability to discover finer-grained morphological details.

Quantization. Fundamental to discrete representation learning (VQ-VAE (van den Oord et al., 2017)), recent advances prioritize reconstruction fidelity and codebook utilization. SoundStream (Zeghidour et al., 2022) recursively quantizes the waveform’s residual, and HiFi-Codec (Yang et al., 2023) performs grouped quantization. FSQ (Mentzer et al., 2024) simplifies quantization to uniform hypercube points. Most relevant to us, SimVQ (Zhu et al., 2025) learns a linear codebook transformation, and the Rotation Trick (Fifty et al., 2025) replaces STE with a rotation. We ablate against both as the closest baselines in Sec. 4.5.

Algorithm 1 Pre-training of AtomECG.

Require: Input signal $\mathbf{x} \in \mathbb{R}^{c \times l}$.

Require: h separate codebooks $\{\mathcal{C}^{(i)}\}_{i=0}^{h-1}$, where $\mathcal{C}^{(i)} = \{\mathbf{c}_1, \dots, \mathbf{c}_N\}$, $\mathbf{c}_n \in \mathbb{R}^d$.

Require: h linear maps $\{\mathbf{E}^{(i)} \in \mathbb{R}^{d \times d}\}_{i=0}^{h-1}$.

```
1:  $\mathbf{f} = \text{Encoder}(\mathbf{x})$ ,  $\mathbf{f} \in \mathbb{R}^{M \times d}$ ,  $M = c \times p$ 
2:  $\mathbf{r}^{(0)} = \mathbf{f}$ ,  $\hat{\mathbf{f}}_{\text{total}} = \mathbf{0} \in \mathbb{R}^{M \times d}$ 
3: for  $i \leftarrow 0$  to  $h - 1$ 
4:    $\mathbf{Q}^{(i)} \in \mathbb{R}^{M \times d}$ ,  $\hat{\mathbf{f}}^{(i)} \in \mathbb{R}^{M \times d}$ 
5:   for  $j \leftarrow 0$  to  $M - 1$ 
6:      $n_j^* = \underset{n=1, \dots, N}{\operatorname{argmin}} \|\mathbf{r}_j^{(i)} - \mathbf{E}^{(i)} \mathbf{c}_n\|$ 
7:      $\mathbf{Q}_j^{(i)} = \mathbf{c}_{n_j^*}$ ,  $\bar{\mathbf{c}}_j^{(i)} = \mathbf{E}^{(i)} \mathbf{Q}_j^{(i)}$ 
8:      $\mathbf{R}_{\theta, j} = \frac{\|\bar{\mathbf{c}}_j^{(i)}\|}{\|\mathbf{r}_j^{(i)}\|} \left( \mathbf{I} - 2 \frac{\mathbf{w} \mathbf{w}^T}{\mathbf{w}^T \mathbf{w} + \varepsilon} \right)$ ,
        $\{\mathbf{w} = \mathbf{u} - \mathbf{v}, \mathbf{u} = \bar{\mathbf{c}}_j^{(i)} / \|\bar{\mathbf{c}}_j^{(i)}\|, \mathbf{v} = \mathbf{r}_j^{(i)} / \|\mathbf{r}_j^{(i)}\|\}$ 
9:      $\hat{\mathbf{f}}_j^{(i)} = \operatorname{sg}[\bar{\mathbf{c}}_j^{(i)}] + \operatorname{sg}[\mathbf{R}_{\theta, j}] (\mathbf{r}_j^{(i)} - \operatorname{sg}[\mathbf{r}_j^{(i)}])$ 
10:   end for
11:    $\mathbf{r}^{(i+1)} = \mathbf{r}^{(i)} - \mathbf{E}^{(i)} \mathbf{Q}^{(i)}$ 
12:    $\hat{\mathbf{f}}_{\text{total}} = \hat{\mathbf{f}}_{\text{total}} + \hat{\mathbf{f}}^{(i)}$ 
13: end for
14:  $\hat{\mathbf{x}}^{c \times l} = \text{Decoder}(\text{Linear}_{p \times d \rightarrow l}(\hat{\mathbf{f}}_{\text{total}}))$ 
15: Return  $\hat{\mathbf{x}}$ 
```

3 Methods

We first describe the patching and encoder–decoder backbone, then introduce GAQ and its theoretical analysis. We use **atom** for a reusable sub-waveform unit, **code** $\mathbf{c}_n \in \mathbb{R}^d$ for its continuous quantizer embedding, and **token (index)** $\tau(\mathbf{r}) \in \{1, \dots, N\}$ for its discrete identifier in the grammar model.

3.1 Patching and Encoder–Decoder Backbone

Patch and Mask. Given an ECG signal $\mathbf{x} \in \mathbb{R}^{c \times l}$ with patch length b and stride s , the number of resulting patches per lead is $p = \lfloor \frac{l-b}{s} \rfloor + 1$. Under masked atom modeling, we randomly mask 40% of the $c \times p$ patches.

Encoder and Decoder. AtomECG adopts an asymmetric encoder–decoder architecture to disentangle high-level semantic abstraction from low-level signal reconstruction. The Encoder uses Multi-Head Attention (Vaswani et al., 2017) to project ECG signals into parallel subspaces, fusing local morphology with global rhythmic context. Following Lopez Alcaraz and Strodthoff (2023), which validates S4 (Gu et al., 2022b) for ECG fore-

casting, we adopt S4D (Gu et al., 2022a) as the decoder backbone (see App. A.3).

Representation extraction and learning. We use patch length $b=16$, stride $s=8$, a six-layer Transformer encoder with hidden size $d=128$, and four residual GAQ levels. Each patch yields one atom index at each level, and the decoder and downstream classifiers receive the sum of the four assigned codes. Training has two stages. Stage 1 masks 40% of the patches and pre-trains the tokenizer with

$$\mathcal{L}_{\text{pre}} = \text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{L}_{\text{vq}} + \mathcal{L}_{\text{commit}}, \quad (1)$$

where the VQ and commitment weights are $\alpha=\beta=0.25$. Stage 2 fine-tunes the tokenizer on the target training split under the same masked objective, freezes it, and trains a six-layer GPT on level-0 indices with next-token cross-entropy and context length 256. No held-out fold enters either stage. The loss definitions and complete training protocol are provided in App. A.1.1 and A.6.

3.2 Geometry Alignment Quantization (GAQ)

We require atoms to satisfy two linguistic-style conditions: *phase-coherent* (same sub-waveform across beats $\rightarrow$ same atom, like a stable subword) and *context-precise* (similar shapes at different phases $\rightarrow$ different atoms, like homographs disambiguated by position). We show below that both can fail under VQ+STE through one geometric mechanism. SimVQ (Zhu et al., 2025) changes assignment geometry and retains STE. The Rotation Trick (Fifty et al., 2025) changes gradient transport and retains the Euclidean assignment metric. These methods isolate the two ingredients that GAQ couples around radial-to-tangential leakage. GAQ couples both ingredients with this leakage as the explicit target (Algo. 1): **(C1)** global metric shaping via a learnable linear map $\mathbf{E}^{(i)}$, and **(C2)** local geometry-aligned gradient transport via a Householder map $\mathbf{R}_{\theta, j}$ replacing STE. Full proofs and implementation notes are in App. A.1.

Composition of GAQ. GAQ has three stages at each residual level. Assignment searches the $\mathbf{E}$ -transformed codebook, which changes the global assignment geometry. The forward pass supplies exactly the selected transformed code to the decoder through the stop-gradient surrogate in Algo. 2. The backward pass replaces the identity Jacobian of STE with the per-sample map $\mathbf{R}_{\theta}$. This separation makes $\mathbf{E}$ responsible for assignment

Algorithm 2 GAQ at one residual level.

Require: Residual $\mathbf{r}$, codebook $\{\mathbf{c}_n\}_{n=1}^N$, map $\mathbf{E}$.

```

1:  $n^* \leftarrow \arg \min_n \|\mathbf{r} - \mathbf{E}\mathbf{c}_n\|_2, \bar{\mathbf{c}} \leftarrow \mathbf{E}\mathbf{c}_{n^*}$ 
2:  $\mathbf{u} \leftarrow \bar{\mathbf{c}}/\|\bar{\mathbf{c}}\|_2, \mathbf{v} \leftarrow \mathbf{r}/\|\mathbf{r}\|_2, \mathbf{w} \leftarrow \mathbf{u} - \mathbf{v}$ 
3:  $\mathbf{H}_\varepsilon \leftarrow \mathbf{I} - 2\mathbf{w}\mathbf{w}^T/(\mathbf{w}^T\mathbf{w} + \varepsilon)$ 
4:  $\mathbf{R}_\theta \leftarrow (\|\bar{\mathbf{c}}\|_2/\|\mathbf{r}\|_2)\mathbf{H}_\varepsilon$ 
5:  $\tilde{\mathbf{q}} \leftarrow \text{sg}[\bar{\mathbf{c}}] + \text{sg}[\mathbf{R}_\theta](\mathbf{r} - \text{sg}[\mathbf{r}])$ 
6: return  $n^*, \tilde{\mathbf{q}}$ 

```

boundaries and $\mathbf{R}_\theta$ responsible for encoder-side gradient transport.

Setup (one residual level). We omit the residual superscript (i) when the context is clear. Let $\mathbf{r} \in \mathbb{R}^d$ be a residual feature for a patch (e.g. $\mathbf{r}_j^{(i)}$ in Algo. 1), and let $\mathcal{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_N\} \subset \mathbb{R}^d$ be the codebook. GAQ performs hard assignment in a transformed code space:

$$\tau(\mathbf{r}) \in \arg \min_{n=1, \dots, N} \|\mathbf{r} - \mathbf{E}\mathbf{c}_n\|_2, \bar{\mathbf{c}}(\mathbf{r}) := \mathbf{E}\mathbf{c}_{\tau(\mathbf{r})}. \quad (2)$$

Define the pointwise quantization error

$$\mathbf{e}(\mathbf{r}) := \mathbf{r} - \bar{\mathbf{c}}(\mathbf{r}). \quad (3)$$

With the usual VQ and commitment losses, we have

$$\mathcal{L}_{\text{vq}} + \mathcal{L}_{\text{commit}} = (\alpha + \beta) \|\mathbf{e}(\mathbf{r})\|_2^2. \quad (4)$$

Thus, stabilizing the assignments in (2) and reducing $\mathbb{E}\|\mathbf{e}\|_2^2$ directly improves the quality of the discrete atom space.

(C1) Global metric shaping via $\mathbf{E}$. Learning $\mathbf{E}$ reshapes the Voronoi tessellation used by hard assignment (2). When features are anisotropic, plain Euclidean nearest-neighbor quantization over-emphasizes high-variance directions, producing phase-inconsistent assignments. A learnable $\mathbf{E}$ adapts the assignment metric to the feature distribution and, being shared across all atoms in a mini-batch, receives dense updates from batch statistics even when individual codes are updated sparsely.

Proposition 1 (Metric-shaping view of $\mathbf{E}$). *Assume $\mathbf{E}$ is invertible and define $\mathbf{G} := \mathbf{E}^T\mathbf{E} \succ 0$ and $\mathbf{z} := \mathbf{E}^{-1}\mathbf{r}$. Then (2) is equivalent to*

$$\tau(\mathbf{r}) \in \arg \min_n \|\mathbf{z} - \mathbf{c}_n\|_{\mathbf{G}}^2, \quad \|\mathbf{x}\|_{\mathbf{G}}^2 := \mathbf{x}^T\mathbf{G}\mathbf{x}.$$

Hence learning $\mathbf{E}$ is equivalent to learning a shared Mahalanobis geometry that globally reshapes the assignment boundaries.

Scale control. Because $(\mathbf{E}, \mathbf{c}_n) \mapsto (s\mathbf{E}, \mathbf{c}_n/s)$ leaves $\mathbf{E}\mathbf{c}_n$ unchanged, we fix the overall scale by $\mathbf{E} \leftarrow \frac{\sqrt{d}}{\|\mathbf{E}\|_F} \mathbf{E}$, applied independently at each residual level.

(C2) Local geometry-aligned gradient transport via $\mathbf{R}_\theta$. Hard assignment (2) is discontinuous. STE substitutes the Jacobian with identity, ignoring the geometry between $\mathbf{r}$ and its assigned code $\bar{\mathbf{c}}$. GAQ instead uses a per-sample *orthogonal alignment map* (a Householder reflection) to transport gradients in a geometry-consistent frame. Let $\mathbf{v} := \mathbf{r}/\|\mathbf{r}\|_2, \mathbf{u} := \bar{\mathbf{c}}/\|\bar{\mathbf{c}}\|_2$ and $\mathbf{w} := \mathbf{u} - \mathbf{v}$. Define the Householder map

$$\mathbf{H}_\varepsilon(\mathbf{r}, \bar{\mathbf{c}}) := \mathbf{I} - 2\frac{\mathbf{w}\mathbf{w}^T}{\mathbf{w}^T\mathbf{w} + \varepsilon},$$

and the scaled alignment used in Algo. 1:

$$\mathbf{R}_\theta(\mathbf{r}, \bar{\mathbf{c}}) := \frac{\|\bar{\mathbf{c}}\|_2}{\|\mathbf{r}\|_2} \mathbf{H}_\varepsilon(\mathbf{r}, \bar{\mathbf{c}}). \quad (5)$$

GAQ uses the exact-forward surrogate in Algo. 2. Its value is $\tilde{\mathbf{q}} = \bar{\mathbf{c}}$, while its Jacobian with respect to $\mathbf{r}$ is $\mathbf{R}_\theta$. Encoder gradients therefore become $\mathbf{R}_\theta^T \nabla_{\tilde{\mathbf{q}}} \mathcal{L}$. In the idealized $\varepsilon=0$ case, the alignment itself also maps $\mathbf{r}$ exactly to $\bar{\mathbf{c}}$ (Lemma 3).

An NLP intuition. Tangential motion on the normalized feature sphere can break phase coherence by sliding a feature along the phase trajectory. Under STE, even a purely radial code-side gradient can induce this motion at the encoder feature. GAQ aligns the gradient frames and removes this leakage in the idealized case.

Proposition 2 (STE tangential leakage and GAQ cancellation). *Let $\mathbf{g} := \nabla_{\tilde{\mathbf{q}}} \mathcal{L}$ be the upstream gradient at the quantized output. Decompose $\mathbf{g}$ into a tangential part at $\mathbf{u}$ and a radial part:*

$$\mathbf{g} = \mathbf{g}_T + a\mathbf{u}, \quad \mathbf{g}_T \perp \mathbf{u}, \quad a = \mathbf{u}^T \mathbf{g}.$$

Let $\Pi_{\mathbf{v}} := \mathbf{I} - \mathbf{v}\mathbf{v}^T$ be the tangent projection at $\mathbf{v}$ and let $\theta := \arccos(\mathbf{u}^T \mathbf{v})$. Under STE, the radial component $a\mathbf{u}$ generally contributes a tangential term of magnitude $|a| \sin \theta$ in $\Pi_{\mathbf{v}} \mathbf{g}$ whenever $\theta \neq 0$. In contrast, in the ideal case $\varepsilon = 0$, GAQ transports gradients by the orthogonal part of $\mathbf{R}_\theta$, mapping $a\mathbf{u}$ to a purely radial direction at $\mathbf{v}$ and hence

$$\Pi_{\mathbf{v}}(\mathbf{H}^T(a\mathbf{u})) = \mathbf{0}, \quad \|\Pi_{\mathbf{v}} \mathbf{H}^T \mathbf{g}_T\|_2 = \|\mathbf{g}_T\|_2.$$

Thus GAQ removes the radial-to-tangential leakage channel in the idealized $\varepsilon=0$ case. Lemma 4 bounds the stabilized deviation used in practice.

A one-step phase drift bound. To connect the cancellation to phase coherence, we use a minimal local chart assumption. The chart is an analytical device that the algorithm never observes, and the result concerns one first-order update.

Assumption 1 (Local phase chart on the feature sphere). *There exists a C^1 curve $\gamma : [0, 1) \rightarrow \mathbb{S}^{d-1}$ describing the phase trajectory of normalized features and a measurable phase readout ψ defined on a neighborhood $\mathcal{N}$ of $\gamma([0, 1))$ such that $\psi(\gamma(\varphi)) = \varphi$ and ψ is L_ψ -Lipschitz on $\mathcal{N}$.*

Theorem 1 (One-step phase drift bound). *Under Assumption 1, let $\mathbf{v} = \mathbf{r}/\|\mathbf{r}\|_2 \in \mathcal{N}$ and consider one SGD update of the encoder feature $\mathbf{r}$ with step size η . Then the induced phase drift satisfies the first-order bound*

$$|\psi(\mathbf{v}^+) - \psi(\mathbf{v})| \leq \frac{L_\psi \eta}{\|\mathbf{r}\|_2} \|\Pi_{\mathbf{v}} \Delta \mathbf{r}\|_2 + O(\eta^2),$$

where $\Delta \mathbf{r}$ is the backpropagated gradient step at $\mathbf{r}$. For STE, $\Delta \mathbf{r} = \mathbf{g}$, so the bound includes an additional leakage term proportional to $|a| \sin \theta$ from Prop. 2. For GAQ with $\varepsilon = 0$, $\Delta \mathbf{r} = \mathbf{R}_\theta^T \mathbf{g}$ and the radial component $a\mathbf{u}$ contributes no tangential motion, yielding the tighter proxy

$$\begin{aligned} |\psi(\mathbf{v}^+) - \psi(\mathbf{v})| &\lesssim \frac{L_\psi \eta}{\|\mathbf{r}\|_2} \|\mathbf{g}_T\|_2 \quad (\text{GAQ}) \text{ vs.} \\ &\lesssim \frac{L_\psi \eta}{\|\mathbf{r}\|_2} (\|\mathbf{g}_T\|_2 + |a| \sin \theta) \quad (\text{STE}). \end{aligned}$$

Measurable phase-coherence criteria. Assumption 1 motivates evaluating phase coherence directly. For a relative phase variable φ and the hard index map $\tau(\cdot)$, we define vector-level and token-level phase coherence:

$$\text{PC} := \mathbb{E}_\varphi \mathbb{E}_{\mathbf{r}, \mathbf{r}' \sim \text{i.i.d. } P_\varphi} \left[\|\bar{\mathbf{c}}(\mathbf{r}) - \bar{\mathbf{c}}(\mathbf{r}')\|_2^2 \right], \quad (6)$$

$$\text{TS} := \mathbb{E}_\varphi \mathbb{P}_{\mathbf{r}, \mathbf{r}' \sim \text{i.i.d. } P_\varphi} (\tau(\mathbf{r}) \neq \tau(\mathbf{r}')). \quad (7)$$

If the transformed codebook has minimum inter-code distance $m := \min_{n \neq n'} \|\mathbf{E} \mathbf{c}_n - \mathbf{E} \mathbf{c}_{n'}\|_2$, then token instability is controlled by vector instability:

Lemma 1 (Token disagreement is controlled by PC). $\text{TS} \leq \text{PC}/m^2$.

Where GAQ helps downstream tasks. Lower quantization error $\mathbb{E}\|\mathbf{e}\|_2^2$ reduces feature-to-code distortion and preserves diagnostic separability. Lower TS makes phase-localized atoms statistically stable across cycles, enabling motif enrichment. Supporting results and implementation limits

are provided in App. A.1.8–A.1.10. Component C1 reduces distortion, and C2 reduces phase-induced assignment noise. GAQ itself never observes phase. The map $\mathbf{H}_\varepsilon$ depends only on $(\mathbf{r}, \bar{\mathbf{c}})$, and Assumption 1 is used only to interpret stability.

4 Experiments

4.1 Datasets and Experimental Settings

We use four datasets from PhysioNet (Goldberger et al., 2000; Pollard et al., 2026) and VitalDB (Lee et al., 2022). All experiments were conducted on a NVIDIA H200 GPU equipped with 141GB memory. Detailed preprocessing and implementation settings are provided in App. A.4 and A.6.

MIMIC-IV-ECG (Gow et al., 2023). The standard MIMIC-IV-ECG dataset contains 800,035 12-lead ECG from 161,352 distinct patients, sampled at 500 Hz for 10 seconds.

PTB-XL (Wagner et al., 2022). The standard PTB-XL dataset contains 21,799 12-lead ECG from 18,869 distinct patients, sampled at 500 Hz for 10 seconds. It can be used for two classification tasks. **Multi-label**: six sub-datasets (label counts): all (71), diagnostic (44), sub-diagnostic (24), super-diagnostic (5), rhythm (19), form (12). **Multi-class**: The “all” sub-dataset is employed, with samples reclassified into: normal, target diagnostic or other. Following (Nonaka and Seita, 2021), all results for all tasks are obtained using five-fold cross-validation.

G12EC (Perez Alday et al., 2022). The standard G12EC dataset contains 10,344 12-lead ECG from 10,344 patients, sampled at 500 Hz for 10 seconds. The label count for multi-label is 30. We follow PTB-XL for result aggregation and reporting.

VitalDB (Lee et al., 2022). The data was obtained from surgery patients. From this cohort, we select 2-lead ECG recordings (leads II and V5) from 3,390 cases, with an average duration of about 2 hours and a sampling rate of 500 Hz. For evaluation, the dataset is partitioned using five-fold cross-validation.

4.2 Evaluation on Classification Tasks

We first verify that the learned atoms encode discriminative information by linear-probing downstream diagnosis, a standard check that the discrete vocabulary is informative. We evaluate two settings: multi-label (overall diagnostic coverage) and multi-class (primary abnormality). All results use

Model	all	diag.	sub.	super.	form	rhythm	G12EC
<i>contrastive self-supervised learning</i>							
MoCo v3	0.704±0.01	0.750±0.01	0.763±0.01	0.777±0.01	0.716±0.01	0.734±0.02	0.929±0.005
SwAV	0.712±0.01	0.744±0.01	0.753±0.01	0.754±0.01	0.730±0.01	0.729±0.02	0.931±0.007
CMSC	0.684±0.01	0.732±0.01	0.742±0.01	0.760±0.01	0.723±0.01	0.721±0.02	0.930±0.006
CMSMLC	0.679±0.01	0.721±0.01	0.751±0.01	0.760±0.01	0.727±0.01	0.758±0.03	0.931±0.008
<i>generative self-supervised learning</i>							
MTAE	0.599±0.02	0.705±0.01	0.623±0.01	0.652±0.02	0.603±0.01	0.671±0.02	0.849±0.009
MTAE+RLM	0.536±0.02	0.688±0.02	0.621±0.02	0.614±0.02	0.472±0.01	0.579±0.03	0.805±0.011
ST-MEM	0.629±0.01	0.702±0.01	0.643±0.01	0.704±0.01	0.658±0.01	0.756±0.02	0.886±0.004
HeartLang	0.796±0.01	0.821±0.01	0.861±0.01	0.853±0.01	0.786±0.01	0.884±0.02	0.926±0.005
A(SimVQ) [†]	0.803±0.01	0.816±0.01	0.858±0.01	0.882±0.01	0.800±0.01	0.876±0.02	0.945±0.003
AtomECG	0.812±0.01	0.825±0.01	0.869±0.01	0.886±0.01	0.807±0.01	0.882±0.02	0.950±0.002

Table 1: Macro AUROC of AtomECG and eight eSSL models via linear probing across seven multi-label tasks on PTB-XL and G12EC. [†] A(SimVQ) replaces GAQ in AtomECG with SimVQ (Zhu et al., 2025), the closest quantization baseline that also learns a linear transformation of the codebook.

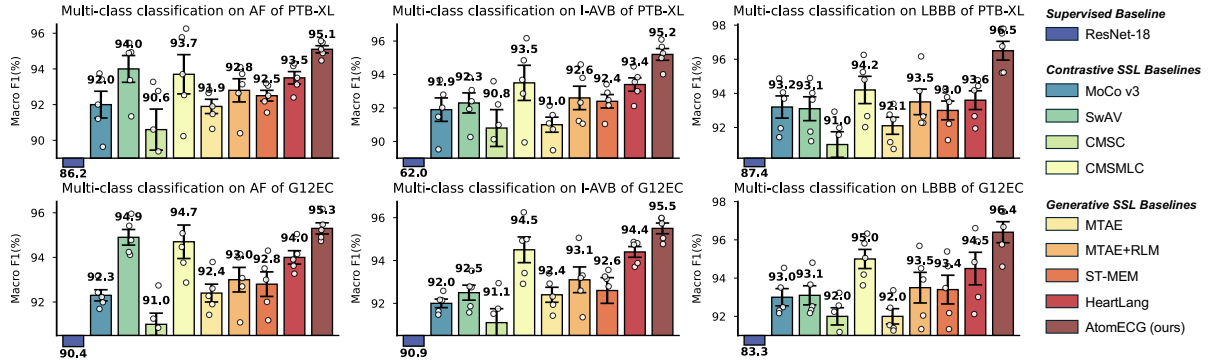


Figure 2: Macro F1 of AtomECG and nine other models under MLP probing across three multi-class tasks on PTB-XL and G12EC.

MIMIC-IV-ECG pre-training. The classifier input is the quantized atom sequence.

Multi-label. As summarized in Table 1, AtomECG sets new SOTA on six of seven subsets via linear probing, without any human-defined augmentation or segmentation. The exception is PTB-XL rhythm, where HeartLang slightly outperforms AtomECG (0.884 vs 0.882), likely benefiting from segmentation-induced inductive bias tailored to rhythm regularities. Replacing GAQ with SimVQ (the strongest concurrent baseline with a linear codebook map) inside the same framework consistently drops Macro AUROC on every subset (e.g. “all” 0.812→0.803, G12EC 0.950→0.945), isolating GAQ’s gain beyond a linear codebook map. Across the seven subsets, a one-sided sign test gives $p=0.0078$. Paired Wilcoxon tests over 15 matched runs are significant on six subsets ($p<10^{-3}$) and borderline on super-diagnostic ($p=0.054$). App. A.7.1 reports confidence intervals and per-subset values.

Multi-class. We further evaluate AF, I-AVB and

LBBB (arrhythmia and conduction abnormalities)¹. Under a lightweight MLP probing (Linear-ReLU-Linear), AtomECG attains the highest Macro F1 on both datasets (Fig. 2), indicating that the discrete atoms encode diagnostically useful morphology and context exploitable by simple classifiers.

4.3 Linguistic Probe #1: Motif-Meaning Correspondence

We adapt the standard χ^2 enrichment test from corpus linguistics. For each cohort label, we test the occurrence of each atom and candidate subsequence, then re-test discovered motifs on an independent validation split (App. A.5). Table 2 summarizes motifs with $p<0.05$ on both splits over 6 major and 12 minor diagnostic classes on PTB-XL.

The learned codebook captures class-enriched motifs. Each major category is associated with a distinct atom set (e.g. atoms {347, 494, 813} across conduction disturbance *CD* subclasses), and many

¹We provide the complete descriptions of all cardiac disease abbreviations used in this manuscript in App. A.2.

	Subclass	Num	Atom usage(ratio)
NORM	NORM	9528	1,24,126,170,269,351,395, 542,696,722,785,822,867, 965,1010. 24-696-722-170(86%)...
CD	RBBB	1660	77,79,346,347,494,497,813. 346-347-77-494(73%)...
	LBBB	613	80,150,346,347,357,494,813. 346-357-80-813-494(72%)...
	AVB	827	2,111,333,347,494,813,815. 111-2-813-494(81%)...
HYP	LVH	2137	52,80,150,357,535,778. 778-357-535(74%)...
	SEHYP	30	22,52,346,357,535,813. 346-357-535(80%)...
MI	AMI	354	113,334,425,436,479,1021. 436-113-1021(76%)...
	IMI	2685	52,81,113,279,425,743,905. 113-905-425(84%)...
STTC	ISCA	1016	2,112,113,456,921. 113-2-456-921(70%)...
	STTC	2329	52,112,113,333,660,743,921, 980,999. 113-743-52(79%)...
Rhythm	AFIB	1514	67,74,162,293,343,586,712, 798,980. 798-712-67-586(95%)...
	STACH	826	6,28,67,333,586,712,798,980. 798-712-67-28(96%)...

Table 2: Atoms and subsequences associated with PTB-XL diagnostic classes (χ^2 /Fisher test, $p < 0.05$ on discovery and validation splits). The analysis is exploratory and the clinical interpretations require expert validation.

subclasses admit compact subsequences with high coverage (*AFIB*: 798–712–67–586 at 95%), giving concise discrete features ready for inspection.

Shared atoms across related cohorts suggest hypotheses for clinical review. Atom 357 is enriched in both *CD-LBBB* and *HYP*, a pattern compatible with the co-occurrence of structural remodeling and conduction abnormalities. Atom 113 is enriched in both *MI* and *STTC*, which may reflect their shared ischemic context. The enrichment establishes cohort association only. It does not validate these pathophysiological interpretations.

4.4 Linguistic Probe #2: Autoregressive Learnability and Long-Context Scaling

We treat next-atom prediction as the ECG analogue of next-token language modeling. We tokenize each ECG, train a GPT to predict the next index, and decode predicted indices to waveforms.

Model	MAE	RMSE	MAE	RMSE
<i>Signal prediction</i>				
	250-length		500-length	
TimeGPT	0.168	0.179	0.274	0.299
Autoformer	0.161	0.177	0.256	0.257
<i>Index prediction</i>				
Random+GPT	0.814	0.875	0.856	0.894
HeartLang+GPT	0.138	0.149	0.280	0.292
AtomECG+GPT	0.022	0.025	0.054	0.061
Model	MAE	RMSE	MAE	RMSE
<i>Signal prediction</i>				
	1000-length		2000-length	
TimeGPT	0.443	0.473	0.605	0.625
Autoformer	0.415	0.434	0.594	0.642
<i>Index prediction</i>				
Random+GPT	0.903	0.926	0.905	0.926
HeartLang+GPT	0.482	0.525	0.617	0.655
AtomECG+GPT	0.062	0.069	0.122	0.143

Table 3: Performance of next-atom prediction on PTB-XL (250/500-signal-length) and VitalDB (1000/2000-signal-length). MAE/RMSE are signal-domain metrics.

Method	FT	Tok.	Seq.	GPU-h
TimeGPT	yes	—	N/A	—
Autoformer	yes	—	12.3 M	5.8
Random+GPT	random	0.52 M	20.1 M	9.2
HeartLang+GPT	yes	2.10 M	20.1 M	34.0
AtomECG+GPT	yes	1.83 M	20.1 M	38.5

Table 4: Adaptation and compute for Table 3. Tok./Seq. denote tokenizer/sequence-model parameters. TimeGPT has closed weights. GPU-hours use one H200.

We report MAE/RMSE on PTB-XL (250/500 samples) and VitalDB (1000/2000 samples), comparing direct signal forecasters (TimeGPT (Garza et al., 2023), Autoformer (Wu et al., 2021)) and index forecasters with alternative tokenizers (Random+GPT, HeartLang+GPT). Every trainable method is adapted on the target training split (Table 4). The three index forecasters use the same six-layer GPT and the same S4D decoder architecture and training recipe. All methods share data splits and sample counts, and each direct baseline uses its recommended hyperparameters. The held-out fold never enters tokenizer adaptation.

As shown in Table 3, AtomECG+GPT yields the lowest MAE/RMSE across both datasets and all horizons. Random+GPT performs poorly, ruling out a pure GPT prior. HeartLang+GPT improves over random discretization yet remains consistently worse. The result suggests that AtomECG’s sub-waveform atoms induce a more regular sequence structure than coarser human-defined units. The advantage persists as the horizon grows, indicating

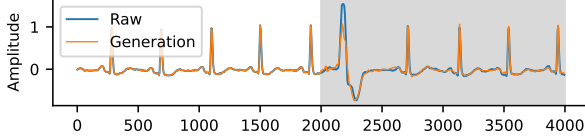


Figure 3: Visualization of next-atom prediction with 2000-signal-length of lead II on VitalDB.

Few-shot	Pre-train on MIMIC		Pre-train on VitalDB	
	MAE	RMSE	MAE	RMSE
$\times$	0.086	0.089	0.099	0.108
1%	0.075	0.076	0.067	0.072
10%	0.056	0.066	0.058	0.064
full	0.049	0.057	0.053	0.058

Table 5: Cross-dataset transfer of next-atom prediction on PTB-XL (Lead II, 500-signal-length).

that atom sequences preserve longer-range dependencies in an autoregressively tractable form. Fig. 3 shows decoded signals that closely follow the raw morphology.

Cross-Dataset Generalization. We pre-train on MIMIC-IV-ECG or VitalDB and evaluate next-atom prediction on PTB-XL lead II (500-signal-length) under zero/few-shot adaptation. Both sources transfer zero-shot and improve monotonically with target data (Table 5), suggesting shared morphological primitives across the evaluated cohorts.

Long-context perplexity scaling. Combined with the motif probe in Sec. 4.3, the next-atom results provide two NLP-native tests of the atom-grammar view. Token-level perplexity drops monotonically from 488 under a unigram model to 24 under a long-context GPT (App. Table 25). Cohort-aligned χ^2 motifs provide a complementary vocabulary-level probe.

4.5 Diagnostics of GAQ

Gradient geometry and phase coherence. Sec. 3.2 predicts that GAQ improves phase-coherent tokenization by suppressing radial-to-tangential gradient leakage. We test this with four diagnostics: LEAK₉₀ (tail leakage severity), TANFRAC (tangential update fraction), TS (within-phase token switch rate across beats), and ORTH₉₀ (deviation from orthogonality of the alignment map, defined in App. A.4). Table 6 shows that GAQ reduces LEAK₉₀ and TANFRAC, consistent with Prop. 2 and Thm. 1. The token switch rate decreases from 0.18 to 0.04, and ORTH₉₀ remains near zero.

Set	Leak ₉₀	TanFrac	TS	Orth ₉₀
STE	0.64	0.51±0.05	0.18±0.002	–
EMAVQ	0.50	0.49±0.04	0.16±0.002	–
SimVQ	0.37	0.48±0.05	0.10±0.003	–
RotVQ	0.62	0.34±0.05	0.12±0.002	–
GAQ	0.28	0.22±0.04	0.04±0.001	0.001

Table 6: Gradient-geometry and phase-coherence diagnostics on MIMIC-IV-ECG. Lower is better for all metrics. RotVQ denotes the rotation trick. Residual VQ denotes multi-level quantization.

Set	CPS $\uparrow$	WPA $\uparrow$	D $\uparrow$	cNMI $\uparrow$
STE	0.67	0.68	0.35	0.06
VQ+E	0.74	0.79	0.53	0.14
VQ+R _{θ}	0.75	0.80	0.55	0.16
GAQ	0.91	0.96	0.87	0.43

Table 7: Morphology-controlled cross-phase disambiguation on MIMIC-IV-ECG. Among morphology-matched patches ($\text{NCC} \geq 0.95$), CPS / WPA measure cross-phase separation / within-phase agreement, $D = \text{CPS} + \text{WPA} - 1$, and cNMI quantifies morphology-conditioned phase dependence. Higher is better.

Cross-phase disambiguation. The diagnostics above show GAQ stabilizes assignments *within* the same phase. The other half of phase-morphology ambiguity requires that morphology-similar patches at *different* phases be *separated* into different atoms. After R-peak detection, each patch gets a within-beat phase $\varphi \in [0, 1)$ discretized into $B=8$ bins. Morphology-matched pairs (z-scored patches with normalized cross-correlation $\text{NCC} \geq 0.95$) are evaluated by cross-phase separation $\text{CPS} = \mathbb{P}(\tau(\mathbf{r}) \neq \tau(\mathbf{r}') \mid \text{match}, \varphi_{\text{bin}}(\mathbf{r}) \neq \varphi_{\text{bin}}(\mathbf{r}'))$ and within-phase agreement $\text{WPA} = \mathbb{P}(\tau(\mathbf{r}) = \tau(\mathbf{r}') \mid \text{match}, \varphi_{\text{bin}}(\mathbf{r}) = \varphi_{\text{bin}}(\mathbf{r}'))$. We combine them into $D = \text{CPS} + \text{WPA} - 1$ and also report morphology-conditioned cNMI between token identity and phase. Table 7 shows GAQ dominates STE and either-component-only ablations on all four metrics (D : 0.35→0.87, cNMI: 0.06→0.43), directly substantiating the context-precise side of the ambiguity.

4.6 Clinical Robustness

We examine three axes on PTB-XL multi-label classification: demographic robustness (4 subgroups stratified by sex and the median age, with 15 paired runs from 5-fold CV $\times$ 3 seeds), calibration via the cross-validated worst-group Macro ECE (CWM), and distortion robustness under lead dropout and additive Gaussian noise (20 dB SNR). On the “all”

Quantizer	Extra	GPU-h	PTB	G12	Leak ₉₀	TS
VQ, $N=1024$	0	27.7	0.767	0.924	0.64	0.18
VQ, $N=2048$	0.524 M	28.8	0.772	0.925	0.63	0.16
VQ + adapter	0.066 M	29.7	0.781	0.929	0.55	0.14
RotVQ	0	28.4	0.793	0.937	0.62	0.12
SimVQ	0.066 M	29.8	0.803	0.945	0.37	0.10
Sim+Rot	0.066 M	30.2	0.807	0.947	0.33	0.07
GAQ	0.066 M	30.0	0.812	0.950	0.28	0.04

Table 8: Controlled comparison with encoder, decoder, data, epochs, batch size, seeds, and four residual levels fixed. Extra parameters are relative to VQ with $N=1024$.

Setting	PTB all	super.	G12EC	TS
HeartLang	0.796	0.853	0.926	0.11
HeartLang+GAQ	0.806	0.861	0.933	0.06
ST-MEM	0.629	0.704	0.886	—
ST-MEM+GAQ	0.659	0.726	0.901	0.06

Table 9: Macro AUROC under linear probing and token switch rate (TS) after inserting GAQ into other backbones.

subset, AtomECG attains mean worst-group AUROC = 0.804 (95% CI [0.801, 0.808]), best–worst gap 0.014, CWM = 0.028 ($\ll 0.05$). Performance drops by ≤ 0.005 AUROC under single-lead removal or AWGN and by ≤ 0.05 AUROC under two-lead removal, consistently across all subgroups. Full subgroup tables and additional tasks are in App. A.7.4.

4.7 Controlled Quantizer Ablation

Table 8 separates the geometric mechanism from capacity. Doubling the codebook adds eight times GAQ’s parameter budget and gains 0.005 AUROC on PTB-XL. A parameter-matched encoder adapter reaches 0.781, while the parameter-free RotVQ reaches 0.793. SimVQ+RotVQ also trails GAQ on both datasets and both geometry diagnostics. GAQ adds four 128×128 matrices, or 65,536 parameters, while $\mathbf{R}_\theta$ has no parameters and costs $O(d)$ per patch. Component, patching, codebook, and grammar ablations remain in App. A.8.

4.8 Cross-Backbone and Cross-Signal Evaluation

To test whether the gain depends on AtomECG’s encoder, we replace HeartLang’s VQ layer with GAQ while preserving its encoder, segmentation, and training recipe. HeartLang+GAQ improves all three classification subsets and nearly halves TS (Table 9), yet remains below AtomECG on PTB-XL all (0.806 vs. 0.812), suggesting that uniform

sub-cycle patching contributes beyond the quantizer. ST-MEM has no native quantizer. Attaching the same GAQ quantizer and decoder to its encoder also yields consistent gains, including +0.030 AUROC on PTB-XL all.

We further isolate signal scope with two non-ECG evaluations (full tables in App. A.7.3). On a limit cycle with known phase and repeated local shapes, GAQ reduces one-step phase drift from 4.9×10^{-2} to 1.1×10^{-2} and TS from 0.18 to 0.05. On VitalDB PPG, it reduces reconstruction MSE from 0.0031 to 0.0016, TS from 0.15 to 0.05, and next-token MAE from 0.041 to 0.024. Together with Table 9, these results separate the geometry-aware quantizer from one particular encoder and support the mechanism on phase-like signals beyond ECG.

5 Conclusion

We introduced GAQ, which addresses phase-morphology ambiguity through coupled Mahalanobis metric shaping and Householder gradient transport. The transport cancels radial-to-tangential leakage in the idealized case, with a controlled stabilization error in practice. AtomECG attains strong diagnostic performance, exhibits long-context perplexity scaling, and supports exploratory motif enrichment over a learned sub-waveform vocabulary. Controlled comparisons isolate the gain from model capacity, while the known-phase limit cycle and PPG experiments provide evidence beyond ECG. These results establish an end-to-end learned tokenization, autoregressive modeling, and linguistic-probing pipeline for physiological signals without treating the motif interpretations as clinical validation.

Limitations

The evidence has four boundaries. First, ECG is the principal testbed. The synthetic limit cycle isolates the proposed mechanism and PPG supplies a second physiological signal, but these experiments do not establish performance on arbitrary sequential modalities.

Second, transfer to highly specialized cohorts remains difficult. Atoms and grammar learned from general clinical data perform poorly on an out-of-distribution Norwegian athlete cohort (Singstad, 2022). Full fine-tuning is needed to recover comparable performance.

Third, AtomECG is stable under demographic

shift, single-lead dropout, and mild Gaussian noise (Sec. 4.6), while two-lead loss and intraoperative artifacts cause larger degradation.

Fourth, the motif analysis is exploratory. Discovery and validation split replication guards against unstable findings, but the reported p -values are not corrected across all candidate motifs. The shared-atom links are statistical associations. Their clinical explanations should not be interpreted as validated mechanisms.

Ethical considerations

This work uses only publicly available, de-identified ECG datasets released with appropriate ethical approvals and data use agreements. No additional human data collection was performed.

Acknowledgments

We utilized a LLM to assist in the preparation of this manuscript. Its role was strictly limited to improving grammar, syntax and clarity. The LLM was not used for generating core research ideas or developing the methodology. All scientific contributions, analyses, and conclusions presented in this paper are entirely our own.

References

- Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924.
- Seokmin Choi, Sajad Mousavi, Phillip Si, Haben G Yhdego, Fatemeh Khadem, and Fatemeh Afghah. 2023. [ECGBERT: Understanding hidden language of ECGs with self-supervised representation learning](#). *arXiv preprint arXiv:2306.06340*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. [Taming transformers for high-resolution image synthesis](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883.
- Christopher Fifty, Ronald Junkins, Dennis Duan, Aniketh Iyengar, Jerry Liu, Ehsan Amid, Sebastian Thrun, and Christopher Re. 2025. Restructuring vector quantization with the rotation trick. In *International Conference on Learning Representations*.
- Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. 2023. [TimeGPT-1](#). *arXiv preprint arXiv:2310.03589*.
- Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. [PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals](#). *circulation*, 101(23):e215–e220.
- Ary L. Goldberger, Zachary D. Goldberger, and Alexei Shvilkin. 2023. *Goldberger’s Clinical Electrocardiography: A Simplified Approach*, 10th edition. Elsevier, Philadelphia, PA.
- Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. 2023. [MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. PhysioNet](#). Version 1.0.
- Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. 2022a. [On the parameterization and initialization of diagonal state space models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 35971–35983.
- Albert Gu, Karan Goel, and Christopher Ré. 2022b. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*.
- William Han, Chaojing Duan, Michael Rosenberg, Emerson Liu, and Ding Zhao. 2025. ECG-Byte: A tokenizer for end-to-end generative electrocardiogram language modeling. In *Proceedings of the 10th Machine Learning for Healthcare Conference*, volume 298. PMLR.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. [Momentum contrast for unsupervised visual representation learning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.
- Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. 2025. Reading your heart: Learning ECG words and sentences via pre-training ECG language model. In *International Conference on Learning Representations*.
- Dani Kiyasseh, Tingting Zhu, and David A Clifton. 2021. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 5606–5615. PMLR.

- Xiang Lan, Dianwen Ng, Shenda Hong, and Mengling Feng. 2022. [Intra-inter subject self-supervised learning for multivariate cardiac signals](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4532–4540.
- Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. 2022. [VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients](#). *Scientific Data*, 9(1):279.
- Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2024. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 31949–31963. PMLR.
- Juan Miguel Lopez Alcaraz and Nils Strodthoff. 2023. Diffusion-based time series imputation and forecasting with structured state space models. *Transactions on Machine Learning Research*.
- Temesgen Mehari and Nils Strodthoff. 2022. [Self-supervised representation learning from 12-lead ECG data](#). *Computers in Biology and Medicine*, 141:105114.
- Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2024. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations*.
- Sajad Mousavi, Fatemeh Afghah, Fatemeh Khadem, and U Rajendra Acharya. 2021. [ECG language processing \(ELP\): A new technique to analyze ECG signals](#). *Computer Methods and Programs in Biomedicine*, 202:105959.
- Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. 2024. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International Conference on Learning Representations*.
- Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*.
- Naoki Nonaka and Jun Seita. 2021. In-depth benchmarking of deep neural network architectures for ECG diagnosis. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149, pages 414–439. PMLR.
- Jungwoo Oh, Hyunseung Chung, Joon-myung Kwon, Dong-gyun Hong, and Edward Choi. 2022. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 338–353. PMLR.
- Erick Andres Perez Alday, Annie Gu, Amit Shah, Chengyu Liu, Ashish Sharma, Salman Seyedi, Ali Bahrami Rad, Matthew Reyna, and Gari Clifford. 2022. [Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020](#). *PhysioNet*. Version 1.0.2.
- Tom Pollard, Benjamin E Moody, Li-wei H Lehman, Brian J Gow, Chrystinne Fernandes, Chen Xie, Alastair Johnson, Roger G Mark, and Thomas Heldt. 2026. [PhysioNet as a global platform for biomedical research](#). *Nature Health*, pages 1–4.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.
- Jürg Schlöpfer and Hein J Wellens. 2017. [Computer-interpreted electrocardiograms: benefits and limitations](#). *Journal of the American College of Cardiology*, 70(9):1183–1192.
- Bjørn-Jostein Singstad. 2022. [Norwegian Endurance Athlete ECG Database](#). *PhysioNet*. Version 1.0.0.
- Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.
- Patrick Wagner, Nils Strodthoff, Ralf-Dieter Boussejot, Wojciech Samek, and Tobias Schaeffter. 2022. [PTB-XL, a large publicly available electrocardiography dataset](#). *PhysioNet*. Version 1.0.3.
- Fuying Wang, Jiacheng Xu, and Lequan Yu. 2025. From token to rhythm: A multi-scale approach for ECG-language pretraining. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 65059–65074. PMLR.
- Ning Wang, Panpan Feng, Zhaoyang Ge, Yanjie Zhou, Bing Zhou, and Zongmin Wang. 2024. [Adversarial spatiotemporal contrastive learning for electrocardiogram signals](#). *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):13845–13859.
- Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 22419–22430.
- Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yueshan Zou. 2023. [HiFi-Codec: Group-residual vector quantization for high fidelity audio codec](#). *arXiv preprint arXiv:2305.02765*.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. [SoundStream: An end-to-end neural audio codec](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Huaicheng Zhang, Wenhan Liu, Jiguang Shi, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. 2023. [MaeFE: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning](#). *IEEE Transactions on Instrumentation and Measurement*, 72:1–15.

Yongxin Zhu, Bocheng Li, Yifei Xin, Zhihua Xia, and Linli Xu. 2025. [Addressing representation collapse in vector quantized models with one linear layer](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22968–22977.

A Appendix

A.1 Proofs and additional details for GAQ

A.1.1 Standard VQ-VAE quantization and the straight-through estimator (STE)

We summarize the standard VQ-VAE quantization mechanism and the STE baseline, using the same notation as the main text. Consider one residual level with encoder feature $\mathbf{r} \in \mathbb{R}^d$ and codebook $\mathcal{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_N\}$.

Hard assignment and quantized value. Standard VQ uses nearest-neighbor assignment in Euclidean space:

$$\tau_{\text{VQ}}(\mathbf{r}) \in \arg \min_n \|\mathbf{r} - \mathbf{c}_n\|_2, \quad \mathbf{q} := \mathbf{c}_{\tau_{\text{VQ}}(\mathbf{r})}.$$

In GAQ, the assigned transformed code is $\bar{\mathbf{c}}(\mathbf{r}) = \mathbf{E}\mathbf{c}_{\tau(\mathbf{r})}$ as in (2). In the baseline (no $\mathbf{E}$), one can simply take $\bar{\mathbf{c}} = \mathbf{q}$.

STE surrogate for backpropagation. Because $\tau_{\text{VQ}}(\cdot)$ is discontinuous, the Jacobian of $\mathbf{q}(\mathbf{r})$ is not defined. The standard STE choice is: *use the quantized value in the forward pass, but pretend it equals $\mathbf{r}$ in the backward pass*. A common implementation writes the forward variable as

$$\tilde{\mathbf{q}}_{\text{STE}} := \mathbf{r} + \text{sg}[\mathbf{q} - \mathbf{r}], \quad (8)$$

so that $\tilde{\mathbf{q}}_{\text{STE}} = \mathbf{q}$ in value, while $\nabla_{\mathbf{r}} \tilde{\mathbf{q}}_{\text{STE}} = \mathbf{I}$ because the stop-gradient term contributes zero derivative. Thus for an upstream gradient $\mathbf{g} = \nabla_{\tilde{\mathbf{q}}} \mathcal{L}$, STE yields

$$\nabla_{\mathbf{r}} \mathcal{L} = \mathbf{g}.$$

GAQ keeps the same forward value but modifies the backward map by replacing $\mathbf{I}$ with $\mathbf{R}_{\theta}^T$.

VQ and commitment losses. Given a reconstruction objective $\mathcal{L}_{\text{rec}}$, VQ-VAE typically uses

$$\begin{aligned} \mathcal{L}_{\text{vq}} &:= \alpha \|\text{sg}[\mathbf{r}] - \mathbf{q}\|_2^2, \\ \mathcal{L}_{\text{commit}} &:= \beta \|\mathbf{r} - \text{sg}[\mathbf{q}]\|_2^2. \end{aligned} \quad (9)$$

Stop-gradient changes gradients but not function values. In particular, $\|\text{sg}[\mathbf{r}] - \mathbf{q}\|^2 = \|\mathbf{r} - \mathbf{q}\|^2$ and $\|\mathbf{r} - \text{sg}[\mathbf{q}]\|^2 = \|\mathbf{r} - \mathbf{q}\|^2$, so $\mathcal{L}_{\text{vq}} + \mathcal{L}_{\text{commit}} = (\alpha + \beta)\|\mathbf{r} - \mathbf{q}\|^2$ in value, matching (4) (and similarly for GAQ with $\mathbf{q}$ replaced by $\bar{\mathbf{c}}$).

A.1.2 Metric shaping via $\mathbf{E}$

Proof of Proposition 1. Let $\mathbf{z} = \mathbf{E}^{-1}\mathbf{r}$ and $\mathbf{x}_n = \mathbf{z} - \mathbf{c}_n$. Then

$$\begin{aligned} \|\mathbf{r} - \mathbf{E}\mathbf{c}_n\|_2^2 &= \|\mathbf{E}\mathbf{z} - \mathbf{E}\mathbf{c}_n\|_2^2 = \|\mathbf{E}(\mathbf{z} - \mathbf{c}_n)\|_2^2 \\ &= (\mathbf{z} - \mathbf{c}_n)^T (\mathbf{E}^T \mathbf{E}) (\mathbf{z} - \mathbf{c}_n) \\ &= \|\mathbf{z} - \mathbf{c}_n\|_{\mathbf{G}}^2, \end{aligned}$$

where $\mathbf{G} = \mathbf{E}^T \mathbf{E} \succ 0$. The argmin over n is therefore identical.

Frozen-assignment least-squares view (optimization support). We formalize the “ $\mathbf{E}$ receives shared/dense updates” statement via a frozen-assignment surrogate.

Proposition 3 (Frozen-assignment least-squares target for $\mathbf{E}$). *Fix a minibatch of residual features $\{\mathbf{r}_j\}_{j=1}^M$ and their hard assignments $\tau_j := \tau(\mathbf{r}_j)$. Let $\mathbf{q}_j := \mathbf{c}_{\tau_j}$ and define matrices $\mathbf{R} = [\mathbf{r}_1, \dots, \mathbf{r}_M] \in \mathbb{R}^{d \times M}$ and $\mathbf{Q} = [\mathbf{q}_1, \dots, \mathbf{q}_M] \in \mathbb{R}^{d \times M}$. Ignoring stop-gradient in values, minimizing $\sum_{j=1}^M \|\mathbf{r}_j - \mathbf{E}\mathbf{q}_j\|_2^2 = \|\mathbf{R} - \mathbf{E}\mathbf{Q}\|_F^2$ over $\mathbf{E}$ yields the normal equation*

$$\mathbf{E}\mathbf{Q}\mathbf{Q}^T = \mathbf{R}\mathbf{Q}^T,$$

and if $\mathbf{Q}\mathbf{Q}^T$ is invertible,

$$\mathbf{E}^* = \mathbf{R}\mathbf{Q}^T(\mathbf{Q}\mathbf{Q}^T)^{-1}.$$

Moreover, the gradient is $\nabla_{\mathbf{E}} \|\mathbf{R} - \mathbf{E}\mathbf{Q}\|_F^2 = 2(\mathbf{E}\mathbf{Q}\mathbf{Q}^T - \mathbf{R}\mathbf{Q}^T)$.

Proof. Expand $\|\mathbf{R} - \mathbf{E}\mathbf{Q}\|_F^2 = \text{Tr}((\mathbf{R} - \mathbf{E}\mathbf{Q})(\mathbf{R} - \mathbf{E}\mathbf{Q})^T)$ and differentiate using standard matrix calculus: $\nabla_{\mathbf{E}} \text{Tr}(\mathbf{E}\mathbf{Q}\mathbf{Q}^T \mathbf{E}^T) = 2\mathbf{E}\mathbf{Q}\mathbf{Q}^T$ and $\nabla_{\mathbf{E}} \text{Tr}(-2\mathbf{R}\mathbf{Q}^T \mathbf{E}^T) = -2\mathbf{R}\mathbf{Q}^T$, yielding the stated gradient. Setting it to zero gives the normal equation and (when invertible) the closed form. $\square$

Interpretation. Prop. 3 does not claim training solves the closed form (assignments evolve). Rather, it highlights a concrete mechanism: $\mathbf{Q}\mathbf{Q}^T$ aggregates all selected atoms in the minibatch, so $\mathbf{E}$ receives dense updates even when individual codes are updated sparsely. This is the non-redundant benefit of keeping $\mathbf{E}$ explicit.

A.1.3 Rank of $\mathbf{E}$

Prop. 1 adopts an invertibility view of $\mathbf{E}$. While exact rank is ill-posed under floating-point arithmetic, we monitor the numerical rank of $\mathbf{E} \in \mathbb{R}^{128 \times 128}$ throughout training to ensure that $\mathbf{E}$ does not collapse to a low-rank map.

Let $\sigma_1 \geq \dots \geq \sigma_{128} \geq 0$ be the singular values of $\mathbf{E}$. For a relative threshold τ_{rank} , we define

$$\text{rank}_\tau(\mathbf{E}) := \#\{k : \sigma_k(\mathbf{E}) \geq \tau_{\text{rank}} \cdot \sigma_1(\mathbf{E})\}.$$

We use $\tau_{\text{rank}} = 10^{-6}$. In our runs, $\text{rank}_\tau(\mathbf{E})$ stays near full rank (typically 128, occasionally 127 depending on the threshold), indicating that $\mathbf{E}$ does not degenerate into a low-dimensional projection.

A.1.4 Householder alignment and forward consistency

We restate the ideal (unregularized) Householder reflection. For $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}$ and $\mathbf{w} = \mathbf{u} - \mathbf{v}$,

$$\mathbf{H} := \mathbf{I} - 2 \frac{\mathbf{w}\mathbf{w}^T}{\mathbf{w}^T \mathbf{w}} (\mathbf{u} \neq \mathbf{v}), \quad \mathbf{H} = \mathbf{I} (\mathbf{u} = \mathbf{v}).$$

Although $\mathbf{H}$ is a reflection, it is an orthogonal map that exactly aligns directions: it sends $\mathbf{v}$ to $\mathbf{u}$ and preserves norms, which is what we need for geometry-consistent gradient transport.

Lemma 2 (Basic properties of the Householder map). *If $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{d-1}$, then $\mathbf{H}$ is symmetric and orthogonal, $\mathbf{H}^T = \mathbf{H}$ and $\mathbf{H}^T \mathbf{H} = \mathbf{I}$, and it swaps $\mathbf{u}$ and $\mathbf{v}$: $\mathbf{H}\mathbf{v} = \mathbf{u}$ and $\mathbf{H}\mathbf{u} = \mathbf{v}$.*

Proof. Symmetry is immediate from the definition. To show orthogonality, write $\mathbf{H} = \mathbf{I} - 2\mathbf{p}\mathbf{p}^T$ where $\mathbf{p} = \mathbf{w}/\|\mathbf{w}\|_2$ has unit norm (when $\mathbf{u} \neq \mathbf{v}$). Then $\mathbf{H}^T \mathbf{H} = (\mathbf{I} - 2\mathbf{p}\mathbf{p}^T)^2 = \mathbf{I} - 4\mathbf{p}\mathbf{p}^T + 4\mathbf{p}(\mathbf{p}^T \mathbf{p})\mathbf{p}^T = \mathbf{I}$. For the alignment identity, note $\mathbf{w}^T \mathbf{v} = \mathbf{u}^T \mathbf{v} - 1$ and $\mathbf{w}^T \mathbf{w} = 2(1 - \mathbf{u}^T \mathbf{v})$, hence

$$\begin{aligned} \mathbf{H}\mathbf{v} &= \mathbf{v} - 2\mathbf{w} \frac{\mathbf{w}^T \mathbf{v}}{\mathbf{w}^T \mathbf{w}} \\ &= \mathbf{v} - 2(\mathbf{u} - \mathbf{v}) \frac{\mathbf{u}^T \mathbf{v} - 1}{2(1 - \mathbf{u}^T \mathbf{v})} = \mathbf{u}. \end{aligned}$$

Finally, $\mathbf{H}^2 = \mathbf{I}$ implies $\mathbf{H}\mathbf{u} = \mathbf{v}$. $\square$

Lemma 3 (Forward consistency of GAQ alignment (ideal case)). *Let $\mathbf{r} \neq \mathbf{0}$ and $\bar{\mathbf{c}} \neq \mathbf{0}$ and define $\mathbf{u} = \bar{\mathbf{c}}/\|\bar{\mathbf{c}}\|$ and $\mathbf{v} = \mathbf{r}/\|\mathbf{r}\|$. In the ideal $\varepsilon = 0$ case, with $\mathbf{R}_\theta = \frac{\|\bar{\mathbf{c}}\|}{\|\mathbf{r}\|} \mathbf{H}$, we have $\mathbf{R}_\theta \mathbf{r} = \bar{\mathbf{c}}$.*

Proof. By Lemma 2, $\mathbf{H}\mathbf{v} = \mathbf{u}$. Then $\mathbf{R}_\theta \mathbf{r} = \frac{\|\bar{\mathbf{c}}\|}{\|\mathbf{r}\|} \mathbf{H}(\|\mathbf{r}\| \mathbf{v}) = \|\bar{\mathbf{c}}\| \mathbf{u} = \bar{\mathbf{c}}$. $\square$

On the stabilization $\varepsilon > 0$. We use $\varepsilon > 0$ only for numerical stability when $\mathbf{u} \approx \mathbf{v}$ so that $\mathbf{w}^T \mathbf{w}$ becomes small. Then $\mathbf{H}_\varepsilon$ is no longer exactly orthogonal, but remains a rank-1 perturbation of $\mathbf{I}$. In particular, the deviation from orthogonality can be quantified as follows.

Lemma 4 (Near-orthogonality of the stabilized Householder map). *Let $\mathbf{H}_\varepsilon = \mathbf{I} - 2 \frac{\mathbf{w}\mathbf{w}^T}{\mathbf{w}^T \mathbf{w} + \varepsilon}$ with $\varepsilon > 0$. Then $\mathbf{H}_\varepsilon$ is symmetric and satisfies*

$$\mathbf{H}_\varepsilon^T \mathbf{H}_\varepsilon = \mathbf{I} - \frac{4\varepsilon}{\mathbf{w}^T \mathbf{w} + \varepsilon} \cdot \frac{\mathbf{w}\mathbf{w}^T}{\mathbf{w}^T \mathbf{w} + \varepsilon},$$

and hence

$$\|\mathbf{H}_\varepsilon^T \mathbf{H}_\varepsilon - \mathbf{I}\|_2 \leq \frac{4\varepsilon}{\mathbf{w}^T \mathbf{w} + \varepsilon}.$$

Therefore, all orthogonality-based identities above hold up to an error that vanishes as $\varepsilon \rightarrow 0$, and are tight whenever $\mathbf{w}^T \mathbf{w} \gg \varepsilon$.

In practice, we choose $\varepsilon = 10^{-8}$ and additionally monitor a direct near-orthogonality diagnostic $\|\mathbf{H}_\varepsilon^T \mathbf{H}_\varepsilon - \mathbf{I}\|$ to confirm that the alignment map remains effectively orthogonal on real training patches. We observe no measurable degradation in forward reconstruction.

A.1.5 Tangential leakage cancellation

Proof of Prop. 2. Let $\Pi_{\mathbf{v}} = \mathbf{I} - \mathbf{v}\mathbf{v}^T$.

(1) What STE does. Under STE, $\nabla_{\mathbf{r}} \mathcal{L} = \mathbf{g}$ (App. A.1.1). Decompose $\mathbf{g} = \mathbf{g}_T + a\mathbf{u}$ with $\mathbf{g}_T \perp \mathbf{u}$. Then

$$\Pi_{\mathbf{v}} \mathbf{g} = \Pi_{\mathbf{v}} \mathbf{g}_T + a \Pi_{\mathbf{v}} \mathbf{u}.$$

The second term is the leakage channel: even if $a\mathbf{u}$ is purely radial w.r.t. the code direction $\mathbf{u}$, it is generally not radial at $\mathbf{v}$ when $\mathbf{u} \neq \mathbf{v}$. Indeed,

$$\begin{aligned} \|\Pi_{\mathbf{v}} \mathbf{u}\|_2^2 &= \|\mathbf{u} - (\mathbf{u}^T \mathbf{v}) \mathbf{v}\|_2^2 \\ &= 1 - (\mathbf{u}^T \mathbf{v})^2 = \sin^2 \theta, \end{aligned}$$

so $\|\Pi_{\mathbf{v}}(a\mathbf{u})\|_2 = |a| \sin \theta$. In general, $\Pi_{\mathbf{v}} \mathbf{g}_T$ may partially cancel $\Pi_{\mathbf{v}}(a\mathbf{u})$, but such cancellation is not systematic. Hence we identify $|a| \sin \theta$ as a robust mechanism by which STE injects tangential motion whenever $\theta \neq 0$ and $a \neq 0$.

(2) What GAQ does (ideal case). In GAQ with $\varepsilon = 0$, the orthogonal part is $\mathbf{H}$ satisfying $\mathbf{H}\mathbf{v} = \mathbf{u}$ and $\mathbf{H}\mathbf{u} = \mathbf{v}$ (Lemma 2). The backpropagated gradient at $\mathbf{r}$ is $\mathbf{H}^T \mathbf{g} = \mathbf{H}\mathbf{g}$ (up to the scalar factor in $\mathbf{R}_\theta$, which does not affect tangential directions). Transporting the radial term gives

$$\begin{aligned}\mathbf{H}^T(a\mathbf{u}) &= a\mathbf{H}\mathbf{u} = a\mathbf{v}, \\ \Rightarrow \Pi_{\mathbf{v}}(\mathbf{H}^T(a\mathbf{u})) &= \Pi_{\mathbf{v}}(a\mathbf{v}) = \mathbf{0}.\end{aligned}$$

For the tangential component, $\mathbf{v}^T(\mathbf{H}^T \mathbf{g}_T) = (\mathbf{H}\mathbf{v})^T \mathbf{g}_T = \mathbf{u}^T \mathbf{g}_T = 0$, so $\mathbf{H}^T \mathbf{g}_T$ is tangent at $\mathbf{v}$ and $\|\Pi_{\mathbf{v}} \mathbf{H}^T \mathbf{g}_T\|_2 = \|\mathbf{H}^T \mathbf{g}_T\|_2 = \|\mathbf{g}_T\|_2$ by orthogonality. This proves the stated cancellation and norm preservation. $\square$

A.1.6 Phase drift bound

We use a standard first-order expansion of the normalization map on $\mathbb{R}^d \setminus \{\mathbf{0}\}$.

Lemma 5 (First-order change of a normalized vector). *Let $\mathbf{r} \neq \mathbf{0}$ and $\mathbf{v} = \mathbf{r}/\|\mathbf{r}\|$. For an update $\mathbf{r}^+ = \mathbf{r} - \eta \Delta \mathbf{r}$ with small η ,*

$$\mathbf{v}^+ - \mathbf{v} = -\frac{\eta}{\|\mathbf{r}\|} \Pi_{\mathbf{v}} \Delta \mathbf{r} + O(\eta^2).$$

Proof. Consider $F(\mathbf{r}) = \mathbf{r}/\|\mathbf{r}\|$. A direct differential calculation gives

$$DF_{\mathbf{r}}[\delta] = \frac{1}{\|\mathbf{r}\|} (\mathbf{I} - \mathbf{v}\mathbf{v}^T) \delta = \frac{1}{\|\mathbf{r}\|} \Pi_{\mathbf{v}} \delta.$$

Apply first-order Taylor expansion at $\delta = -\eta \Delta \mathbf{r}$ to obtain the claim. $\square$

Proof of Thm. 1. By Assumption 1, ψ is L_ψ -Lipschitz on $\mathcal{N}$, hence $|\psi(\mathbf{v}^+) - \psi(\mathbf{v})| \leq L_\psi \|\mathbf{v}^+ - \mathbf{v}\|_2$. Lemma 5 yields

$$|\psi(\mathbf{v}^+) - \psi(\mathbf{v})| \leq \frac{L_\psi \eta}{\|\mathbf{r}\|} \|\Pi_{\mathbf{v}} \Delta \mathbf{r}\|_2 + O(\eta^2).$$

For STE, $\Delta \mathbf{r} = \mathbf{g}$ (App. A.1.1), so the proxy depends on $\|\Pi_{\mathbf{v}} \mathbf{g}\|$. Decompose $\mathbf{g} = \mathbf{g}_T + a\mathbf{u}$ as in Prop. 2. Then $\|\Pi_{\mathbf{v}}(a\mathbf{u})\| = |a| \sin \theta$ (App. A.1.5), explaining the additional leakage contribution. For GAQ in the ideal case $\varepsilon = 0$, $\Delta \mathbf{r} = \mathbf{R}_\theta^T \mathbf{g}$ and the orthogonal part transports $a\mathbf{u}$ to $a\mathbf{v}$, hence contributes no tangential term: $\Pi_{\mathbf{v}}(\mathbf{H}^T(a\mathbf{u})) = \mathbf{0}$. Moreover $\|\Pi_{\mathbf{v}} \mathbf{H}^T \mathbf{g}_T\| = \|\mathbf{g}_T\|$ by orthogonality (App. A.1.5). Substituting these into the proxy completes the bound. $\square$

A.1.7 Phase coherence metrics: PC and TS

Proof of Lemma 1. Fix φ and draw $\mathbf{r}, \mathbf{r}' \stackrel{\text{i.i.d.}}{\sim} P_\varphi$. If $\tau(\mathbf{r}) \neq \tau(\mathbf{r}')$, then the selected transformed codes are distinct, hence $\|\bar{\mathbf{c}}(\mathbf{r}) - \bar{\mathbf{c}}(\mathbf{r}')\|_2 \geq m$ by definition of m . Therefore

$$\begin{aligned}\mathbb{P}(\tau(\mathbf{r}) \neq \tau(\mathbf{r}')) &\leq \mathbb{P}(\|\bar{\mathbf{c}}(\mathbf{r}) - \bar{\mathbf{c}}(\mathbf{r}')\|_2^2 \geq m^2) \\ &\leq \frac{\mathbb{E}\|\bar{\mathbf{c}}(\mathbf{r}) - \bar{\mathbf{c}}(\mathbf{r}')\|_2^2}{m^2},\end{aligned}$$

by Markov's inequality. Taking expectation over φ yields $\text{TS} \leq \text{PC}/m^2$. $\square$

A.1.8 Quantization error $\rightarrow$ pairwise distortion $\rightarrow$ separation

Proposition 4 (Quantization error controls pairwise distortion). *Let $\mathbf{r}, \mathbf{r}'$ be i.i.d. and let $\bar{\mathbf{c}}, \bar{\mathbf{c}}'$ be their assigned transformed codes. Define $\mathbf{e} = \mathbf{r} - \bar{\mathbf{c}}$, $\mathbf{e}' = \mathbf{r}' - \bar{\mathbf{c}}'$ and $Z = \|\mathbf{r} - \mathbf{r}'\|_2^2 - \|\bar{\mathbf{c}} - \bar{\mathbf{c}}'\|_2^2$. Assume $\|\mathbf{r} - \mathbf{r}'\|_2 \leq D$ almost surely. Then*

$$\mathbb{E}[Z^2] \leq 16D^2 \mathbb{E}\|\mathbf{e}\|_2^2 + 32 \mathbb{E}\|\mathbf{e}\|_2^4.$$

Proof. Let $\mathbf{d} = \mathbf{r} - \mathbf{r}'$ and $\delta = \mathbf{e} - \mathbf{e}'$. Then $\bar{\mathbf{c}} - \bar{\mathbf{c}}' = (\mathbf{r} - \mathbf{e}) - (\mathbf{r}' - \mathbf{e}') = \mathbf{d} - \delta$ and

$$Z = \|\mathbf{d}\|^2 - \|\mathbf{d} - \delta\|^2 = 2\langle \mathbf{d}, \delta \rangle - \|\delta\|^2.$$

Hence $|Z| \leq 2\|\mathbf{d}\|\|\delta\| + \|\delta\|^2 \leq 2D\|\delta\| + \|\delta\|^2$, so $Z^2 \leq 8D^2\|\delta\|^2 + 2\|\delta\|^4$. Now use i.i.d. symmetry: $\mathbb{E}\|\delta\|^2 = \mathbb{E}\|\mathbf{e} - \mathbf{e}'\|^2 \leq 2\mathbb{E}\|\mathbf{e}\|^2$ and the inequality $(a+b)^4 \leq 8(a^4 + b^4)$ to get $\mathbb{E}\|\delta\|^4 \leq 16\mathbb{E}\|\mathbf{e}\|^4$. Substituting proves the claim. $\square$

A.1.9 Motif identifiability (token enrichment)

Separation alone does not imply that a disease corresponds to enriched atoms. We therefore state a mild mechanism assumption: diseases often manifest as phase-localized morphological events.

Assumption 2 (Phase-localized events). *Fix a phase interval $I \subset [0, 1]$ with nonzero measure. There exists a measurable event set $\mathcal{U} \subset \mathbb{R}^d$ (a local morphological pattern in feature space) such that, for control and disease cohorts $S_{\text{ctrl}}, S_{\text{dis}}$, $\mathbb{P}(\mathbf{r} \in \mathcal{U} \mid \varphi \in I, S_{\text{dis}}) = \mathbb{P}(\mathbf{r} \in \mathcal{U} \mid \varphi \in I, S_{\text{ctrl}}) + \rho$ for some $\rho > 0$, and membership in $\mathcal{U}$ is stable across cycles conditional on $\varphi \in I$.*

Let $\mathcal{T} \subset \{1, \dots, N\}$ be a token set intended to represent $\mathcal{U}$ in the discrete space. The next proposition formalizes when cohort-level enrichment of tokens/motifs is guaranteed.

Proposition 5 (Motif identifiability via phase coherence). Assume Assumption 2. Suppose there exist $\eta, \eta' \in [0, 1)$ and a token set $\mathcal{T}$ such that

$$\begin{aligned}\mathbb{P}(\tau(\mathbf{r}) \in \mathcal{T} \mid \varphi \in I, \mathbf{r} \in \mathcal{U}) &\geq 1 - \eta, \\ \mathbb{P}(\tau(\mathbf{r}) \in \mathcal{T} \mid \varphi \in I, \mathbf{r} \notin \mathcal{U}) &\leq \eta'.\end{aligned}$$

Then the cohort-level token-frequency gap in I satisfies

$$\begin{aligned}\mathbb{P}(\tau(\mathbf{r}) \in \mathcal{T} \mid \varphi \in I, S_{\text{dis}}) \\ - \mathbb{P}(\tau(\mathbf{r}) \in \mathcal{T} \mid \varphi \in I, S_{\text{ctrl}}) \geq \rho(1 - \eta - \eta').\end{aligned}$$

Proof of Proposition 5. Let $p_{\text{ctrl}} = \mathbb{P}(\mathbf{r} \in \mathcal{U} \mid \varphi \in I, S_{\text{ctrl}})$ and $p_{\text{dis}} = \mathbb{P}(\mathbf{r} \in \mathcal{U} \mid \varphi \in I, S_{\text{dis}}) = p_{\text{ctrl}} + \rho$. For cohort $S \in \{S_{\text{ctrl}}, S_{\text{dis}}\}$, apply the law of total probability conditional on $\{\varphi \in I\}$:

$$\begin{aligned}\mathbb{P}(\tau \in \mathcal{T} \mid \varphi \in I, S) &= \mathbb{P}(\tau \in \mathcal{T} \mid \varphi \in I, \mathcal{U}) p_S \\ &\quad + \mathbb{P}(\tau \in \mathcal{T} \mid \varphi \in I, \mathcal{U}^c) (1 - p_S).\end{aligned}$$

The phase-stability clause in Assumption 2 ensures the two conditional accuracy bounds are comparable across cohorts. Using the hypotheses $\mathbb{P}(\tau \in \mathcal{T} \mid \varphi \in I, \mathcal{U}) \geq 1 - \eta$ and $\mathbb{P}(\tau \in \mathcal{T} \mid \varphi \in I, \mathcal{U}^c) \leq \eta'$ gives

$$\mathbb{P}(\tau \in \mathcal{T} \mid \varphi \in I, S) \geq (1 - \eta)p_S + \eta'(1 - p_S).$$

Subtracting the control expression ($p_S = p_{\text{ctrl}}$) from the disease expression ($p_S = p_{\text{ctrl}} + \rho$) yields

$$\Delta \geq ((1 - \eta) - \eta')\rho = \rho(1 - \eta - \eta'),$$

which proves the claim. $\square$

A.1.10 Implementation details relevant to GAQ

Efficient computation of $\mathbf{R}_{\theta}\mathbf{r}$. Although (5) is written as a matrix, $\mathbf{H}_{\varepsilon}$ is a rank-1 update, and $\mathbf{H}_{\varepsilon}\mathbf{r}$ can be computed in $O(d)$ time:

$$\mathbf{H}_{\varepsilon}\mathbf{r} = \mathbf{r} - 2\mathbf{w} \frac{\mathbf{w}^T \mathbf{r}}{\mathbf{w}^T \mathbf{w} + \varepsilon}.$$

This avoids forming any $d \times d$ matrices in practice.

Scale fixing for $\mathbf{E}$. We apply $\mathbf{E} \leftarrow \frac{\sqrt{d}}{\|\mathbf{E}\|_F} \mathbf{E}$ after each optimizer step, independently per residual level. This removes the non-identifiable global scaling degree of freedom and empirically prevents $\mathbf{E}$ from becoming ill-conditioned.

Abbreviation	Full name (description)
AF	Atrial fibrillation
I-AVB	First-degree atrioventricular (AV) block
LBBS	Left bundle branch block
RBBB	Right bundle branch block
CD	Conduction disturbance
HYP	Hypertrophy
LVH	Left ventricular hypertrophy
SEHYP	Septal hypertrophy
MI	Myocardial infarction
AMI	Anterior myocardial infarction
IMI	Inferior myocardial infarction
STTC	ST/T change
ISCA	Ischemic in anterior leads
AFIB	Atrial fibrillation
STACH	Sinus tachycardia

Table 10: Abbreviations of cardiac conditions and their full names

Phase-free in operation and graceful degradation. Although our analysis interprets the cancellation as removing motion on a phase manifold, GAQ itself never observes phase: $\mathbf{H}_{\varepsilon}$ depends only on the local pair $(\mathbf{r}, \bar{\mathbf{c}})$, and Assumption 1 is used only to interpret the resulting stability. The algorithm requires no R-peak detection, no segmentation, and no domain-specific fiducials. In highly nonstationary regimes, $\mathbf{H}_{\varepsilon} \rightarrow \mathbf{I}$ smoothly as $\varepsilon \rightarrow \infty$, so GAQ degrades gracefully to STE.

A.2 Cardiac Disease Abbreviations

A.3 Details of the Decoder

As the primary component of AtomECG decoder, the computation of S4D(Gu et al., 2022a) is shown in Algo. 3.

Algorithm 3 S4D in AtomECG Decoder

Require: Feature $\mathbf{g} \in \mathbb{R}^{c \times l}$.

Require: SSM Parameters $(\mathbf{A}, \mathbf{C}, \Delta\mathbf{t}, \mathbf{D})$.

- 1: $\tilde{\mathbf{C}} = \mathbf{C} \odot (\exp(\mathbf{A} \odot \Delta\mathbf{t}) - \mathbf{1}) \oslash \mathbf{A}$
- 2: **for** $i \leftarrow 0$ **to** $l - 1$
- 3: $\mathbf{k}_{:,i} = \text{Re} \left(\sum_n \tilde{\mathbf{C}}_{:,n} \odot \exp(\mathbf{A}_{:,n} \odot \Delta\mathbf{t} \cdot i) \right)$
- 4: **end for**
- 5: $\mathbf{k}_{\text{fft}} \leftarrow \text{RFFT}(\mathbf{k}, n = 2l)$
- 6: $\mathbf{g}_{\text{fft}} \leftarrow \text{RFFT}(\mathbf{g}, n = 2l)$
- 7: $\mathbf{y} \leftarrow \text{iRFFT}(\mathbf{g}_{\text{fft}} \odot \mathbf{k}_{\text{fft}})[:, : l]$
- 8: $\hat{\mathbf{x}}^{c \times l} = \text{Linear}(\text{GELU}(\mathbf{y} + \mathbf{g} \odot \mathbf{D}))$
- 9: **Return** $\hat{\mathbf{x}} \xrightarrow{c \rightarrow c}$

- (Lines 1-4): A convolutional kernel $\mathbf{k}$ of length l is generated from the SSM parameters $(\mathbf{A}, \mathbf{C}, \Delta\mathbf{t}, \mathbf{D})$, embedding the model’s dynamics into the kernel.

- (Lines 5-7): Convolution is performed via Fast Fourier Transform(FFT), where the input $\mathbf{g}$ and kernel $\mathbf{k}$ are transformed into the frequency domain, multiplied element-wise, and then mapped back to the time domain with an inverse FFT. RFFT denotes the FFT for real-valued inputs.
- (Line 8): The result of the convolution is added to a skip connection controlled by the parameter $\mathbf{D}$. Finally, the output undergoes a GELU activation followed by a linear layer.

A.4 Datasets Preprocessing and Metric

Pre-train Preprocessing. The preprocessing pipeline sequentially applies NaN/Inf value handling, amplitude outlier exclusion (threshold ± 11), statistical filtering (mean $\in[-1.5, 1.5]$, variance $\in[0.05, 2]$), and temporal variation filtering (threshold 10 units for adjacent differences).

Metric.

- Macro AUROC: In multi-label classification, the area under the receiver operating characteristic curve is computed for each class independently by varying the classification threshold, and the final score is obtained by averaging the per-class AUCs.
- Macro F1: Computes the harmonic mean of precision and recall for each class individually, then averages these F1 scores.
- Perplexity: Quantifies the uniformity of vector usage within a codebook. Defined as: $\text{Perplexity} = \exp\left(-\sum_{c=1}^K p(c) \log p(c)\right)$. Here, $-\sum_{c=1}^K p(c) \log p(c)$ represents the entropy of the codebook distribution. Higher perplexity indicates greater uncertainty in vector selection. When vectors are uniformly used ($p(c) = 1/K$), perplexity equals the codebook size K . Conversely, concentrated usage reduces the perplexity.

GAQ diagnostics: LEAK, TANFRAC, TS, and ORTH. To empirically validate the gradient-geometry claims of GAQ, we report four scale-stable diagnostics computed on MIMIC-IV-ECG during pre-training. Unless otherwise stated, all statistics are aggregated over sampled patches from

lead II and are computed on the first residual level ($i=0$), which is also the level used for grammar modeling.

Notation. For a patch residual $\mathbf{r} \in \mathbb{R}^d$, let $\bar{\mathbf{c}} = \mathbf{E}\mathbf{c}_{\tau(\mathbf{r})}$ be the assigned transformed code (with $\mathbf{E} = \mathbf{I}$ if $\mathbf{E}$ is ablated), $\mathbf{u} := \bar{\mathbf{c}}/\|\bar{\mathbf{c}}\|_2$, $\mathbf{v} := \mathbf{r}/\|\mathbf{r}\|_2$, and $\theta := \arccos(\mathbf{u}^\top \mathbf{v})$. Let $\tilde{\mathbf{q}}$ denote the quantizer output fed to the decoder (Eq. (5) for GAQ and Eq. (8) for STE), and let $\mathbf{g} := \nabla_{\tilde{\mathbf{q}}} \mathcal{L}$ be the upstream gradient from the reconstruction objective $\mathcal{L}$. We use a small numerical constant $\varepsilon = 10^{-8}$ when normalizing vectors.

Leak severity and Leak₉₀. Proposition 2 isolates a radial-to-tangential “leakage channel” under STE whose magnitude scales with $|a| \sin \theta$, where $a = \mathbf{u}^\top \mathbf{g}$. We form a scale-invariant proxy by normalizing the upstream gradient $\hat{\mathbf{g}} := \mathbf{g}/(\|\mathbf{g}\|_2 + \varepsilon)$ and defining

$$\text{LEAK}(\mathbf{r}) := \left| \mathbf{u}^\top \hat{\mathbf{g}} \right| \sin \theta, \quad \theta = \arccos(\mathbf{u}^\top \mathbf{v}). \quad (10)$$

This statistic lies in $[0, 1]$ and summarizes how strongly a code-direction component can induce tangential motion at $\mathbf{v}$ when $\theta \neq 0$. We report **Leak₉₀** as the 90th percentile of LEAK over evaluated patches (computed per seed and then averaged across seeds), emphasizing tail misalignment cases.

Tangential update fraction (TanFrac). To quantify how much of the encoder-side update lies in the tangent space of the unit sphere at $\mathbf{v}$, define the projector $\Pi_{\mathbf{v}} := \mathbf{I} - \mathbf{v}\mathbf{v}^\top$ and

$$\text{TANFRAC}(\mathbf{r}) := \frac{\|\Pi_{\mathbf{v}} \Delta \mathbf{r}\|_2}{\|\Delta \mathbf{r}\|_2 + \varepsilon}, \quad (11)$$

where $\Delta \mathbf{r}$ is the backpropagated encoder-side update direction at $\mathbf{r}$. Under STE, $\Delta \mathbf{r} = \mathbf{g}$ (App. A.1.1), whereas under GAQ we use $\Delta \mathbf{r} = \mathbf{R}_\theta^\top \mathbf{g}$ with $\mathbf{R}_\theta$ defined in Eq. (5). Lower TanFrac indicates more radial (less tangential) updates, consistent with reduced one-step phase drift in Theorem 1. We report TanFrac as mean $\pm$ std across random seeds.

Token switch rate (TS) within phase bins. To quantify token-level phase coherence, we estimate within-beat phase and measure how often tokens disagree across beats when the phase is approximately matched. We obtain beat boundaries on lead II using a standard R-peak detector, define the within-beat phase $\varphi(t) \in [0, 1)$ by normalizing time between consecutive R peaks, and assign each patch the phase of its center timestamp. We discretize $[0, 1)$ into uniform bins ($B = 20$ uniform

phase bins). Within each bin we sample pairs of patches from different beats and compute the disagreement indicator $\mathbb{1}\{\tau(\mathbf{r}) \neq \tau(\mathbf{r}')\}$. TS is the empirical average disagreement probability across bins, matching Eq. (7). We report TS as mean $\pm$ std across random seeds.

Near-orthogonality diagnostic (Orth and Orth₉₀). GAQ uses the stabilized Householder map

$$\mathbf{H}_\varepsilon(\mathbf{r}, \bar{\mathbf{c}}) = \mathbf{I} - 2 \frac{\mathbf{w}\mathbf{w}^\top}{\mathbf{w}^\top \mathbf{w} + \varepsilon}, \quad \mathbf{w} = \mathbf{u} - \mathbf{v},$$

whose deviation from exact orthogonality vanishes as $\varepsilon \rightarrow 0$ (Lemma 4). We quantify numerical non-orthogonality per patch by the normalized Frobenius deviation ($\|\mathbf{I}\|_F = \sqrt{d}$)

$$\text{ORTH}(\mathbf{r}) := \frac{1}{\|\mathbf{I}\|_F} \left\| \mathbf{H}_\varepsilon(\mathbf{r}, \bar{\mathbf{c}})^\top \mathbf{H}_\varepsilon(\mathbf{r}, \bar{\mathbf{c}}) - \mathbf{I} \right\|_F. \quad (12)$$

We report **Orth₉₀** as the 90th percentile of ORTH over evaluated patches (computed per seed and then averaged across seeds). For STE (which does not form $\mathbf{H}_\varepsilon$), Orth is not defined.

Estimation protocol. For each validation batch, we sample 256 patches uniformly over time (lead II), compute LEAK, TANFRAC, TS and ORTH, and aggregate over batches. Leak₉₀ and Orth₉₀ are computed per seed and then averaged across seeds. TanFrac and TS are reported as mean $\pm$ std across five random seeds.

A.5 Cohort-enriched atoms and common subsequences via chi-squared tests

We identify class-associated atom indices and common subsequences from AtomECG token sequences using a χ^2 test of independence, and report significant motifs in Table 2.

Token sequences and motifs. For each individual $i \in \{1, \dots, n\}$, AtomECG yields an index sequence $\mathbf{z}_i = (z_{i1}, \dots, z_{iT_i})$, where $z_{it} \in \{1, \dots, N\}$. We consider two motif types: (i) a single atom $m = a$, and (ii) a common subsequence $m = (a_1, \dots, a_L)$ (order-preserving, not necessarily contiguous). We enforce a maximum window (span) $W = 30$ tokens: a subsequence m occurs in $\mathbf{z}_i$ if there exist indices $t_1 < \dots < t_L$ such that

$$z_{it_\ell} = a_\ell \quad (\ell = 1, \dots, L), \quad t_L - t_1 + 1 \leq W (= 30). \quad (13)$$

This captures local, phase-like compositions while allowing short gaps within the window.

Individual-level occurrence and usage ratio.

We use a binary indicator per individual:

$$I_i(m) := \mathbb{1}\{m \text{ occurs in } \mathbf{z}_i \text{ under (13)}\}. \quad (14)$$

For cohort (subclass) label ℓ , let $y_i^{(\ell)} \in \{0, 1\}$ denote membership. We report the usage ratio as

$$\begin{aligned} \text{Usage}(m, \ell) &= \mathbb{P}(I(m) = 1 \mid y^{(\ell)} = 1) \\ &= \frac{\sum_i \mathbb{1}\{y_i^{(\ell)} = 1, I_i(m) = 1\}}{\sum_i \mathbb{1}\{y_i^{(\ell)} = 1\}}. \end{aligned} \quad (15)$$

2×2 contingency table and χ^2 test. For each pair (m, ℓ) , we form counts $n_{rc} = \sum_i \mathbb{1}\{y_i^{(\ell)} = r, I_i(m) = c\}$ for $r, c \in \{0, 1\}$, with total $n = \sum_{r,c} n_{rc}$ and expected counts under independence $e_{rc} = \frac{(\sum_c n_{rc})(\sum_r n_{rc})}{n}$. We compute the Pearson statistic

$$\chi^2(m, \ell) = \sum_{r \in \{0,1\}} \sum_{c \in \{0,1\}} \frac{(n_{rc} - e_{rc})^2}{e_{rc}}, \quad (16)$$

and obtain a p -value using a χ^2 distribution with 1 degree of freedom. When expected counts are sparse (e.g. $\min_{r,c} e_{rc} < 5$), we replace χ^2 by Fisher’s exact test on the same table. We report motifs with $p < 0.05$.

Mining candidate common subsequences (span $W=30$). Enumerating all subsequences is intractable, so we mine candidates by frequent sequential pattern mining with a span constraint. Concretely, we collect subsequences satisfying (13) and retain only those with sufficient document frequency $\text{df}(m) = \sum_i I_i(m)$ (minimum support) to ensure stable testing. We then run the above χ^2 /Fisher procedure for each cohort ℓ and report significant atoms and subsequences together with their usage ratios.

Because we test a large number of motifs, we additionally require that a reported motif is reproducible. Concretely, we perform motif testing on a discovery split and re-test the same motifs on an independent validation split. A motif is reported for cohort ℓ only if it satisfies $p < 0.05$ on both splits (using the same χ^2 /Fisher procedure). In K -fold settings, we alternatively report motifs that achieve $p < 0.05$ in at least $K - 1$ folds.

A.6 Implementation Details

Pre-training settings. The performance of all baselines and eSSL models was evaluated under consistent preprocessing, strictly following their publicly

available code and training configurations. During pre-training on MIMIC-IV-ECG, 95% is used for training and 5% for validation, where the validation set is used to monitor training and prevent code-book collapse. For training on VitalDB, we segment each patient’s data into 1-minute clips, with a processing window of 1 minute. All other procedures remain identical to those used for MIMIC-IV-ECG. The overall architecture of AtomECG comprises three main components: an encoder, a quantization module and a decoder. The encoder architecture draws inspiration from PatchTST (Nie et al., 2023). The detailed hyperparameters are as follows:

- Encoder: Number of layers: $m = 6$
- Encoder: Number of attention heads: 4
- Encoder: Hidden dimension (d_{model}): 128
- Encoder: Feed-forward network dimension: 512
- Quantization: Residual layer: $h = 4$
- Decoder: Number of S4D layers: 4
- Decoder: State-space dimension: 64
- Decoder: Dropout rate: 0.1
- Decoder: Minimum and maximum values for time step: 0.001 and 0.1
- Training on MIMIC-IV-ECG: Epoch-150, Optimizer-AdamW-1e-4, Batch-128, Masking rate-40%
- Training on VitalDB: Epoch-100, Optimizer-AdamW-5e-5, Batch-64, Masking rate-40%

The loss function in pre-training can be written as $\mathcal{L} = \text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{L}_{\text{vq}} + \mathcal{L}_{\text{commit}}$, where

$$\mathcal{L}_{\text{vq}} = \sum_{i=0}^{h-1} \left(\alpha_i \|\text{sg}(\mathbf{r}^{(i)}) - \mathbf{E}^{(i)} \mathbf{Q}^{(i)}\|_2^2 \right),$$

$$\mathcal{L}_{\text{commit}} = \sum_{i=0}^{h-1} \left(\beta_i \|\mathbf{r}^{(i)} - \text{sg}(\mathbf{E}^{(i)} \mathbf{Q}^{(i)})\|_2^2 \right).$$

And for simplicity, we set the vq and commitment weights to $\alpha = \beta = 0.25$.

Cross-validation protocol (no test leakage).

For every result that involves tokenizer fine-tuning, we exclude test-fold data from all stages of representation learning. This includes multi-label and

multi-class classification (Table 1 and Fig. 2) and next-atom prediction (Tables 3 and 5). We first pre-train the tokenizer on the unlabeled MIMIC-IV-ECG or VitalDB pool. For each downstream fold, we then fine-tune the tokenizer with the masked reconstruction objective using only the training split. Atom indices, downstream classifiers, and the GPT grammar are trained on the same split and evaluated on the held-out fold. The five folds are run independently. This nested protocol prevents test exposure through either the vocabulary or decoder.

Downstream classification. Different classification tasks require different learning rates. We report the choices for multi-label and multi-class settings under linear and MLP probing.

Task	lr
all	1e-4
diag.	1e-4
sub.	1e-4
super.	5e-5
form	1e-5
rhythm	5e-5
G12EC	1e-4

Table 11: The learning rate on multi-label classification tasks under linear probing.

Class	lr on PTB-XL	lr on G12EC
AF	4e-5	8e-5
I-AVB	5e-5	6e-5
LBBB	5e-5	6e-5
RBBB	5e-5	6e-5

Table 12: The learning rate on multi-class classification tasks under MLP probing.

We adopt different probing strategies across the two tasks for the following reasons. For the multi-label task, linear probing is sufficient to compare the representational strength of different eSSL models. However, according to the supervised results from ECG baselines (Nonaka and Seita, 2021), no eSSL model consistently outperforms ResNet-18. Therefore, in the subsequent multi-class task, we additionally evaluate nonlinear probing. The results show that all eSSL models significantly surpass ResNet-18. We attribute this to the diverse diagnostic labels and their strong inter-label dependencies, where linear probing alone struggles to achieve accurate discrimination on small-scale datasets. Hence, we adopt different probing strategies for the two tasks, with the only difference being that the linear classifier is replaced by an

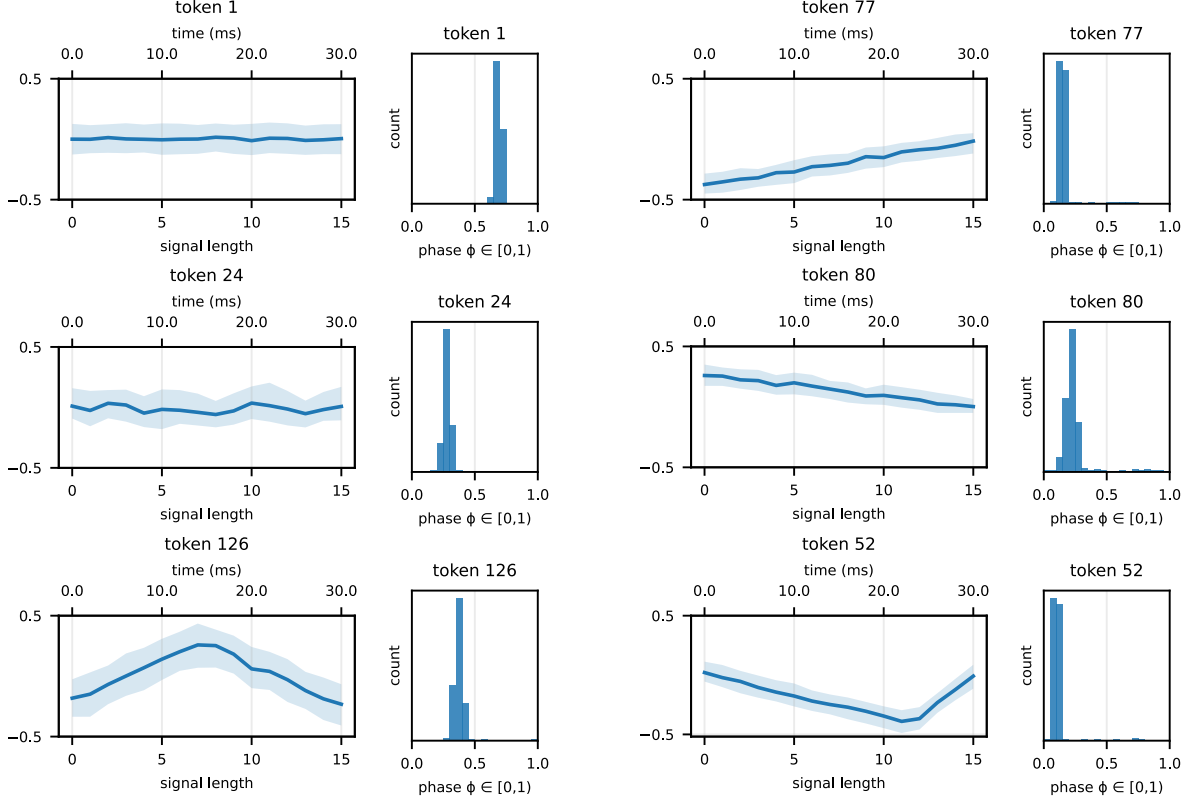


Figure 4: Visualization of representative ECG atoms on PTB-XL (lead II) using the AtomECG tokenizer fine-tuned on PTB-XL (first residual level). For each atom, the left panel summarizes amplitude-normalized patches by the 25–75% quantile band (shaded) with the median curve in bold. The right panel shows the normalized histogram of within-beat phase $\phi \in [0, 1)$ of patch centers computed between consecutive R peaks, indicating where the atom tends to occur within the cardiac cycle.

MLP (Linear–ReLU–Linear).

Next-atom prediction. We evaluate whether the learned atom sequence admits an autoregressive grammar by training a causal Transformer (GPT) to predict the next atom index and decoding predicted indices back to waveform samples.

Let $\mathbf{z} = (z_1, \dots, z_T)$ be the atom-index sequence extracted from an ECG (we use the first residual-level codebook to define the vocabulary). The training and evaluation pipeline is:

- **Tokenizer fine-tuning.** Fine-tune AtomECG on the target dataset with the masked reconstruction objective.
- **Token extraction.** Freeze the tokenizer (encoder + codebook) and convert each ECG into an index sequence $\mathbf{z}$.
- **GPT training (context length 256).** Train a causal Transformer on index sequences with standard next-token cross-entropy. We form training pairs by sliding a window of length

$C=256$ over $\mathbf{z}$:

$$\text{input } (z_{t-C}, \dots, z_{t-1}) \rightarrow \text{target } z_t,$$

and optimize $\sum_t -\log p(z_t | z_{t-C:t-1})$.

- **Autoregressive generation and decoding.** At inference time, given an observed context of 256 indices, GPT autoregressively generates future indices. The generated indices are mapped to code vectors by codebook lookup and then decoded by the (fine-tuned) AtomECG decoder to produce waveform samples.

The reported horizons $\{250, 500\}$ on PTB-XL and $\{1000, 2000\}$ on VitalDB refer to the number of waveform samples decoded in the prediction window (i.e. the signal-domain continuation length). Accordingly, we generate the required number of future indices (under the fixed patching configuration) so that decoding yields the desired number of samples.

GPT hyperparameters. Unless otherwise stated, we use: layers= 6, heads= 8, hidden size $d_{\text{model}} =$

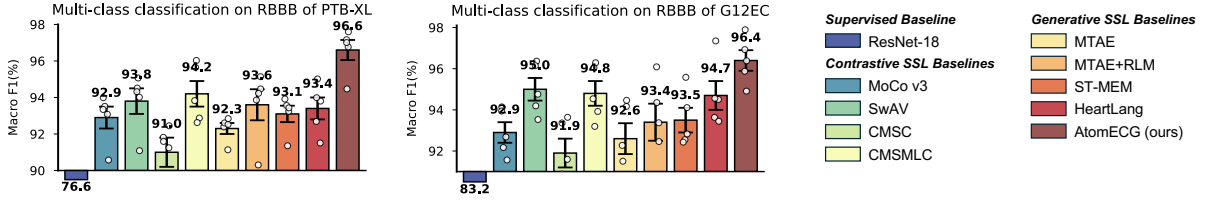


Figure 5: Macro F1 of AtomECG and nine other models under MLP probing across multi-class of RBBB on PTB-XL and G12EC.

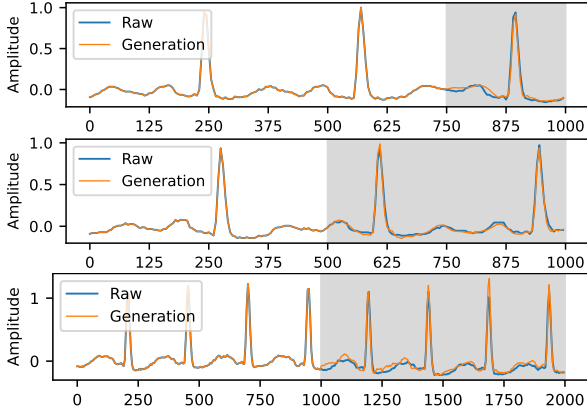


Figure 6: Next-atom prediction (lead II) with 250/500-signal-length on PTB-XL and 1000-signal-length on VitalDB.

512, optimizer AdamW with learning rate 3×10^{-4} , batch size 64, and train for 100 epochs.

A.7 Supplementary Experiments

A.7.1 Paired tests and tokenizer-sensitive metrics

Table 13 reports paired comparisons between GAQ and SimVQ over 15 matched runs from five folds and three seeds. GAQ improves all seven subsets. Six gains are significant under a paired Wilcoxon signed-rank test, while the super-diagnostic subset has $p=0.054$.

Classification is less sensitive to tokenizer geometry than autoregressive prediction and motif stability. Table 14 fixes the GPT and decoder and changes only the tokenizer. GAQ yields the lowest forecasting error and token-level perplexity, and the highest split-replicated motif reproducibility.

A.7.2 Forecasting adaptation, capacity, and compute

All trainable methods in Table 3 use target-domain adaptation, identical folds, and identical sample counts. The index forecasters share a six-layer, eight-head GPT with hidden size 512 and an S4D decoder of identical architecture and training recipe.

Each tokenizer has its own codebook and therefore its own decoder parameters, but decoder capacity and optimization are held fixed. TimeGPT parameter counts are unavailable because its weights are closed. All GPU-hours below were measured on one NVIDIA H200.

GAQ adds four 128×128 matrices, totaling 65,536 parameters. This is 0.30% of the 21.9 M AtomECG+GPT pipeline. The Householder transport has no trainable parameters and is implemented as a rank-one update with $O(d)$ cost per patch. Table 15 shows that target adaptation helps HeartLang, while doubling its GPT depth and training budget closes only a small part of the gap. Two continuous forecasters also remain behind AtomECG+GPT.

A.7.3 Cross-signal evaluation

Cross-backbone results now appear in Table 9 of the main paper. Here we provide the complete cross-signal breakdowns.

Known-phase limit cycle. We construct a synthetic limit cycle containing two nearly identical local morphologies at different phases. Its known phase permits a direct measurement of the one-step proxy in Theorem 1. GAQ reduces phase drift by about 4.5 times relative to STE (Table 16). The leakage proxy is not monotone for RotVQ because it changes gradient transport without reshaping the assignment metric. This separation motivates coupling the two components.

Photoplethysmography. We apply the same patching, encoder, decoder, and training pipeline to the VitalDB PLETH channel. PPG supplies a second physiological signal with a phase-like trajectory. GAQ improves reconstruction, phase coherence, and next-token prediction over STE (Table 17).

Visualization of representative atoms and their phase locations. To qualitatively inspect the learned atom space, we visualize the waveform morphology and the within-beat phase location

Subset	GAQ	SimVQ	Paired Δ	95% CI	p
all	0.812	0.803	+0.009	[0.007, 0.011]	$< 10^{-3}$
diag.	0.825	0.816	+0.009	[0.006, 0.012]	$< 10^{-3}$
sub.	0.869	0.858	+0.011	[0.008, 0.014]	$< 10^{-3}$
super.	0.886	0.882	+0.004	[-0.001, 0.009]	0.054
form	0.807	0.800	+0.007	[0.005, 0.009]	$< 10^{-3}$
rhythm	0.882	0.876	+0.006	[0.004, 0.008]	$< 10^{-3}$
G12EC	0.950	0.945	+0.005	[0.004, 0.006]	$< 10^{-3}$

Table 13: Paired GAQ versus SimVQ tests over 15 matched runs. The one-sided sign test across the seven subsets gives $p=0.0078$.

Tokenizer	MAE ₂₅₀	RMSE ₂₅₀	MAE ₅₀₀	PPL	Motif
VQ	0.055	0.063	0.128	102	47%
RotVQ	0.035	0.040	0.083	40	71%
SimVQ	0.043	0.047	0.104	44	73%
GAQ	0.022	0.025	0.054	24	86%

Table 14: Tokenizer-sensitive metrics with the grammar and decoder fixed. Motif reproducibility is the fraction of Table 2 motifs with $p<0.05$ on both splits.

of selected atoms on PTB-XL. Unless otherwise stated, we use the AtomECG tokenizer fine-tuned on PTB-XL with the masked reconstruction objective, and we visualize atoms from the first residual level.

Patch extraction and preprocessing. We tokenize ECGs under the default patching configuration (patch length $b=16$). The patches used for visualization are extracted from signals after the same preprocessing pipeline used during training, and we further apply amplitude normalization to each extracted patch. We then assign each patch to an atom index using hard assignment at level $i=0$, i.e. $\tau(\mathbf{r})$.

Representative atom selection. We select representative atoms based on their frequent occurrence in the statistically significant atoms reported in Table 2, aiming to cover diverse and commonly observed morphologies.

Morphology summary (left panels). For an atom n , let $\{x^{(m)}\}_{m=1}^{M_n}$ denote all collected patches assigned to n , where each patch is a length- b waveform segment. For every discrete time position $t \in \{0, \dots, b-1\}$, we compute the empirical 25th, 50th and 75th percentiles across $\{x_t^{(m)}\}_{m=1}^{M_n}$. We plot the interquartile range (25–75%) as a shaded band and the median (50%) as a bold curve. With a 500Hz sampling rate, $b=16$ corresponds to 32 ms.

Phase histogram (right panels). To show where an atom tends to occur within a cardiac cycle, we compute a within-beat phase $\phi \in [0, 1)$ from

detected R-peak locations. Let $\{R_k\}$ be R-peak timestamps. For a patch centered at time t_c with $R_k \leq t_c < R_{k+1}$, we define

$$\phi(t_c) = \frac{t_c - R_k}{R_{k+1} - R_k}.$$

We exclude patches whose centers are not bracketed by two consecutive R peaks. For each atom, we plot a histogram of $\phi(t_c)$ over its assigned patches using uniform bins on $[0, 1)$, while the y-axis reports normalized frequency.

Multi-class classification. Fig. 5 reports Macro F1 under MLP probing for multi-class classification within the RBBB cohort on PTB-XL and G12EC. AtomECG consistently achieves the best performance among supervised, contrastive and generative SSL baselines. The result supports our central claim that discrete sub-waveform atoms capture diagnostically relevant morphologies and context, enabling strong discrimination even under a lightweight probe.

Visualization of the next-atom prediction. Moreover, for next-atom prediction, Fig. 6 shows visualization on lead II for PTB-XL (250/500-signal-length) and VitalDB (1000-signal-length). AtomECG produces continuations that closely follow the raw morphology. These qualitative examples provide additional evidence that the learned atoms form reusable building blocks whose compositions can be modeled as a coherent ECG grammar across different datasets and horizons.

Epinephrine case study. Fig. 7 compares atom usage during the 10-second intervals immediately before and after an epinephrine injection in one VitalDB case. The post-injection window shows increased use of atoms associated with sinus tachycardia. This shift is compatible with the known stimulant effect of epinephrine. The example is illustrative and does not establish event-detection performance.

Method	Seq. params	GPU-h	PTB-XL 500	VitalDB 1000
HeartLang, no FT	20.1 M	25.0	0.341 / 0.358	0.560 / 0.605
HeartLang, target FT	20.1 M	34.0	0.280 / 0.292	0.482 / 0.525
HeartLang, 12L, 2× epochs	39.0 M	61.5	0.264 / 0.279	0.451 / 0.494
PatchTST forecaster	8.9 M	6.4	0.183 / 0.201	0.331 / 0.362
DLinear	0.05 M	1.3	0.229 / 0.248	0.402 / 0.431
AtomECG+GPT	20.1 M	38.5	0.054 / 0.061	0.062 / 0.069

Table 15: Additional adaptation and capacity controls for forecasting. Entries for PTB-XL and VitalDB are MAE / RMSE.

Quantizer	Drift ($\times 10^{-2}$)	Leak P90	TS
STE	4.9	0.71	0.18
SimVQ	3.2	0.49	0.11
RotVQ	2.3	0.69	0.12
GAQ	1.1	0.31	0.05

Table 16: Known-phase limit-cycle results. Lower is better for every metric.

Quantizer	Rec. MSE	TS	Next MAE
STE	0.0031	0.15	0.041
GAQ	0.0016	0.05	0.024

Table 17: PPG results on VitalDB PLETH. Lower is better.

A.7.4 Clinical Robustness: detailed results

This appendix details the demographic robustness, calibration, and distortion-robustness analyses summarized in Sec. 4.6. All numbers are computed on PTB-XL multi-label classification under linear probing, with 15 paired runs from 5-fold cross-validation combined with 3 random seeds. Within each fold, the cohort is stratified into four demographic subgroups by sex and the median age (62): Male<62, Male≥62, Female<62, Female≥62. We report mean±std together with two-sided 95% bootstrap confidence intervals (5000 resamples).

Subgroup performance. Table 18 reports per-subgroup Macro AUROC on the “all”, “sub.” and “form” subsets. Across all three tasks the differences between subgroups remain within ± 0.003 AUROC.

Group	AUROC (all)	AUROC (sub.)	AUROC (form)
M<62	0.812±0.012	0.865±0.009	0.806±0.012
M≥62	0.811±0.012	0.870±0.010	0.807±0.012
F<62	0.813±0.012	0.869±0.009	0.807±0.013
F≥62	0.812±0.013	0.873±0.009	0.806±0.013

Table 18: Per-subgroup Macro AUROC of AtomECG on PTB-XL.

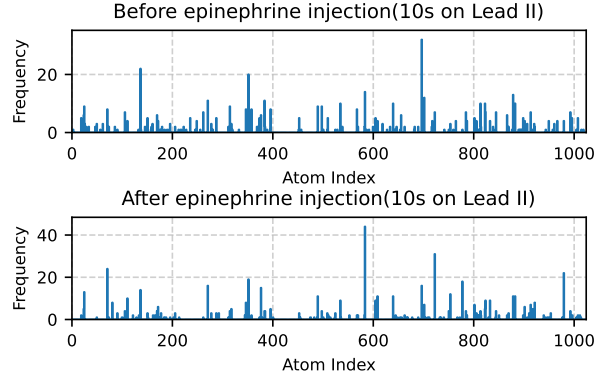


Figure 7: Atom usage frequency before and after epinephrine injection from VitalDB.

Overall robustness and calibration. We further report for each task the mean worst-subgroup AUROC, the best–worst gap, and the cross-validated worst-group Macro Expected Calibration Error (CWM), which takes the maximum subgroup ECE per label (10 confidence bins) and macro-averages across labels. CWM below 0.05 is typically considered strong (Table 19).

Task	Mean Worst	95% CI (Worst)	Mean Gap	CWM
all	0.804	[0.801, 0.808]	0.014	0.028
sub.	0.862	[0.861, 0.864]	0.010	0.030
form	0.800	[0.795, 0.804]	0.012	0.027

Table 19: Overall subgroup robustness and worst-group calibration of AtomECG.

Distortion robustness. Table 20 reports per-subgroup AUROC on the “all” subset under lead-removal scenarios (remove Lead I, remove a random lead, remove Leads I+V6, remove two random leads) and additive Gaussian noise at 20 dB SNR. AtomECG degrades by at most 0.05 AUROC under removal of two leads and stays within 0.005 AUROC under single-lead loss or AWGN. The variation is consistent across all subgroups.

Group	Base	rm Lead I	rm 1 random Lead	rm Leads I+V6	rm 2 random Leads	AWGN (20 dB)
M<62	0.812±0.012	0.808±0.016	0.809±0.018	0.764±0.035	0.783±0.029	0.809±0.014
M≥62	0.811±0.012	0.808±0.017	0.808±0.017	0.761±0.033	0.782±0.031	0.808±0.014
F<62	0.813±0.012	0.810±0.015	0.810±0.016	0.764±0.030	0.786±0.024	0.811±0.013
F≥62	0.812±0.013	0.809±0.016	0.808±0.017	0.763±0.034	0.785±0.026	0.809±0.013

Table 20: Distortion robustness on PTB-XL “all” subset.

Patch Size	MSE	VQ-loss
<i>stride-8</i>		
8	0.00390	5.71e-5
16	0.00217	2.65e-5
24	0.00180	5.17e-4
32	0.00245	2.15e-3
<i>stride-16</i>		
16	0.00762	2.01e-3
24	0.00567	1.24e-3
32	0.00564	1.03e-3
<i>stride-32</i>		
32	0.01505	1.45e-3

Table 21: Ablation of patch size and stride during pre-training on MIMIC-IV-ECG.

Atom Size	Perplexity	VQ-loss
<i>codebook-512</i>		
64	125/70/70/65	9.19e-5
128	113/101/85/79	8.75e-5
256	139/91/85/84	2.12e-4
<i>codebook-1024</i>		
64	294/203/178/171	6.01e-5
128	317/200/185/187	2.65e-5
256	299/205/184/187	7.79e-5
<i>codebook-2048</i>		
64	444/363/250/239	7.47e-5
128	606/300/288/202	7.05e-5
256	585/376/178/164	1.96e-4

Table 22: Ablation of codebook size during pre-training on MIMIC-IV-ECG.

A.8 Ablation Studies

We perform ablations to clarify how AtomECG’s core design choices affect (i) pre-training stability and reconstruction fidelity, (ii) downstream diagnostic transfer, and (iii) the learnability of the induced discrete grammar. Unless otherwise specified, all ablations follow the main setting: pre-training on MIMIC-IV-ECG, linear probing on PTB-XL and G12EC for multi-label classification, and GPT-based next-atom prediction on PTB-XL.

Patch size and stride. AtomECG tokenizes ECGs by uniformly patching the waveform. The patch length b and stride s control the granularity and redundancy of morphological units: smaller b

increases atomic resolution but may fragment clinically meaningful micro-structures, while larger s reduces overlap but risks losing fine temporal dynamics and harms masked reconstruction. Table 21 reports the reconstruction MSE and the quantization objective (VQ-loss) during pre-training on MIMIC-IV-ECG. Two trends are consistent. First, increasing stride substantially degrades reconstruction (e.g. MSE rises from $\approx 10^{-3}$ under stride-8 to $\approx 10^{-2}$ under stride-32), suggesting that dense coverage and overlap are crucial to preserve subtle morphologies under masked modeling. Second, patch length exhibits a trade-off: very small patches (e.g. $b=8$) yield higher MSE than intermediate lengths, while overly large patches (e.g. $b=32$) increase VQ-loss, indicating harder assignment and less stable discretization. Overall, we adopt the $b=16$ and $s=8$ configuration as the default tokenizer granularity.

Codebook size and atom size. We ablate the vocabulary capacity by varying the codebook size N and the atom embedding dimension d (Table 22). We report (i) codebook-usage perplexity (entropy-based) at each residual level, which measures the effective number of atoms actually used, and (ii) VQ-loss, which reflects assignment difficulty and quantization stability. Note that this perplexity is upper bounded by N and equals N only under perfectly uniform usage. In practice, it is governed by the intrinsic diversity of sub-waveform morphologies and the optimization dynamics under hard assignment. Therefore, increasing N does not imply a proportional increase in perplexity: once N exceeds the effective support of the data distribution, the additional codes tend to remain rarely used, leading to a sublinear growth (or saturation) of perplexity.

To compare codebook sizes more fairly, we can view the utilization ratio $\rho := \text{Perplexity}/N$. We observe that $N=1024$ yields the best overall utilization (e.g. 317/1024 at the first residual level) together with the lowest VQ-loss, indicating a good balance between expressiveness and stable assign-

Quantization	E	R_θ	Residual	all	diag.	sub.	super.	form	rhythm	G12EC
✓	-	-	-	0.765	0.764	0.805	0.809	0.750	0.842	0.921
✓	✓	✓	-	0.799	0.804	0.848	0.836	0.777	0.856	0.939
✓	-	✓	✓	0.782	0.786	0.849	0.821	0.766	0.845	0.932
✓	✓	-	✓	0.770	0.779	0.835	0.825	0.765	0.844	0.939
✓	✓	✓	✓	0.812	0.825	0.869	0.886	0.807	0.882	0.950

Table 23: Ablation of the GAQ on multi-label classification via linear probing on PTB-XL and G12EC

Quantization	E	R_θ	Residual	MAE	RMSE	Perplexity
✓	-	-	-	0.063	0.075	210
✓	✓	✓	-	0.032	0.034	425
✓	-	✓	✓	0.029	0.033	$\left(\begin{array}{c} 225/176 \\ 145/135 \end{array} \right)$
✓	✓	-	✓	0.045	0.047	$\left(\begin{array}{c} 420/315 \\ 213/187 \end{array} \right)$
✓	✓	✓	✓	0.022	0.025	$\left(\begin{array}{c} 476/324 \\ 202/199 \end{array} \right)$

Table 24: Ablation of the GAQ on next-atom prediction(250-signal-length on PTB-XL).

ments. In contrast, enlarging the codebook to $N=2048$ increases the nominal perplexity, but the gain is not commensurate with the doubled capacity and is uneven across residual levels, suggesting redundancy and less consistent usage. Regarding atom size, increasing d enlarges per-atom expressiveness but can also make alignment harder under VQ, which may raise VQ-loss without improving utilization. Overall, we adopt $N=1024$ and $d=128$ as the default configuration, as it achieves the most favorable trade-off between utilization and quantization stability.

Disentangling GAQ components. Geometry Alignment Quantization (GAQ) contains two mechanisms: (i) a learnable global linear transformation **E** that reshapes the codebook geometry, and (ii) a local geometry-aware transport **R_θ** used in the backward surrogate to align gradients under hard assignments. In contrast, residual quantization (multi-level codebooks) is another trick that can be combined with GAQ. Accordingly, we ablate three factors: **E**, **R_θ**, and residual quantization. We evaluate their effects on diagnostic transfer and grammar learnability.

(1) Downstream diagnosis. Table 23 reports linear probing AUROC on PTB-XL and G12EC. Starting from a standard VQ baseline (Quantization only), adding GAQ yields a consistent and sizable improvement across *all* subsets (e.g. PTB-XL: 0.765 $\rightarrow$ 0.799, G12EC: 0.921 $\rightarrow$ 0.939), supporting that geometry reshaping plus geometry-consistent gradient alignment produces a more

discriminative and clinically faithful atom space. Residual quantization alone improves some subsets but is less consistent, suggesting that additional capacity does not automatically translate to better discrete representations when the geometry mismatch remains unresolved. When combining residual quantization with GAQ, performance improves further and achieves the best overall results, indicating that GAQ primarily addresses assignment quality and stability, while residual levels mainly contribute representational capacity. The two are complementary.

(2) Next-atom prediction. Table 24 evaluates next-atom prediction on PTB-XL. Notably, we only use the first-level codebook to form the token sequence for GPT training and inference. Under this protocol, one might expect residual quantization to decrease index-prediction accuracy, since the first-level indices now represent a coarser approximation of the signal and discard information carried by higher residual levels.

Interestingly, we observe the opposite: enabling residual quantization improves waveform-level forecasting after decoding (e.g. MAE/RMSE improves from 0.032/0.034 to 0.022/0.025 when combined with GAQ). We attribute this gain to a denoising effect induced by the residual decomposition. Specifically, the first-level codebook is encouraged to capture the most stable, high-energy morphological structure, while finer-scale variations and transient artifacts are pushed into later residual levels. As a result, the first-level token sequence becomes a cleaner and more regular ab-

Grammar Model	Context	Layers	d_{model}	PPL	MAE / RMSE
Unigram (freq)	–	–	–	488	0.850 / 0.887
Bigram (Markov-1)	1	–	–	378	0.337 / 0.380
Trigram (Markov-2)	2	–	–	340	0.294 / 0.326
GPT (small)	64	3	256	58	0.067 / 0.077
GPT (medium)	128	6	512	40	0.058 / 0.069
GPT (deep)	128	12	512	32	0.052 / 0.062
GPT (long context)	256	6	512	24	0.049 / 0.057

Table 25: Ablation of grammar modeling for next-atom prediction on PTB-XL (Lead II, 500-signal-length, pre-training on MIMIC-IV-ECG). We report token-level predictability as standard language-model perplexity $\text{PPL} = \exp\left(-\frac{1}{N} \sum_{t=1}^N \log p(x_t | x_{<t})\right)$ and waveform-level forecasting error after decoding (MAE/RMSE). Note this differs from the codebook-usage perplexity (entropy-based) used in the main text.

straction of the ECG, which is easier for an autoregressive grammar model to learn and extrapolate. Under this interpretation, residual quantization filters fine-scale variability into later levels. The first-level sequence becomes more regular and easier to predict over long contexts.

Ablation of grammar modeling. Finally, we ablate the grammar model used for autoregressive next-atom prediction (Table 25). We compare shallow Markov baselines (unigram/bigram/trigram) against GPT variants with increasing context length and capacity. Here we report token-level language-model perplexity (PPL) computed from next-token log-likelihood, which measures how predictable the atom sequence is under the grammar model. Note that this PPL is different from the codebook-usage perplexity reported elsewhere in the paper, which is an entropy-based statistic over the empirical distribution of atom indices and quantifies how uniformly the codebook entries are utilized.

Trigram models reduce PPL relative to unigram, but remain far from adequate: their decoded forecasts have large MAE/RMSE, indicating that short-range statistics cannot capture the long-horizon compositional rules of ECG dynamics. In contrast, GPT dramatically improves token predictability and signal accuracy, and performance scales with longer context. The long-context configuration achieves the best results, highlighting that clinically meaningful ECG grammar is inherently long-range and benefits from modeling extended rhythmic and morphological dependencies.

A.9 Reproducibility Statement

We have made extensive efforts to ensure the reproducibility of our work. The datasets used in our experiments are publicly available, and the pre-processing steps are detailed in App. A.4. The

architecture details, training hyperparameters, and evaluation protocols are described in App. A.6.