SPARK: Lightweight Adaptation of Healthcare Foundation Models via Steering Knowledge Circuits

Bing Li BLMCG@MST.EDU
Department of Computer Science
Missouri University of Science and Technology
Rolla, Missouri, USA

Qiang Fu QFBQT@MST.EDU
Department of Computer Science
Missouri University of Science and Technology
Rolla, Missouri, USA

Yuwei Long YLW22@MST.EDU
Department of Computer Science
Missouri University of Science and Technology
Rolla, Missouri, USA

Huiyuan Yang HYANG@MST.EDU
Department of Computer Science
Missouri University of Science and Technology
Rolla, Missouri, USA

Abstract

Large-scale foundation models (FMs) have demonstrated immense potential in medical time-series analysis, yet their static, "one-size-fits-all" nature limits their efficacy in patient-specific clinical settings. Traditional adaptation paradigms, such as full fine-tuning or standard parameter-efficient fine-tuning (PEFT), are computationally prohibitive for edge-device deployment and often overfit when faced with the chronic data scarcity and long-tail distributions characteristic of healthcare. In this work, we reveal that pre-trained medical FMs inherently possess sparse, multi-layer "Physiological Knowledge Circuits" dedicated to representing distinct clinical patterns. Leveraging this mechanistic insight, we propose **SPARK** (**S**teering **P**ersonalized **A**daptation via **R**outing **K**nowledge), a novel, ultra-lightweight adaptation framework. Instead of updating the model's global weights, SPARK utilizes a Multi-Layer Personalized Steering Hub (M-PSH) to surgically inject targeted perturbations into the hidden state activations of these specific knowledge circuits. Experiments across multiple clinical ECG benchmarks demonstrate that SPARK achieves a favorable parameter-to-performance ratio and excels in extreme data-scarce environments, yielding an approximately 6.8% accuracy improvement in 1-shot adaptation scenarios. Furthermore, by modulating internal representations rather than relearning them from scratch, SPARK significantly enhances model robustness against rare diseases, paving the way for privacy-preserving, dynamic, and on-device personalization in continuous health monitoring. The code of this work is available at: https://github.com/icedleo/SPARK.

1. Introduction

In recent years, large-scale Foundation Models (FMs) have demonstrated remarkable potential in health sensing and medical time-series analysis [Sun et al. (2025); Yang et al. (2026); McKeen et al. (2025); Cui et al. (2024); Abbaspourazad et al. (2023)]. By pre-training on vast, heterogeneous datasets encompassing diverse physiological signals, these models are capable of capturing universal features across domains. However, this pursuit of generality often comes at a cost: a significant trade-off in specialization when the model is applied to specific clinical datasets or specialized tasks. In healthcare contexts, physiological heterogeneity represents a core challenge, as baseline physiological metrics, metabolic levels, and disease manifestations vary substantially across individuals. Most existing foundation models are inherently static and follow a "one-size-fits-all" paradigm, failing to capture the dynamic intra-individual changes. Consequently, efficiently achieving the personalization and specialization of pre-trained models has become a critical bottleneck for the real-world deployment of digital health technologies.

Existing methods for model adaptation typically fall into two main paradigms: full fine-tuning or the integration of external auxiliary modules (e.g., linear probes or online ensembling Lee et al. (2025b)). While full fine-tuning remains the standard for domain adaptation, its application in personalized healthcare is severely hindered by practical obstacles. First, the medical domain suffers from a chronic scarcity of high-quality labeled data; clinical annotation requires expensive domain expertise, and the long-tail distribution of rare diseases often causes fully fine-tuned models to severely overfit to limited samples Ghassemi et al. (2020). Second, full-parameter updates demand substantial computational resources that are inherently incompatible with the power and memory constraints of continuous health monitoring systems and edge devices (e.g., wearable sensors) Hartmann et al. (2022). Furthermore, personalizing a foundation model for thousands of patients via full fine-tuning would require maintaining a separate, gigabyte-scale model copy for every individual, leading to an unsustainable storage explosion. To mitigate these overheads, Parameter-Efficient Fine-Tuning (PEFT) techniques—such as Low-Rank Adaptation (LoRA) Hu et al. (2022) and prompt tuning—have been increasingly adopted Konigorski et al. (2026); Garikipati et al. (2024); Le et al. (2025); Liu et al. (2025). However, while PEFT reduces the number of trainable parameters, it still relies on computationally expensive backpropagation through the model's global computational graph and introduces additional architectural overhead. This remains suboptimal for ultra-lightweight, dynamic personalization where only sparse, patient-specific physiological traits need to be adjusted on the fly.

Recent research in Natural Language Processing (NLP) Yu and Ananiadou (2024); Dai et al. (2022) introduced the concept of "*knowledge neurons*", proving that specific neurons within Large Language Models (LLMs) are responsible for storing and activating factual knowledge. Inspired by this, we investigate a pivotal question: *do similar mechanisms exist in foundation models for health time-series*? We hypothesize that for any given user, only a small, sparse subset of neurons within a large foundation model's hidden layers are critical for capturing their unique physiological traits and specific clinical manifestations.

To address the limitations of current adaptation methods, we propose a novel, lightweight framework called SPARK built upon this hypothesis. Unlike previous approaches that utilize external ensemble predictors or modify the model's global weights (e.g., ELF Lee

et al. (2025b)), our method operates directly and surgically within the internal parameter space of a frozen foundation model. We first employ attribution techniques to pinpoint a multi-layer **Physiological Knowledge Circuit**—comprising shallow "Query Neurons" that act as physiological feature detectors, and deep "Value Neurons" that dictate the final clinical prediction. We then introduce the **Multi-Layer Personalized Steering Hub (M-PSH)**. The M-PSH is a user-specific, highly lightweight module designed to learn a personalized modulation strategy that mathematically "steers" the hidden state activations of this identified circuit. This bypasses complex global gradient computations, achieving a rapid transition from "generic" to "specialized" and "personalized" with minimal data and computational overhead.

The primary contributions of this work are summarized as follows:

- **Discovery of Physiological Knowledge Circuits in Health Time-Series**: We bridge the gap between NLP and time-series analysis by successfully migrating the concept of knowledge neurons. Through experimentation, we demonstrate that health time-series models contain multi-layer, sparse sub-networks (Knowledge Circuits) dedicated to representing distinct physiological patterns. We categorize these into functional Query Neurons and Value Neurons, enhancing the mechanistic interpretability of time-series models and providing a precise target for intervention.
- **A Efficient Paradigm for Lightweight Personalization via SPARK**: To overcome the prohibitive computational and storage costs of traditional adaptation, we design the SPARK framework, powered by a Multi-Layer Personalized Steering Hub (M-PSH). Instead of fine-tuning the network's weights, M-PSH learns to inject targeted perturbations (steering vectors) directly into the hidden states of the identified Knowledge Circuit. Compared to standard PEFT techniques, this targeted steering acts more directly on the model's internal causal reasoning chain, achieving a favorable parameter-to-performance ratio for on-device deployment.
- **Favorable Performance in Personalization and Adaptation while Addressing Long-tail Challenges**: We validate our framework on two core tasks: individual-level personalization and cross-dataset rapid specialization. Our results demonstrate that steering the Knowledge Circuit significantly improves performance while maintaining minimal computational overhead. Crucially, because the framework modulates pre-existing knowledge rather than learning from scratch, it excels in extreme few-shot scenarios—achieving an approximately 6.8% improvement with just a 1-shot adaptation. Furthermore, ablation studies confirm that our surgical intervention enhances the model's robustness in detecting rare diseases, effectively mitigating the prevalent long-tail distribution issue in medical data.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work provides several foundational insights for the broader machine learning and healthcare community, moving beyond standard model application to address the structural bottlenecks of deploying Foundation Models (FMs) in clinical settings:

- **Mechanistic Interpretability Extends to Physiological Time-Series**: We demonstrate that deep time-series FMs are not entirely opaque black boxes. Similar to Large Language Models, specific physiological traits and disease patterns are functionally localized within sparse, multi-layer "Knowledge Circuits".

- **Activation Steering as a Paradigm for Edge Personalization**: In continuous health monitoring, maintaining separate fine-tuned model copies for individual patients is computationally and physically impossible. We establish that directly modulating hidden state activations (steering) provides a vastly favorable parameter-to-performance ratio compared to standard PEFT, offering a generalizable blueprint for deploying highly adaptive FMs on resource-constrained wearable devices.
- **Mitigating Extreme Data Scarcity and Long-Tail Distributions**: Rare diseases and patient-specific anomalies inherently suffer from data scarcity. By demonstrating that we can adapt models by mathematically shifting pre-existing internal knowledge—rather than learning new mappings from scratch via backpropagation—this work provides a novel, ultra-lightweight strategy for achieving extreme few-shot (e.g., 1-shot) adaptation and improving algorithmic robustness against the long-tail distributions common in clinical datasets.

2. Related Work

2.1. Foundation Models for Time-Series and Physiological Signals

Recent foundation models (FMs) have revolutionized time-series analysis by learning highly transferable representations from massive, heterogeneous datasets, such as Chronos Ansari et al. (2024), MOMENT Goswami et al. (2024), and Timer-XL Liu et al. (2024b). This paradigm is rapidly expanding to healthcare, with emerging models tailored for optical physiological signals Pillai et al. (2024) and wearable sensing Lee et al. (2025a). However, despite their strong zero-shot generalization, most existing physiological FMs are deployed merely as static feature extractors. They lack the intrinsic flexibility to accommodate the dynamic, patient-specific physiological shifts encountered in real-world clinical deployments. Our work addresses this critical gap: rather than treating the FM as a rigid black box, the SPARK framework transforms static models into adaptive systems, unlocking personalized downstream adjustments without disrupting the robust pre-trained backbone.

2.2. Parameter-Efficient Adaptation

To adapt large pre-trained models within strict computational and privacy constraints—a common scenario in healthcare—PEFT has become the standard approach. Traditional PEFT methods, such as LoRA Hu et al. (2022), DoRA Liu et al. (2024a), and adapter-based modules Karimi Mahabadi et al. (2021), reduce the trainable parameter count by injecting structural updates into frozen networks. However, these techniques remain fundamentally coarse-grained; they operate "blindly" at the layer or block level without explicitly identifying which internal parameters are responsible for specific clinical predictions. In contrast, our proposed Multi-Layer Personalized Steering Hub (M-PSH) is mechanistically informed. By exclusively targeting and steering a sparse subset of clinically relevant neurons, our method achieves a favorable parameter-to-performance ratio and enables extreme few-shot (e.g., 1-shot) adaptation that standard PEFT struggles to achieve.

2.3. Mechanistic Knowledge Localization and Editing in Transformers

A parallel line of research in Natural Language Processing (NLP) investigates how knowledge is physically stored within deep Transformers. Studies have revealed that factual associations and specific capabilities can often be localized to sparse "Knowledge Neurons"

Dai et al. (2022); Wang et al. (2022) or structured, multi-component "Knowledge Circuits" Yao et al. (2024). This has inspired model editing techniques focused on localized, surgical interventions rather than global weight updates Pan et al. (2025); Wang et al. (2024). Thus far, this mechanistic perspective has been virtually absent in physiological time-series modeling. Our framework bridges this interdisciplinary divide. We are the first to demonstrate the existence of structured multi-layer knowledge circuits (Query and Value neurons) within health FMs, and we uniquely leverage this discovery not for static factual editing, but as the precise mathematical foundation for dynamic, on-device patient personalization.

2.4. Personalized ECG Modeling and Patient-Specific Adaptation

Personalization has long been recognized as essential in ECG analysis due to the extreme inter-patient variability in cardiac morphology, rhythm patterns, and noise profiles. Existing approaches typically tackle this challenge via data-level augmentations (e.g., personalized digital twins Hu et al. (2024)), subject-specific sequence calibration Boda et al. (2023), or standard transfer learning and domain adaptation Ding et al. (2025, 2024). However, these methods predominantly personalize at the input level or the final output head, leaving the internal representation process largely ignored. **SPARK** presents an adaptation framework designed for personalized ECG analysis: internal representation steering. By directly modulating the deeper reasoning circuits responsible for clinical feature extraction, our method handles inter-patient variability more natively, significantly improving the robustness of the model, especially when detecting rare, long-tail diseases.

3. Methods

Our approach, SPARK is built on the hypothesis that for any given user, only a small, sparse subset of neurons within a large foundation model's hidden layers are critical for capturing their unique physiological traits and specific clinical manifestations (e.g., rare disease patterns). We term this specific subset of interconnected nodes the **Physiological Knowledge Circuit**. Instead of attempting to fine-tune the entire network—which is prone to overfitting in few-shot scenarios and computationally prohibitive—our framework focuses exclusively on learning to "steer" the activations of these specific KNs. This aligns the model's internal reasoning process with the user's ground truth while preserving the foundation model's generalizability. The framework is composed of three core components: (1) a static Health Time-Series Foundation Model, (2) a Knowledge Neuron Identification module, and (3) an online-learning Multi-Layer Personalized Steering Hub (M-PSH). The overall framework of SPARK is illustrated in Figure 1.

3.1. The Health Time-Series Foundation Model (FM)

The FM is a pre-trained deep learning architecture (e.g., a Transformer-based time-series model) that serves as the backbone of our system. It processes a window of physiological features x_t and, at a specific intermediate layer l, produces a hidden state representation $h_t^{(l)} \in \mathbb{R}^D$. This hidden state is then passed through the remaining layers of the network to produce a general prediction. To strictly minimize computational overhead and prevent catastrophic forgetting of general medical knowledge, the parameters of the entire FM are frozen during deployment. Our precise point of intervention is the hidden state $h_t^{(l)}$.

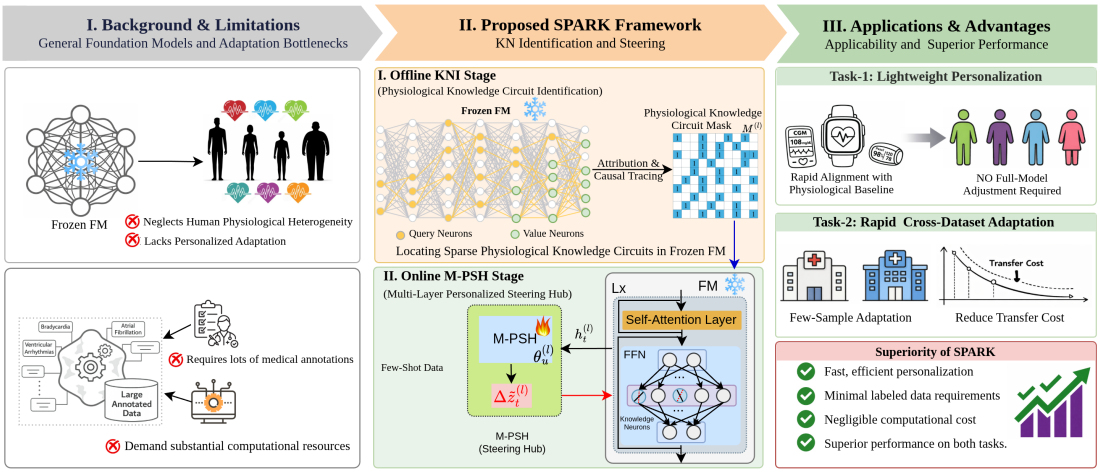


Figure 1: **Overview of SPARK. (Left)** Existing foundation-model adaptation paradigms are limited by physiological heterogeneity, label scarcity, and high computational cost. **(Middle)** SPARK uses offline Knowledge Neuron Identification (KNI) to construct a sparse physiological knowledge circuit, followed by online M-PSH to generate steering vectors from few-shot user data while keeping the foundation model frozen. **(Right)** SPARK enables efficient personalization and cross-dataset adaptation with reduced computational cost and limited forgetting.

3.2. Knowledge Neuron Identification (KNI) Module

Inspired by neuron-level attribution studies (Yu and Ananiadou (2024); Dai et al. (2022)), we define a **Physiological Knowledge Circuit** as a sparse sub-network within the foundation model's (FM) feed-forward networks (FFNs). Let d be the model's hidden dimension and N be the intermediate dimension of the FFN. For any layer l, the FFN operation is defined as $FFN\left(h^{(l)}\right) = W_{fc2}^{(l)}\sigma\left(W_{fc1}^{(l)}h^{(l)}\right)$, where $W_{fc1}^{(l)} \in \mathbb{R}^{N\times d}$ and $W_{fc2}^{(l)} \in \mathbb{R}^{d\times N}$ are the weight matrices. We define the k-th neuron in layer l by its associated vectors: (1)**Subkey** $w_{in,k}^{(l)} \in \mathbb{R}^{1\times d}$: The k-th row of $W_{fc1}^{(l)}$, which acts as a pattern detector. (2)**Subvalue** $w_{\text{out},k}^{(l)} \in \mathbb{R}^{d\times 1}$: The k-th column of $W_{fc2}^{(l)}$, which encodes the contribution to the representation stream. The identification of the circuit proceeds in two offline stages:

Stage 1: Identifying Value Neurons. We identify neurons in deeper layers that directly influence the clinical prediction y. Using Integrated Gradients (IG) Sundararajan et al. (2017), we compute an attribution score for each subvalue:

$$S_{val}(l,k) = \text{IG}(w_{out,k}^{(l)}, y) \tag{1}$$

Neurons with the highest S_{val} across calibration samples form the Value Neuron set $\mathcal{V}$.

Stage 2: Tracing Query Neurons. For each $(L,k) \in \mathcal{V}$, we trace its activators in preceding layers $l < L$ by calculating the inner product between the activator's subvalue and the target's subkey:

$$S_{query}(l,j \mid L,k) = \langle (w_{out,j}^{(l)})^\top, (w_{in,k}^{(L)})^\top \rangle \tag{2}$$

Neurons with top S_{query} scores form the Query Neuron set $\mathcal{Q}$. The full circuit is the set of indices $\mathcal{I} = \mathcal{V} \cup \mathcal{Q}$. For each layer l, we define a binary diagonal mask matrix $M^{(l)} \in \{0,1\}^{N\times N}$, where the i-th diagonal element $M_{i,i}^{(l)} = 1$ if $(l,i) \in \mathcal{I}$, and 0 otherwise.

To reduce the sensitivity of Integrated Gradients to calibration samples, we construct three calibration subsets by independently sampling 70% of the training data. Value and Query Neurons are ranked in each subset using S_{val} and S_{query}, respectively. Neurons that are consistently highly ranked across the three subsets are retained and re-ranked using their average attribution scores. The top $N(30)$ neurons are then selected to form a label-specific Physiological Knowledge Circuit.

3.3. Multi-Layer Personalized Steering Hub (M-PSH)

The M-PSH is the sole trainable component during on-device deployment, designed to learn a personalized modulation strategy. It consists of a set of lightweight, layer-wise MLPs $\left\{f_{PSH}^{(l)}\left(\cdot;\theta_u^{(l)}\right)\right\}_{l\in\mathcal{L}}$ that predict optimal perturbations for the identified Knowledge Circuit.

At each layer l, the sub-hub takes the hidden state $h_t^{(l)} \in \mathbb{R}^d$ as input and outputs a raw steering vector $\Delta z_t^{(l)} \in \mathbb{R}^N$. This vector is constrained by the circuit mask $M^{(l)}$ to ensure targeted intervention:

$$\Delta\tilde{z}_t^{(l)} = M^{(l)} \cdot f_{PSH}^{(l)}(h_t^{(l)};\theta_u^{(l)}) \tag{3}$$

The steered FFN output is then computed by injecting the masked perturbation into the intermediate activation $a_t^{(l)} = \sigma\left(W_{fc1}^{(l)} h_t^{(l)}\right)$:

$$FFN_{steered}(h_t^{(l)}) = W_{fc2}^{(l)}(a_t^{(l)} + \Delta\tilde{z}_t^{(l)}) \tag{4}$$

Learning Objective: The M-PSH is trained to mimic the optimal gradient that minimizes the clinical prediction error on a small labeled dataset. Let $\mathcal{L}_{CE}(y,\hat{y})$ be the cross-entropy loss for a labeled sample. The ground-truth target for the MLP is the gradient of this loss with respect to the FFN activation:

$$g_t^{(l)} = \nabla a_t^{(l)} \mathcal{L}_{CE} \tag{5}$$

The optimization goal for each sub-hub $f_{PSH}^{(l)}$ is to minimize the Mean Squared Error (MSE) between its masked output and the target gradient:

$$\min_{\theta_u^{(l)}} \sum_t |M^{(l)} f_{PSH}^{(l)}(h_t^{(l)};\theta_u^{(l)}) - M^{(l)} g_t^{(l)}|_2^2 \tag{6}$$

By learning to predict these "*pseudo-gradients*", the M-PSH enables the model to dynamically correct its reasoning path based on user-specific physiological patterns without updating the frozen FM weights.

4. Experimental Analysis

4.1. Datasets and Tasks

We evaluate the proposed SPARK framework on four public 12-lead ECG benchmarks: PTB-XL (Wagner et al. (2020)), CPSC2018 (Liu et al. (2018)), Chapman-Shaoxing-Ningbo (CSN) (Zheng et al. (2020)), and MIMIC-IV-ECG (Gow et al. (2023)). These datasets cover diverse clinical sources, label granularities, and levels of diagnostic difficulty, providing a

comprehensive testbed for evaluating both cross-dataset generalization and patient-specific adaptation. Detailed dataset descriptions are provided in Appendix A.1. For classification, we follow the standard split protocol commonly used in the literature. For personalization, we evaluate the model at the individual-patient level and report the final performance as the average across patients. Detailed dataset split statistics are provided in Appendix A.2.

We consider two downstream tasks: (1) **ECG classification** evaluates general diagnostic modeling under standard supervised learning. We compare SPARK against a diverse set of eSSL baselines, including SimCLR (Chen et al. (2020)), BYOL (Grill et al. (2020)), BarlowTwins (Zbontar et al. (2021)), MoCo-v3 (Chen et al. (2021)), SimSiam (Chen and He (2021)), TS-TCC (Eldele et al. (2021)), CLOCS (Kiyasseh et al. (2021)), ASTCL (Wang et al. (2023)), CRT (Zhang et al. (2023)), ST-MEM (Na et al. (2024)), and HeartLang (Jin et al. (2025)); (2) **Personalization** evaluates whether SPARK can adapt to patient-specific physiological heterogeneity in a few-shot setting, where a small number of historical ECGs from a target patient are used for adaptation and the adapted model is evaluated on the remaining recordings of the same patient.

4.2. Evaluation Metrics

We use Macro F1 and Macro AUROC as the primary evaluation metrics for both classification and personalization. These macro-averaged metrics are well suited to ECG benchmarks with substantial class imbalance, as they assign equal importance to each label. For MIMIC-IV-ECG, we additionally report AUPRC and Recall for a more comprehensive evaluation.

4.3. Experimental Settings

All experiments are built on ECG-FM (McKeen et al. (2025)), initialized from the publicly released PhysioNet-pretrained checkpoint unless otherwise specified. Raw ECG recordings are converted into a unified 12-lead format, resampled to 500Hz, segmented into 5-second windows, and normalized on a per-segment basis. The classifier is optimized with Adam using a learning rate of 10^{-3} and weight decay of 10^{-4}. Additional implementation details are provided in Appendix A.3. For all experiments, the calibration data are sampled only from the training partition. We construct three calibration subsets, each containing 70% of the training samples. After stability filtering and score aggregation, we select the top $N = 30$ Value Neurons and Query Neurons for each diagnostic label. The resulting circuit masks are identified offline and fixed during downstream adaptation. The same masks are used for all few-shot settings, where $K \in \{1, 3, 5\}$ denotes the number of labeled patient-specific records.

5. Results and Discussion

5.1. Mechanistic Evidence for Knowledge Neurons in ECG Modeling

A central question in this work is whether the identified KNs reflect a meaningful internal mechanism in ECG-FM rather than an artifact of attribution or an arbitrary sparse subset of units. To examine this, we study the identified neurons from three complementary perspectives under the 10% label setting on PTB-XL: causal intervention, ablation-based necessity, and run-to-run stability. The results are summarized in Figure 2.

Figure 2(a) shows that explicitly suppressing or amplifying the identified KNs induces clear and structured changes in the target-category logit, whereas applying the same inter-

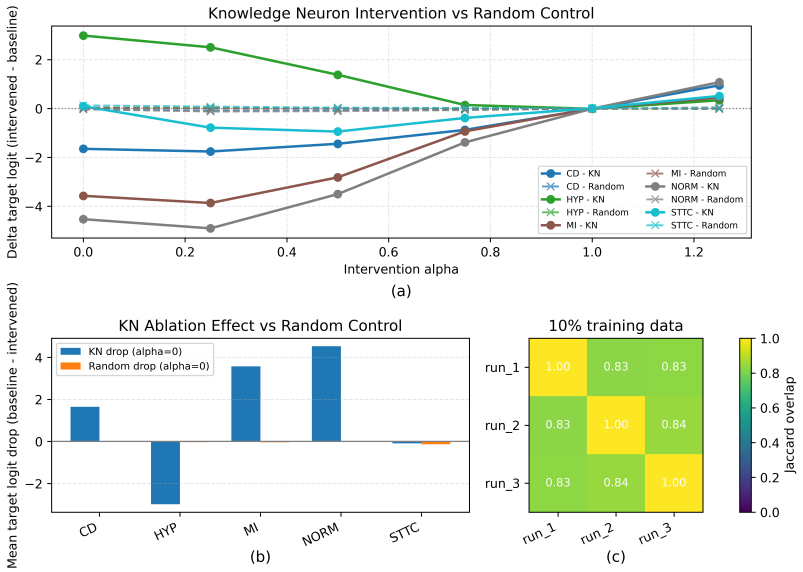


Figure 2: Mechanistic evidence for KNs in ECG-FM on PTB-XL under the 10% label setting.

vention to random control neurons yields effects close to zero. Figure 2(b) further shows that fully ablating the identified KNs causes substantially larger target-logit degradation than ablating random neurons, indicating that these units are not only able to influence prediction when manipulated, but are also important for preserving disease-relevant evidence during inference. Figure 2(c) shows that the recovered KN sets remain highly consistent across independent runs with different random seeds, with pairwise Jaccard overlap around 0.83–0.84. Taken together, these results support the view that ECG-FM contains a sparse subset of neurons with reproducible and causal influence on disease-specific prediction, which we refer to as KNs. Additional results under the 1% and 100% label settings are provided in Appendix B.1.

5.2. Results on ECG Classification

5.2.1. Benchmark Comparison Across Multiple ECG Datasets

We first evaluate SPARK across a broad range of benchmark settings to assess whether its classification performance remains consistent under varying datasets and supervision levels. The comparison includes six public ECG datasets and three label ratios (1%, 10%, and 100%), spanning different clinical sources, label granularities, and task difficulties. The full results are reported in Table 1. SPARK achieves the best performance in most settings, suggesting that the proposed knowledge-guided steering mechanism generalizes well across datasets and supervision levels. It consistently outperforms competing methods on PTBXL-Super, PTBXL-Form, PTBXL-Rhythm, CPSC2018, and most CSN settings; meanwhile, on PTBXL-Sub, it reaches the best results at 10% and 100% label ratios. This advantage is already visible in limited-label regimes, where SPARK performs best on several benchmarks at 1%, indicating strong label efficiency. As supervision increases, these gains remain stable rather than diminishing, which suggests that SPARK improves not only sample efficiency but also the final quality of the learned representations. The improvements are especially ev-

Table 1: Classification performance by Macro AUROC for SPARK and other eSSL methods. The best results are highlighted in **bold**, while the second-best results are indicated with *underlined*.

Method	PTBXL-Super			PTBXL-Sub			PTBXL-Form			PTBXL-Rhythm			CPSC2018			CSN		
	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
SimCLR	63.41	69.77	73.53	60.84	68.27	73.39	54.98	56.97	62.52	51.41	69.44	77.73	59.78	68.52	76.54	59.02	67.26	73.20
BYOL	71.70	73.83	76.45	57.16	67.44	71.64	48.73	61.63	70.82	41.99	74.40	77.17	60.88	74.42	78.75	54.20	71.92	74.69
BarlowTwins	72.87	75.96	78.41	*62.57*	70.84	74.34	52.12	60.39	66.14	50.12	73.54	77.62	55.12	72.75	78.39	**60.72**	71.64	77.43
MoCo-v3	73.19	76.65	78.26	55.88	69.21	76.69	50.32	63.71	71.31	51.38	71.66	74.33	*62.13*	76.74	75.29	54.61	*74.26*	77.68
SimSiam	73.15	72.70	75.63	62.52	69.31	76.38	55.16	62.91	71.31	49.30	69.47	75.92	58.35	72.89	75.31	58.25	68.61	77.41
TS-TCC	70.73	75.88	78.91	53.54	66.98	77.87	48.04	61.79	71.18	43.34	69.48	78.23	57.07	73.62	78.72	55.26	68.48	76.79
CLOCS	68.94	73.36	76.31	57.94	72.55	76.24	51.97	57.96	72.65	47.19	71.88	76.31	59.59	*77.78*	77.49	54.38	71.93	76.13
ASTCL	72.51	77.31	81.02	61.86	68.77	76.51	44.14	60.93	66.99	52.38	71.98	76.05	57.90	77.01	79.51	56.40	70.87	75.79
CRT	69.68	78.24	77.24	61.98	70.82	78.67	46.41	59.49	68.73	47.44	73.52	74.41	58.01	76.43	*82.03*	56.21	73.70	78.80
ST-MEM	61.12	66.87	71.36	54.12	57.86	63.59	55.71	59.99	66.07	51.12	65.44	74.85	56.69	63.32	70.39	59.77	66.87	71.36
HeartLang	*78.94*	*85.59*	*87.52*	**64.68**	*79.34*	*88.91*	*58.70*	*63.99*	*80.23*	*62.08*	*76.22*	*90.34*	60.44	66.26	77.87	57.94	68.93	*82.49*
SPARK	**80.23**	**86.53**	**89.38**	60.58	**82.10**	**90.30**	**58.86**	**68.55**	**83.84**	**67.50**	**78.20**	**91.53**	**67.51**	**85.97**	**89.92**	*59.97*	**79.52**	**90.92**

Table 2: Per-class performance of SPARK on MIMIC-IV-ECG. Each cell reports the SPARK score, with the subscript indicating the absolute change relative to ECG-FM on the same metric (↑: improvement; ↓: decrease). Rows shaded in light color denote rare disease categories discussed in the main text (Diagnostic labels with fewer than 50 positive cases are long-tail conditions).

Label	AUROC	AUPRC	Recall	FPR
Sinus rhythm	$0.969_{\uparrow 0.001}$	$0.989_{\downarrow 0.002}$	$0.930_{\downarrow 0.029}$	$0.096_{\downarrow 0.041}$
Premature ventricular contraction	$0.899_{\downarrow 0.009}$	$0.591_{\downarrow 0.019}$	$0.897_{\uparrow 0.176}$	$0.346_{\uparrow 0.275}$
Tachycardia	$0.967_{\downarrow 0.001}$	$0.911_{\uparrow 0.001}$	$0.913_{\uparrow 0.010}$	$0.060_{\uparrow 0.020}$
Ventricular tachycardia	$0.989_{\uparrow 0.041}$	$0.593_{\uparrow 0.407}$	$0.932_{\uparrow 0.518}$	$0.005_{\uparrow 0.003}$
Supraventricular tachycardia with aberrancy	$0.971_{\uparrow 0.010}$	$0.735_{\uparrow 0.062}$	$0.935_{\uparrow 0.250}$	$0.109_{\uparrow 0.091}$
Atrial fibrillation	$0.978_{\uparrow 0.002}$	$0.896_{\uparrow 0.013}$	$0.946_{\uparrow 0.049}$	$0.092_{\uparrow 0.054}$
Atrial flutter	$0.974_{\uparrow 0.043}$	$0.577_{\uparrow 0.106}$	$0.979_{\uparrow 0.399}$	$0.141_{\uparrow 0.127}$
Bradycardia	$0.961_{\downarrow 0.004}$	$0.843_{\downarrow 0.017}$	$0.897_{\uparrow 0.049}$	$0.065_{\uparrow 0.037}$
Accessory pathway conduction	$0.978_{\uparrow 0.006}$	$0.907_{\uparrow 0.019}$	$0.928_{\uparrow 0.059}$	$0.067_{\uparrow 0.037}$
Atrioventricular block	$0.973_{\uparrow 0.002}$	$0.828_{\downarrow 0.005}$	$0.906_{\uparrow 0.050}$	$0.069_{\uparrow 0.034}$
1st degree atrioventricular block	$0.980_{\uparrow 0.002}$	$0.848_{\uparrow 0.003}$	$0.924_{\uparrow 0.065}$	$0.060_{\uparrow 0.038}$
Bifascicular block	$0.992_{\uparrow 0.005}$	$0.848_{\uparrow 0.048}$	$0.987_{\uparrow 0.204}$	$0.076_{\uparrow 0.068}$
Right bundle branch block	$0.991_{\uparrow 0.004}$	$0.928_{\uparrow 0.019}$	$0.961_{\uparrow 0.082}$	$0.047_{\uparrow 0.030}$
Left bundle branch block	$0.989_{\uparrow 0.003}$	$0.863_{\uparrow 0.027}$	$0.972_{\uparrow 0.124}$	$0.071_{\uparrow 0.056}$
Infarction	$0.930_{\uparrow 0.002}$	$0.843_{\uparrow 0.023}$	$0.946_{\uparrow 0.134}$	$0.307_{\uparrow 0.186}$
Electronic pacemaker	$0.984_{\uparrow 0.008}$	$0.854_{\uparrow 0.042}$	$0.966_{\uparrow 0.103}$	$0.081_{\uparrow 0.069}$
Poor data quality	$0.833_{\uparrow 0.050}$	$0.252_{\uparrow 0.038}$	$0.975_{\uparrow 0.579}$	$0.741_{\uparrow 0.692}$

ident on more challenging settings such as PTBXL-Rhythm, CPSC2018, and CSN, pointing to a stronger ability to capture clinically meaningful structure across heterogeneous ECG datasets. The main exception arises on PTBXL-Sub at 1%, where HeartLang performs best, implying that large-scale pretrained representations may still retain an advantage under extremely limited supervision for finer-grained subclass prediction. Once modest task-specific supervision becomes available, however, SPARK surpasses HeartLang, further supporting the effectiveness of the proposed steering mechanism.

5.2.2. Per-Class Analysis on MIMIC-IV-ECG

Beyond aggregate performance, it is also important to examine whether the advantage of SPARK extends to clinically important diagnostic categories in ECG analysis, including underrepresented and relatively rare abnormalities. To this end, we conduct a per-class analysis on MIMIC-IV-ECG and compare SPARK with the downstream ECG-FM baseline using AUROC, AUPRC, and Recall for each label, as reported in Table 2. SPARK improves

one or more key metrics for most diagnostic categories, indicating stronger class-wise robustness on heterogeneous clinical data. The gains are especially pronounced for several clinically challenging and relatively rare abnormalities. For example, ventricular tachycardia improves by 0.041 AUROC, 0.407 AUPRC, and 0.518 Recall, while atrial flutter improves by 0.043, 0.106, and 0.399, respectively. Consistent gains are also observed for bifascicular block and left bundle branch block. Notably, these improvements are often more evident in AUPRC and Recall than in AUROC alone, suggesting that SPARK is particularly effective at improving positive-case retrieval under severe class imbalance. This is clinically important because rare and difficult abnormalities are often the categories for which sensitivity and precision-recall behavior matter most. Although a few categories, such as sinus rhythm, premature ventricular contraction, and bradycardia, show modest trade-offs in some metrics, these decreases remain small relative to the gains observed on more difficult categories. Taken together, the per-class results suggest that the benefit of SPARK extends beyond dominant classes and translates into a more clinically useful operating profile, particularly for underrepresented yet high-value diagnostic labels.

5.3. Results on Few-Shot Personalization

5.3.1. Overall Personalization Performance

Having established the advantage of SPARK in classification, we further evaluate its behavior in few-shot personalization. For this purpose, we conduct a controlled comparison of adaptation strategies on PTB-XL and MIMIC-IV-ECG using the same ECG-FM backbone, isolating the effects of backbone fine-tuning and knowledge-neuron-based personalization. The results are summarized in Figure 3, where the blue bars denote our KN-only adaptation strategy. SPARK remains consistently competitive with, and often favorable to, fine-tuning alone across both datasets. Its advantage also becomes more pronounced as K increases, suggesting that knowledge-neuron adaptation can effectively exploit additional patient-specific evidence. This pattern indicates that strong personalization does not necessarily require updating the full foundation model; instead, targeted adaptation on a sparse set of KNs already provides a lightweight and effective solution. Although the combined strategy achieves the strongest overall performance, the consistently strong results of the blue bars confirm the effectiveness of the proposed method on its own. Full numerical results corresponding to Figure 3 are provided in Appendix B.2.

5.3.2. Rare-Disease Analysis on MIMIC-IV-ECG

A more stringent test of few-shot personalization is whether the gains of SPARK also extend to underrepresented yet clinically important categories, rather than being driven mainly by common labels. To examine this, Table 3 reports the per-class ΔAUROC of knowledge-neuron adaptation over the baseline on several rare diseases in MIMIC-IV-ECG across different shot settings. The gains remain positive for all reported rare diseases at all values of K, indicating that the proposed adaptation mechanism consistently improves discrimination on underrepresented categories. Several classes show especially strong improvements. Bifascicular block reaches a particularly large gain of 0.3685 at $K = 5$, while left bundle branch block improves from 0.0700 at $K = 1$ to 0.2200 at $K = 5$, suggesting that the benefit of personalization becomes more pronounced as additional patient-specific records are available. Atrial flutter also remains consistently improved across all shot settings, with gains of 0.0829, 0.1813, and 0.1732 at $K = 1, 3, 5$, respectively. Even for supraventricular

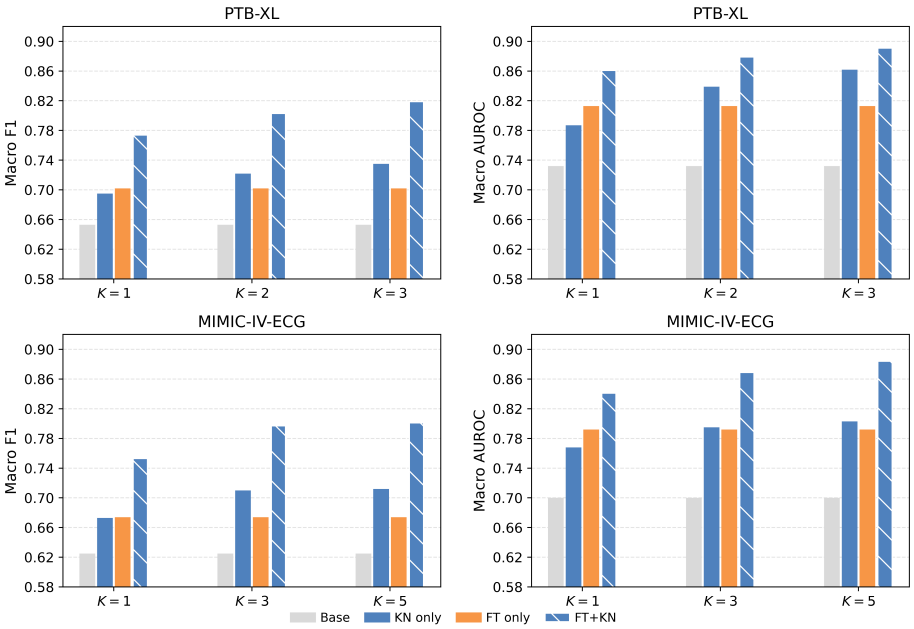


Figure 3: Few-shot personalization performance on PTB-XL and MIMIC-IV-ECG under four adaptation strategies. **Base**: frozen ECG-FM backbone with a trained classification head. **KN Only**: frozen Base model with M-PSH-based Knowledge Neuron modulation. **FT Only**: full-parameter fine-tuning of the backbone and classification head. **FT + KN**: full-parameter fine-tuning followed by M-PSH-based Knowledge Neuron modulation.

Table 3: Per-class ΔAUROC of knowledge neuron adaptation over the baseline on rare diseases under few-shot personalization on MIMIC-IV-ECG.

Rare Disease	ΔAUROC		
	$K = 1$	$K = 3$	$K = 5$
Bifascicular block	0.1270	0.0575	0.3685
Left bundle branch block	0.0700	0.0970	0.2200
Atrial flutter	0.0829	0.1813	0.1732
Electronic pacemaker	0.2024	0.0268	0.1444
Supraventricular tachycardia with aberrancy	0.0729	0.0906	0.0938
Poor data quality	0.3683	0.4148	0.2200
Average	0.1539	0.1447	0.2033

tachycardia with aberrancy, the gains remain stable across all three regimes. Averaged over the reported rare diseases, the improvement is 0.1539 at $K = 1$, 0.1447 at $K = 3$, and 0.2033 at $K = 5$, indicating that SPARK remains effective under low-shot personalization while continuing to benefit from richer patient-specific evidence.

6. Ablation Studies

6.1. Effective KNs Adaptation Without Full Backbone Fine-Tuning

This ablation study disentangles the effect of KN adaptation from that of conventional backbone fine-tuning. Figure 4 shows a controlled comparison across six ECG benchmarks

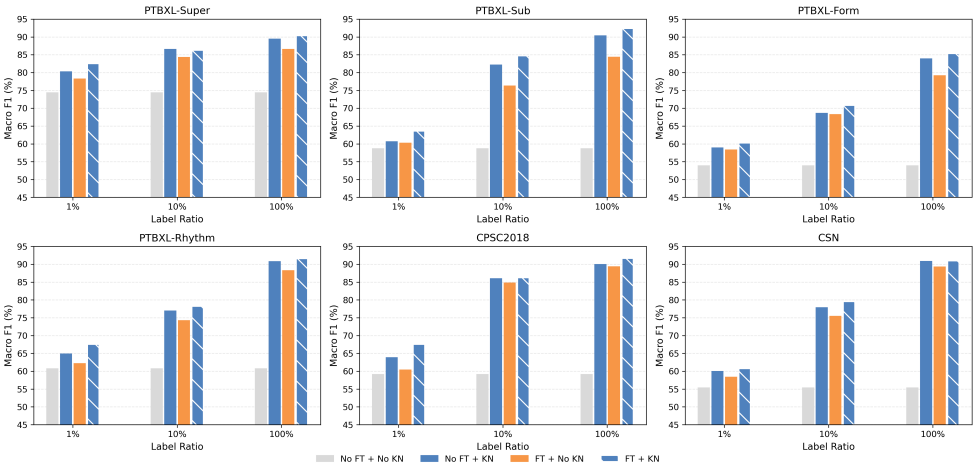


Figure 4: Controlled comparison of adaptation strategies under different label ratios across multiple ECG benchmarks.

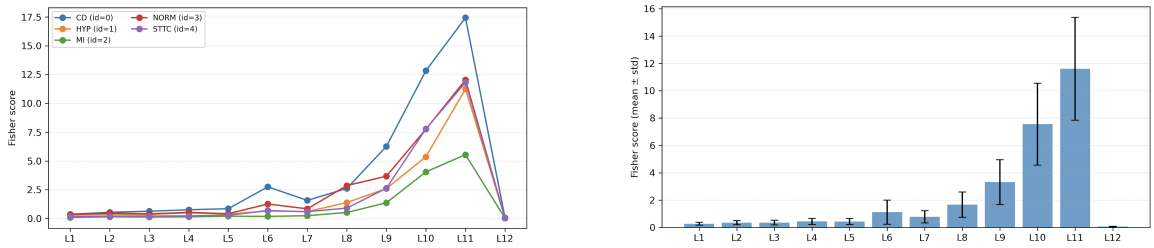
Table 4: Comparison of fine-tuning strategies and KN-only adaptation for few-shot personalization on MIMIC-IV-ECG. Best results are in **bold**; second-best are *underlined*.

Adaptation Strategy	$K = 1$		$K = 3$		$K = 5$	
	Macro F1	Macro AUROC	Macro F1	Macro AUROC	Macro F1	Macro AUROC
No Adaptation	0.624	0.699	0.624	0.699	0.624	0.699
LoRA Fine-Tuning	0.662	0.779	0.662	0.779	0.662	0.779
Supervised Fine-Tuning	**0.673**	**0.791**	*0.673*	*0.791*	*0.673*	*0.791*
KN-only (Ours)	*0.672*	*0.767*	**0.709**	**0.794**	**0.711**	**0.802**

under 1%, 10%, and 100% label ratios, varying only whether the ECG-FM backbone is fine-tuned and whether KN adaptation is enabled. Across datasets and supervision levels, enabling KN adaptation consistently improves the corresponding non-KN variants, indicating that the gains of SPARK are not simply due to generic parameter updating. Although combining backbone fine-tuning with KN adaptation (FT+KN) usually yields the best overall performance, the KN-only setting without backbone fine-tuning already performs remarkably well: it consistently surpasses the frozen baseline and remains highly competitive with, and sometimes favorable to, standard fine-tuning alone, while avoiding the cost of updating the full ECG-FM backbone. The same trend appears in the controlled comparison on MIMIC-IV-ECG in Table 4, where KN-only adaptation remains competitive with standard fine-tuning strategies and exceeds LoRA and supervised fine-tuning once more support records are available. Taken together, these results suggest that the effectiveness of SPARK comes primarily from selectively adapting task-relevant knowledge rather than from full-model updating alone. Detailed numerical results corresponding to Figure 4 are provided in Appendix B.3.

6.2. Layer-Wise Analysis of Knowledge Circuit Localization

Having shown that KN-based adaptation remains effective without full backbone fine-tuning, we next examine where the most informative knowledge circuits are located within ECG-FM. To this end, we analyze the layer-wise distribution of Fisher discriminative scores



(a) (b)

Figure 5: Layer-wise Fisher-score analysis on PTBXL_Super. (a) Per-label Fisher scores across layers. (b) Mean Fisher score across labels for each layer, with standard deviation shown as error bars. Higher values indicate stronger discriminative structure.

on PTBXL_Super and summarize the results in Figure 5. Both the per-label analysis in Figure 5(a) and the layer-wise summary in Figure 5(b) show a consistent pattern: discriminative structure is concentrated in the deeper layers rather than uniformly distributed across the backbone. Notably, the Fisher score does not increase monotonically to the final layer, but instead peaks in the upper layers and then shows a mild drop at the top, suggesting that the final layer may be more strongly aligned with output-facing representations where information becomes more compressed or entangled. These results motivate our knowledge-circuit design: targeted adaptation should focus on the most informative deep layers rather than treating all layers as equally relevant or assuming that the final layer is always optimal. The variation across labels further suggests that the structure of the knowledge circuit is disease-dependent, supporting the need for label-specific circuit identification. Additional ablation and robustness experiments are provided in Appendix C.

6.3. Efficiency Analysis of KN-Based Adaptation

The practical appeal of KN-based adaptation depends not only on its effectiveness, but also on whether it can reduce the computational burden of personalization relative to full backbone updating. Table 5 therefore compares the training and inference cost of four adaptation settings: no adaptation, KN-only adaptation, full fine-tuning, and full fine-tuning combined with KN adaptation. A clear efficiency gap appears during training. Full fine-tuning of ECG-FM requires 1.13×10^{17} FLOPs, whereas KN-only adaptation requires only 1.13×10^{15} FLOPs, corresponding to roughly a $100\times$ reduction in training cost. Even when KN adaptation is added on top of full fine-tuning, the extra KN cost remains only 1.13×10^{15} FLOPs, which is small relative to the full fine-tuning budget. This shows that the dominant computational burden comes from updating the entire backbone rather than from the proposed KN-based mechanism itself. The same conclusion holds at inference time: KN adaptation adds only 7.67×10^{6} extra FLOPs, corresponding to an overhead of just 0.00691, while the settings without KN adaptation incur no additional cost. Taken together, these results show that SPARK preserves negligible inference overhead while substantially reducing training cost, making KN-only adaptation a practical lightweight alternative to full-backbone fine-tuning.

Table 5: Computational cost of different adaptation settings during training and inference.

Fine-Tuning	KN Adaptation	Training		Inference	
		FT Cost	KN Cost	Extra FLOPs	Overhead
✗	✗	0	0	0	0
✗	✓	0	1.13×10^{15}	7.67×10^{6}	0.00691
✓	✗	1.13×10^{17}	0	0	0
✓	✓	1.13×10^{17}	1.13×10^{15}	7.67×10^{6}	0.00691

7. Conclusion

We presented SPARK, a lightweight adaptation framework for ECG foundation models that combines disease-specific sensitive subspace identification with targeted knowledge neuron adaptation. Across multiple public ECG benchmarks, SPARK consistently improved dataset-level classification performance under different label ratios, demonstrating robustness across heterogeneous clinical sources and label granularities. On MIMIC-IV-ECG, the method further improved per-class AUROC, AUPRC, and Recall for most diagnostic categories, with especially meaningful gains on several clinically challenging and relatively rare abnormalities. Under few-shot personalization, SPARK showed that updating only the identified KNs can already outperform standard fine-tuning alone in several settings, while the combination of fine-tuning and knowledge neuron adaptation yielded the strongest overall performance. Finally, computational analysis showed that KN-only adaptation remains substantially more cost beneficial than full backbone fine-tuning during training while introducing negligible inference overhead. Taken together, these results suggest that targeted internal adaptation of a frozen ECG foundation model offers a practical and effective path toward personalized ECG modeling in healthcare.

8. Discussion

This study proposes the SPARK framework, opening a new pathway for the lightweight adaptation of healthcare foundation models by "steering" rather than "retraining" physiological knowledge circuits. The experimental results validate the method's favorable performance in extreme few-shot scenarios and its robustness in handling long-tail clinical data. Below, we discuss the profound implications of this research from both technical innovation and clinical application perspectives.

▶ **Technical Implications: From Parameter Updating to Internal Representation Steering** Traditional adaptation paradigms (such as full fine-tuning or LoRA) essentially attempt to learn new knowledge by altering the model's "weights". However, this study demonstrates that medical foundation models already contain rich physiological knowledge internally; this knowledge simply fails to be properly activated when encountering specific individuals or rare diseases. **I) The Superiority of Mechanistic Adaptation**: The core contribution of SPARK lies in the discovery and utilization of "Physiological Knowledge Circuits". By generating pseudo-gradients through the M-PSH module to "steer" hidden layer activations, we achieve a finer-grained intervention than standard PEFT methods. This "surgical" adjustment avoids the catastrophic forgetting caused by global gradient descent, explaining why SPARK achieves a significant improve-

ment of approximately 5% in the 1-shot setting—it steers the model to "recall" existing relevant representations rather than learning them from scratch. **II) Decoupling Computation and Storage**: From a technical architecture perspective, SPARK achieves a deep decoupling of the foundation model backbone from user-specific features. Each user only needs to maintain extremely lightweight M-PSH parameters (steering vectors), completely eliminating the need to store independent model copies for thousands of patients. This architecture provides a highly sustainable technical solution for ultra-large-scale personalized healthcare systems.

▶ **Clinical Implications: Towards Precise and Robust Edge Healthcare** The success of SPARK represents more than just an algorithmic performance enhancement; it specifically addresses core bottlenecks in the real-world deployment of healthcare AI. **I) Breakthroughs in Rare Diseases and Long-Tail Distributions**: In clinical practice, rare diseases and atypical cases are often marginalized by "one-size-fits-all" models due to inherent data scarcity. By precisely pinpointing and amplifying "Value Neurons" associated with specific pathological features, SPARK significantly improves the model's capability to identify long-tail samples. This implies that when facing atypical patients, clinical AI systems will no longer merely default to "average performance" but can provide much more precise, personalized diagnostic insights. **II) Privacy-Preserving Edge Personalization**: The highly sensitive nature of clinical data dictates that processing should occur on edge devices (such as wearable ECG monitors) whenever possible. The extremely low computational overhead of SPARK makes real-time, dynamic patient adaptation feasible even on power-constrained embedded devices. Because the adaptation process only involves the lightweight tuning of internal activations—without requiring the upload of sensitive raw physiological data to the cloud for retraining—it inherently aligns with strict medical privacy protection requirements.

Limitations While SPARK demonstrates significant improvements in the lightweight adaptation of medical time-series, several limitations must be acknowledged. First, because our approach fundamentally relies on "steering" pre-existing internal representations, **it may not apply** to clinical scenarios involving entirely novel diseases or extreme physiological anomalies that were completely absent from the foundation model's pre-training corpus. If a specific pathological pattern was never learned during pre-training, the requisite "knowledge circuit" simply does not exist to be steered, necessitating traditional fine-tuning to acquire the new knowledge. Second, regarding aspects **we could not evaluate**, our empirical validation was strictly constrained to Transformer-based ECG foundation models. We could not check whether similar sparse, multi-layer knowledge circuits naturally emerge in alternative architectures that are increasingly used for time-series forecasting, such as State Space Models (e.g., Mamba Gu and Dao (2023)). Furthermore, while we proved the efficacy of few-shot personalization over standard episodic windows, the lack of multi-year, continuous patient datasets meant we could not check the long-term longitudinal stability of these personalized steering vectors—specifically, how M-PSH parameters should evolve as a patient's baseline physiology organically drifts over years of aging or chronic disease progression. Finally, the structural topology of knowledge circuits across different or multi-modal physiological signals (e.g., continuous EEG or fused ECG-PPG data) remains unverified and warrants future investigation.

References

Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Somaraju Boda, Manjunatha Mahadevappa, and Pranab Kumar Dutta. An automated patient-specific ecg beat classification using lstm-based recurrent neural networks. *Biomedical Signal Processing and Control*, 84:104756, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, 2022.

Cheng Ding, Tianliang Yao, Chenwei Wu, and Jianyuan Ni. Deep learning for personalized electrocardiogram diagnosis: A review. *arXiv preprint arXiv:2409.07975*, 2024.

Cheng Ding, Tianliang Yao, Chenwei Wu, and Jianyuan Ni. Advances in deep learning for personalized ecg diagnostics: A systematic review addressing inter-patient variability and generalization constraints. *Biosensors and Bioelectronics*, 271:117073, 2025.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

Anurag Garikipati, Jenish Maharjan, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Qingqing Mao, and Ritankar Das. Openmedlm: Prompt engineering can out-perform fine-tuning in medical question-answering with open-source

large language models. In *AAAI 2024 spring symposium on clinical foundation models*, 2024.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *PhysioNet*, September 2023. Version 1.0.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Morghan Hartmann, Umair Sajid Hashmi, and Ali Imran. Edge computing in smart health care systems: Review, challenges, and research directions. *Transactions on Emerging Telecommunications Technologies*, 33(3):e3710, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Yaojun Hu, Jintai Chen, Lianting Hu, Dantong Li, Jiahuan Yan, Haochao Ying, Huiying Liang, and Jian Wu. Personalized heart disease detection via ecg digital twin generation. *arXiv preprint arXiv:2404.11171*, 2024.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ecg words and sentences via pre-training ecg language model. *arXiv preprint arXiv:2502.10707*, 2025.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in neural information processing systems*, 34:1022–1035, 2021.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space. *Time, and Patients. arXiv*, 2005, 2021.

Stefan Konigorski, Johannes E Vedder, Babajide Alamu Owoyele, and İbrahim Özkan. Personalization of large foundation models for health interventions. *arXiv preprint arXiv:2601.03482*, 2026.

Chenqian Le, Ziheng Gong, Chihang Wang, Haowei Ni, Panfeng Li, and Xupeng Chen. Instruction tuning and cot prompting for contextual medical qa with llms. In *2025 International Conference on Artificial Intelligence, Human-Computer Interaction and Natural Language Processing (ICAHN)*, pages 43–46. IEEE, 2025.

Simon A Lee, Cyrus Tanade, Hao Zhou, Juhyeon Lee, Megha Thukral, Baiying Lu, and Sharanya Arcot Desai. Towards on-device foundation models for raw wearable signals. In *NeurIPS 2025 Workshop on Learning from Time Series for Health*, 2025a.

Thomas L Lee, William Toner, Rajkarn Singh, Artjom Joosen, and Martin Asenov. Lightweight online adaption for time series foundation model forecasts. *arXiv preprint arXiv:2502.12920*, 2025b.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

Jialin Liu, Fang Liu, Changyu Wang, and Siru Liu. Prompt engineering in clinical practice: tutorial for clinicians. *Journal of Medical Internet Research*, 27:e72644, 2025.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024a.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-xl: Long-context transformers for unified time series forecasting. *arXiv preprint arXiv:2410.04803*, 2024b.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *Jamia Open*, 8(5):ooaf122, 2025.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450*, 2024.

Haowen Pan, Xiaozhi Wang, Yixin Cao, Zenglin Shi, Xun Yang, Juanzi Li, and Meng Wang. Precise localization of memories: A fine-grained neuron-level knowledge editing technique for llms. *arXiv preprint arXiv:2503.01090*, 2025.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. *arXiv preprint arXiv:2410.20542*, 2024.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE journal of biomedical and health informatics*, 25(5):1519–1528, 2020.

Jingyuan Sun, Riza Theresa Batista-Navarro, Hao Li, Viktor Schlegel, and Yizheng Sun. Mira: Medical time series foundation model for real-world health data. In *Proceedings of Neural Information Processing Systems*. Neural information processing systems foundation, 2025.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.

Ning Wang, Panpan Feng, Zhaoyang Ge, Yanjie Zhou, Bing Zhou, and Zongmin Wang. Adversarial spatiotemporal contrastive learning for electrocardiogram signals. *IEEE transactions on neural networks and learning systems*, 35(10):13845–13859, 2023.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *Advances in Neural Information Processing Systems*, 37:53764–53797, 2024.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, 2022.

Jinning Yang, Wenjie Sun, and Wen Shi. Heartllm: Discretized ecg tokenization for llm-based diagnostic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34250–34258, 2026.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. *Advances in neural information processing systems*, 37:118571–118602, 2024.

Zeping Yu and Sophia Ananiadou. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280, 2024.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

Wenrui Zhang, Ling Yang, Shijia Geng, and Shenda Hong. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16129–16138, 2023.

Jianwei Zheng, Huimin Chu, Daniele Struppa, Jianming Zhang, Sir Magdi Yacoub, Hesham El-Askary, Anthony Chang, Louis Ehwerhemuepha, Islam Abudayyeh, Alexander Barrett, et al. Optimal multi-stage arrhythmia classification approach. *Scientific reports*, 10(1): 2898, 2020.

Appendix A. Experimental Details

A.1. Detailed Dataset Description

To comprehensively evaluate the proposed method, we conduct experiments on multiple publicly available 12-lead ECG benchmarks that exhibit substantial diversity in clinical sources, label granularity, and task complexity. Specifically, our evaluation includes PTB-XL (Wagner et al. (2020)), CPSC2018 (Liu et al. (2018)), Chapman-Shaoxing-Ningbo (CSN) (Zheng et al. (2020)), and MIMIC-IV-ECG (Gow et al. (2023)). Collectively, these datasets span both coarse- and fine-grained diagnostic taxonomies, encompass relatively controlled benchmark recordings as well as highly heterogeneous real-world hospital data, and present varying levels of classification difficulty.

PTB-XL. PTB-XL is a large-scale open clinical 12-lead ECG benchmark containing 21,837 recordings from 18,885 patients. All signals are 10 seconds long and sampled at 500 Hz. Following the SCP-ECG annotation system, the dataset supports several label granularities, including **Superclass** (5 classes), **Subclass** (23 classes), **Form** (19 classes), and **Rhythm** (12 classes). In our experiments, we adopt the officially recommended split protocol when applicable Strodthoff et al. (2020).

CPSC2018. The CPSC2018 dataset is a public 12-lead ECG benchmark released for the China Physiological Signal Challenge, consisting of 6,877 recordings annotated with 9 rhythm categories. Each ECG recording is sampled at 500 Hz, with lengths varying from 6 to 60 seconds.

Chapman-Shaoxing-Ningbo (CSN). The Chapman-Shaoxing-Ningbo dataset is a large multi-center 12-lead ECG resource with 45,152 recordings acquired at 500 Hz, each with a duration of 10 seconds. Following common preprocessing practice, we exclude samples associated with "unknown" annotations. After filtering, 23,026 recordings with 38 diagnostic labels are retained.

MIMIC-IV-ECG. MIMIC-IV-ECG is a large-scale public clinical 12-lead ECG database derived from routine care at Beth Israel Deaconess Medical Center and linked to the MIMIC-IV clinical database. It contains approximately 800,000 diagnostic ECG recordings from nearly 160,000 unique patients, with each study consisting of a standard 10-second waveform sampled at 500 Hz. Compared with relatively controlled benchmark datasets, MIMIC-IV-ECG exhibits substantially greater clinical heterogeneity, spanning emergency, inpatient, intensive care, and outpatient settings.

A.2. Detailed Data Split Statistics

As the two tasks serve different objectives, they differ in both the datasets used and the strategies for partitioning the training, validation, and test sets.

Classification task split. For the ECG classification task, we evaluate on four public datasets: PTB-XL, CPSC2018, Chapman-Shaoxing-Ningbo (CSN), and MIMIC-IV-ECG. To ensure fair comparison and consistent experimental settings, all classification experiments use a unified 70%:10%:20% split for training, validation, and test, respectively. Detailed split statistics are reported in Table 6.

Personalization task split. For the personalization task, the goal is to evaluate patient-specific adaptation from a limited amount of individual historical data. This requires datasets in which each patient has multiple ECG recordings. Among the datasets considered in this study, only PTB-XL and MIMIC-IV-ECG exhibit clear multi-record-per-

Table 6: Statistics of datasets and data splits for the classification task.

Dataset	#Categories	Total	Train	Valid	Test
PTB-XL-Super	5	21,388	17,084	2,146	2,158
PTB-XL-Sub	23	21,388	17,084	2,146	2,158
PTB-XL-Form	19	8,978	7,197	901	880
PTB-XL-Rhythm	12	21,030	16,832	2,100	2,098
CPSC2018	9	6,877	4,950	551	1,376
CSN	38	23,026	16,546	1,860	4,620
MIMIC-IV-ECG	17	800,035	568,914	71,114	160,007

Table 7: Patient-level split statistics for the personalization task.

Dataset	#Patients ($\geq$ 5 ECGs)	Total	Train	Valid	Test
PTB-XL	72	21,837	17,142	4,280	415
MIMIC-IV-ECG	46,012	800,035	176,211	47,771	576,053

patient characteristics; accordingly, personalization experiments are conducted on these two datasets only.

To avoid information leakage, data partitioning is performed at the patient level rather than the individual recording level, ensuring that recordings from the same patient do not appear in both the training and test sets. We also use a patient-and-time-aware protocol: all ECGs from the same patient are assigned to only one split to avoid patient leakage, and within each patient, the earliest K ECGs are used as support context while later ECGs are used for evaluation. Specifically, patients with at least five ECG recordings are assigned to the test set, while patients with fewer than five recordings are split into training and validation sets with an 80%:20% ratio and are used for global model training, model selection, and related module learning. This evaluation protocol is designed to assess few-shot adaptation on previously unseen patients with sufficient within-patient history, rather than to enforce record-level balance across splits. As a result, the personalization test cohort in MIMIC-IV-ECG naturally contains more ECG records than the corresponding training cohort, since patients with richer longitudinal histories contribute substantially more recordings on average. This record-level imbalance is therefore an inherent property of the task construction and reflects the intended deployment setting of few-shot personalization. Detailed split statistics are reported in Table 7.

A.3. Implementation Details

Unless otherwise specified, all experiments are conducted with a fixed global random seed. For the label-ratio experiments (1%, 10%, and 100%), the reduced training subsets are obtained by random subsampling from the corresponding training split under the same seed, while validation and test splits remain unchanged.

For few-shot personalization, let $K \in \{1, 3, 5\}$ denote the number of historical ECG records used for patient-specific KN updating. For each patient in the personalization test cohort, we collect all original ECG records for that patient, shuffle them using a fixed random

Table 8: Hyperparameters for classification and personalization.

Hyperparameters	**Values**
optimizer	Adam
Learning rate	$1e-3$
Weight decay	$1e-4$
dropout	0.3
Embedding batch size	1,024
Raw ECG batch size	512
KN identification batch size	32
LR scheduler	Cosine
Total epochs	200
Early stopping	5

seed, and select K records for adaptation; the remaining records are used for personalized evaluation. To prevent information leakage, data partitioning is performed at the patient level: all ECG records from the same patient are assigned to exactly one of the train, validation, or test splits and never appear across multiple splits.

Table 8 summarizes the key hyperparameters used in downstream classification and few-shot personalization.

Appendix B. Additional Experimental Results

B.1. Mechanistic Evidence Across Label Ratios

For completeness, we additionally report the intervention, ablation, and stability analyses under the 1% and 100% label settings in Figures 6, complementing the 10% results shown in the main text. Across all three label regimes, the same overall pattern remains consistent: intervening on identified KNs induces substantially larger target-logit shifts than random controls, ablating KNs causes markedly stronger degradation than random neuron removal, and the recovered KN sets show substantial run-to-run overlap. These results suggest that the existence of sparse and functionally important KNs is not specific to a single label regime.

At the same time, the strength and stability of the evidence vary somewhat with label availability. Under the 1% setting, the intervention and ablation effects remain clear but are somewhat less uniform across labels, and the KN sets are slightly less stable across repeated runs, indicating that KN identification remains meaningful even under extreme supervision scarcity, albeit with weaker reproducibility. Under the 100% setting, the intervention and ablation patterns remain strong, while the recovered KN sets continue to exhibit substantial overlap across runs. Compared with the 10% regime, however, the exact KN subsets show slightly greater variability, suggesting that stronger supervision sharpens functional effects but does not necessarily force identical sparse solutions across repeated runs. Overall, these results reinforce the main conclusion that disease-relevant computation in ECG-FM is at least partially localized to a sparse subset of neurons across a wide range of supervision levels.

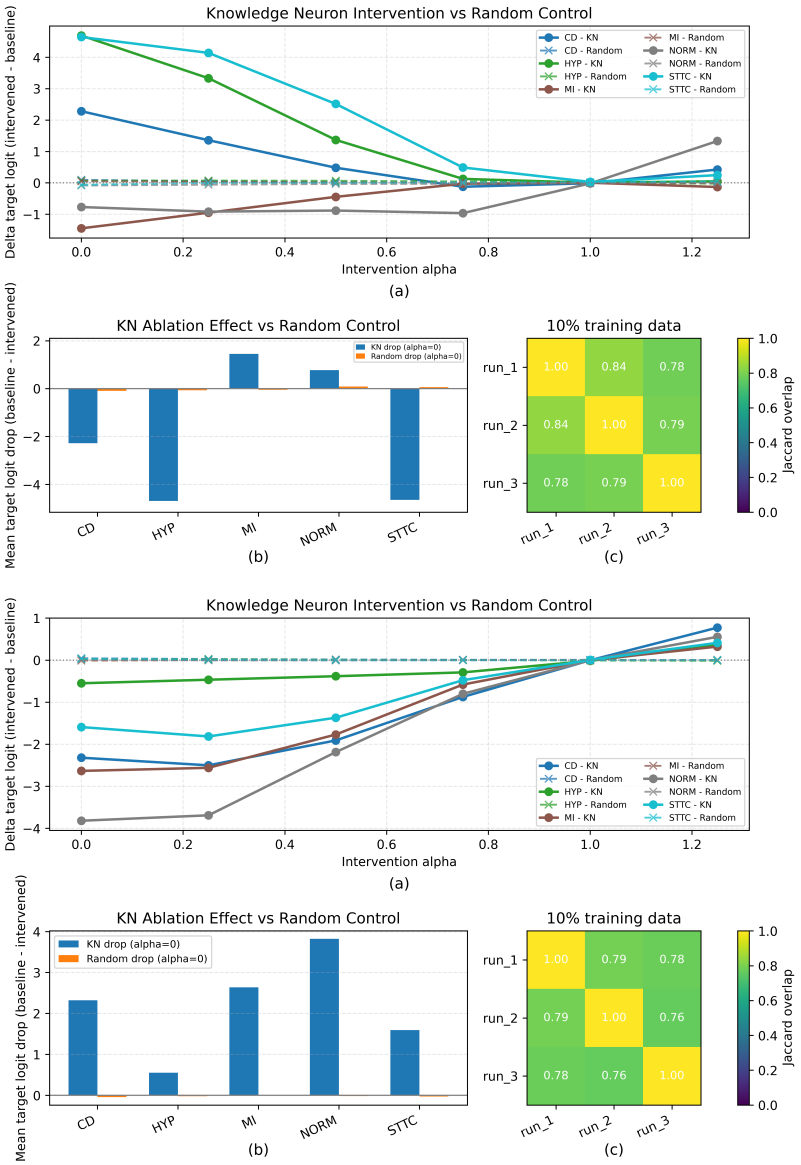


Figure 6: Mechanistic evidence for KNs in ECG-FM on PTB-XL under the 1% (top) and 100% (bottom) label settings.

B.2. Detailed Personalization Results

To complement the visual comparison in Figure 3, we report the full numerical results in Tables 9 and 10. These tables list the exact Macro F1 and Macro AUROC values for all adaptation strategies under different few-shot settings on PTB-XL and MIMIC-IV-ECG, enabling a more detailed comparison of the underlying performance differences.

Table 9: Detailed numerical results for the controlled comparison of adaptation strategies on PTB-XL under different few-shot settings.

Fine-Tuning	KN Adaptation	$K = 1$		$K = 2$		$K = 3$	
		Macro F1	Macro AUROC	Macro F1	Macro AUROC	Macro F1	Macro AUROC
✗	✗	0.652	0.731	0.652	0.731	0.652	0.731
✗	✓	0.694	0.786	0.721	0.838	0.734	0.861
✓	✗	0.701	0.812	0.701	0.812	0.701	0.812
✓	✓	0.774	0.861	0.803	0.879	0.819	0.891
Δ over No Adaptation (KN-only)		+0.042	+0.055	+0.069	+0.107	+0.082	+0.130
Δ over No Adaptation (FT+KN)		+0.122	+0.130	+0.151	+0.148	+0.167	+0.160

Table 10: Detailed numerical results for the controlled comparison of adaptation strategies on MIMIC-IV-ECG under different few-shot settings.

Fine-Tuning	KN Adaptation	$K = 1$		$K = 3$		$K = 5$	
		Macro F1	Macro AUROC	Macro F1	Macro AUROC	Macro F1	Macro AUROC
✗	✗	0.624	0.699	0.624	0.699	0.624	0.699
✗	✓	0.672	0.767	0.709	0.794	0.711	0.802
✓	✗	0.673	0.791	0.673	0.791	0.673	0.791
✓	✓	0.753	0.841	0.797	0.869	0.801	0.884
Δ over No Adaptation (KN-only)		+0.048	+0.068	+0.085	+0.095	+0.087	+0.103
Δ over No Adaptation (FT+KN)		+0.129	+0.142	+0.173	+0.170	+0.177	+0.185

B.3. Detailed Results for Label-Ratio Ablation

Table 11 reports the full numerical results corresponding to Figure 4. It lists the exact performance values for all adaptation strategies under 1%, 10%, and 100% label ratios across the six ECG benchmarks, complementing the visual trends shown in the main text.

Appendix C. Extended Ablation Studies

C.1. Stability Across Random Support-Set Sampling

Since KN-only personalization explicitly depends on the support ECGs selected for each target patient, we further evaluate its robustness to support-set sampling under multiple random seeds. We focus on the KN-only setting because it is the adaptation strategy that directly reflects the stability of the proposed knowledge-neuron personalization mechanism under few-shot adaptation.

For both PTB-XL and MIMIC-IV-ECG, we keep the patient-level train/validation/test partition fixed and independently reconstruct the patient-specific support/query episodes for each random seed. For each shot setting, we report the mean and standard deviation over five runs.

As shown in Tables 12, KN-only adaptation remains stable under repeated support-set resampling on both datasets, with relatively small standard deviations in both Macro F1 and Macro AUROC. The performance trends across different shot settings are also consistent with the main personalization results, indicating that the gains of KN-only personalization

Table 11: Detailed numerical results for the controlled comparison of adaptation strategies under different label ratios across multiple ECG benchmarks.

Fine-Tuning	KN Adaptation	PTBXL-Super			PTBXL-Sub			PTBXL-Form			PTBXL-Rhythm			CPSC2018			CSN		
		1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
✗	✗	74.31	74.31	74.31	58.66	58.66	58.66	53.87	53.87	53.87	60.67	60.67	60.67	59.06	59.06	59.06	55.32	55.32	55.32
✗	✓	80.23	86.53	89.38	60.58	82.10	90.30	58.86	68.55	83.84	64.85	76.90	90.75	63.80	85.97	89.92	59.97	77.80	90.80
✓	✗	78.23	84.25	86.52	60.23	76.23	84.32	58.32	68.22	79.12	62.12	74.21	88.20	60.34	84.76	89.30	58.32	75.42	89.23
✓	✓	82.45	86.23	90.34	63.56	84.67	92.35	60.23	70.76	85.32	67.50	78.20	91.53	67.51	86.20	91.60	60.70	79.52	90.92

Table 12: Stability of KN-only personalization under repeated random support-set sampling. Results are reported as mean ± std over 5 random seeds.

Dataset	$K^{(1)}$		$K^{(2)}$		$K^{(3)}$	
	Macro F1	Macro AUROC	Macro F1	Macro AUROC	Macro F1	Macro AUROC
PTB-XL	0.694 ± 0.010	0.786 ± 0.008	0.722 ± 0.008	0.839 ± 0.006	0.735 ± 0.006	0.862 ± 0.005
MIMIC-IV-ECG	0.672 ± 0.009	0.767 ± 0.007	0.708 ± 0.006	0.793 ± 0.005	0.712 ± 0.005	0.803 ± 0.004

Note: For PTB-XL, $K^{(1)}$=1, $K^{(2)}$=2, and $K^{(3)}$=3. For MIMIC-IV-ECG, $K^{(1)}$=1, $K^{(2)}$=3, and $K^{(3)}$=5.

are not tied to a particular favorable support-set realization, but instead reflect a robust advantage of targeted knowledge-neuron adaptation in the few-shot setting.

C.2. Ablation on the Number of Updated KNs

To test whether effective personalization requires updating a large number of internal parameters, we vary the budget of updated KNs while keeping all other settings fixed. Specifically, we evaluate KN-only personalization on MIMIC-IV-ECG under different KN budgets, where the budget is defined as the proportion of identified KNs enabled for adaptation. Results under $K \in \{1, 3, 5\}$ are shown in Figure 7.

As the KN budget increases from 1% to moderate values, performance improves consistently across all shot settings in both Macro F1 and Macro AUROC. The gains then gradually saturate beyond 10%–20%, with only marginal improvement at larger budgets. This trend suggests that the effectiveness of SPARK does not rely on broad parameter updates, but can already be achieved by adapting a relatively small subset of task-relevant KNs.

C.3. Sensitivity to KN Selection Strategy

To test whether the gains of SPARK arise from knowledge-guided neuron selection rather than from arbitrary sparse updates, we compare the proposed attribution-based KN selection strategy against several alternatives under a matched adaptation budget. Specifically, we consider random neuron selection, magnitude-based top-k, and a last-layer-only heuristic. Results on MIMIC-IV-ECG are shown in Figure 8.

The proposed KN selection strategy consistently outperforms all alternatives across different shot settings. In particular, random neuron selection under the same update budget yields substantially weaker performance, indicating that the gains of SPARK cannot be explained solely by sparse adaptation. The results further suggest that identifying task-relevant KNs is important for achieving effective personalization.

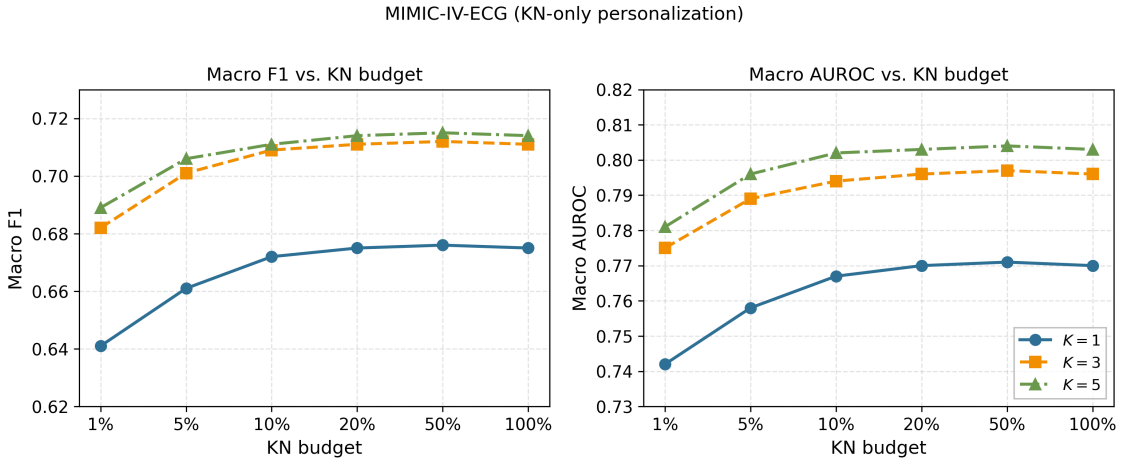


Figure 7: Effect of the KN update budget on KN-only personalization performance on MIMIC-IV-ECG. The budget is defined as the proportion of identified KNs enabled for adaptation. Left: Macro F1 under different shot settings. Right: Macro AUROC under different shot settings. Performance improves rapidly when moving from very small budgets to moderate budgets, and then gradually saturates, indicating that effective personalization can already be achieved by updating only a relatively small subset of identified KNs.

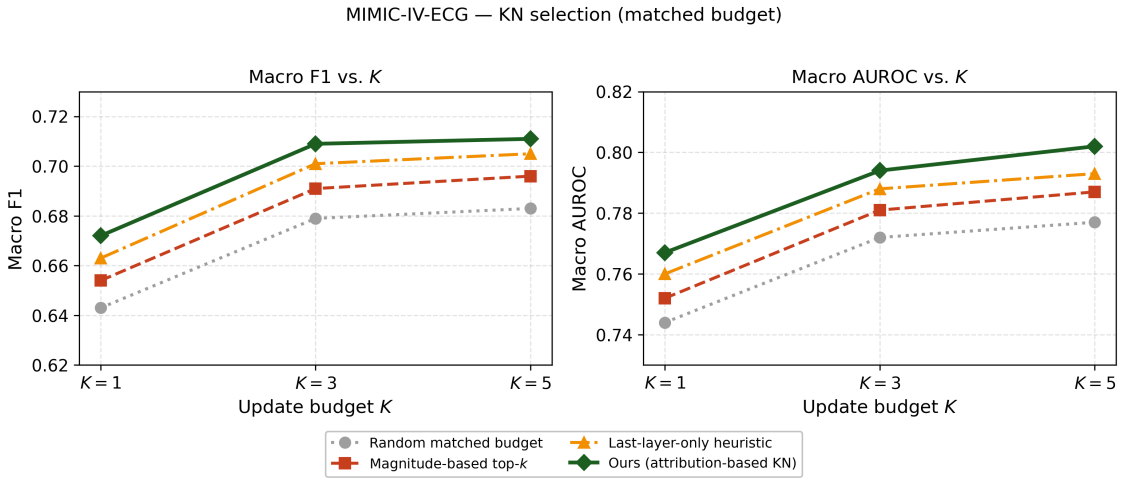


Figure 8: Sensitivity to KN selection strategy on MIMIC-IV-ECG under matched update budget. The proposed attribution-based KN selection consistently outperforms random matched-budget selection, magnitude-based top-k, and the last-layer-only heuristic across different shot settings in both Macro F1 and Macro AUROC. These results indicate that the gains of SPARK are not explained by arbitrary sparse updating alone, but depend on identifying clinically relevant KNs.