

CaReFit: Cost-Effective Cardiorespiratory Fitness Assessment Using Multimodal Sensors

CHI XU, The Hong Kong University of Science and Technology, Hong Kong SAR

WENTAO XIE, The Hong Kong University of Science and Technology, Hong Kong SAR

HUANING TAN, Southern University of Science and Technology, China

ZIRUI HUANG, The Hong Kong University of Science and Technology (Guangzhou), China

YANBIN GONG, The Hong Kong University of Science and Technology, Hong Kong SAR

JIN ZHANG, Southern University of Science and Technology, China

SHIFANG YANG, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), China

QIAN ZHANG, The Hong Kong University of Science and Technology, Hong Kong SAR

Cardiorespiratory fitness (CRF) is crucial for maintaining overall health, with $VO_2\text{max}$ (maximal oxygen consumption) and $VO_2\text{AT}$ (oxygen consumption at anaerobic threshold) serving as key physiological indicators. The current gold standard for assessing these metrics is Cardiopulmonary Exercise Testing (CPET), which requires bulky equipment, trained personnel, and significant time, putting it out of reach for most people. In this study, we introduce CaReFit, an innovative and low-cost system that utilizes a single-lead electrocardiogram (ECG) and muscle oxygenation (SmO_2) sensors to estimate $VO_2\text{max}$ and $VO_2\text{AT}$. Our method leverages the coupled behavior of ECG and SmO_2 signals during exercise to predict these critical performance metrics. To mitigate exercise-induced noise in ECG signals, we implement a diffusion-based ECG denoising model trained with a tailored ECG-guided loss function on combined public datasets. Then, our multimodal neural network is specifically designed to leverage physiological insights for accurate CRF predictions. We conducted a deployment of CaReFit with 50 patients in a clinical setting, and evaluated CaReFit under a strict subject-disjoint protocol. Under this leakage-free evaluation, CaReFit achieves a mean absolute percentage error (MAPE) of 9.30% (95% CI: 7.06–11.80) for $VO_2\text{max}$ and 9.49% (95% CI: 7.02–12.12) for $VO_2\text{AT}$. We further conduct a head-to-head comparison with a commercial portable device, and our system demonstrates similar $VO_2\text{max}$ performance trends under the same session protocol while additionally supporting $VO_2\text{AT}$ estimation. Furthermore, we make our dataset and evaluation code publicly available to facilitate further research and development in this domain.

CCS Concepts: • **Human-centered computing** → **Mobile devices**; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: cardiorespiratory fitness, ECG, muscle oxygenation, multimodal signals fusion

Authors' Contact Information: [Chi Xu](#), Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, cxubs@cse.ust.hk; [Wentao Xie](#), Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, wentaiox@cse.ust.hk; [Huaning Tan](#), National Graduate College for Engineers, Southern University of Science and Technology, Shenzhen, China, 12440067@mail.sustech.edu.cn; [Zirui Huang](#), Internet of Things Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, zhuang285@connect.hkust-gz.edu.cn; [Yanbin Gong](#), Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, yonggae@cse.ust.hk; [Jin Zhang](#), Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China, zhangj4@sustech.edu.cn; [Shifang Yang](#), Department of Pulmonary and Critical Care Medicine, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China, yangshifang@gdph.org.cn; [Qian Zhang](#), Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, qianzh@cse.ust.hk.



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

© 2026 Copyright held by the owner/author(s).

ACM 2474-9567/2026/9-ART176

<https://doi.org/10.1145/3831985>

ACM Reference Format:

Chi Xu, Wentao Xie, Huaning Tan, Zirui Huang, Yanbin Gong, Jin Zhang, Shifang Yang, and Qian Zhang. 2026. CaReFit: Cost-Effective Cardiorespiratory Fitness Assessment Using Multimodal Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 3, Article 176 (September 2026), 31 pages. <https://doi.org/10.1145/3831985>

1 INTRODUCTION

Cardiorespiratory fitness (CRF) refers to the ability of the cardiorespiratory system to supply oxygen to skeletal muscles during sustained physical activity, which summarizes the integrated performance of the heart, lungs, and circulatory system [26]. According to medical guidance, CRF indicates exercise endurance, and assessing CRF is critical for personalized rehabilitation prescription, preoperative assessment, and even disease diagnosis [33]. Due to its importance, the American Heart Association has designated CRF as the fifth vital sign, complementing body temperature, pulse rate, respiration rate, and blood pressure [37].

There are two important metrics to characterize CRF – VO_2max and VO_2AT [21, 37]. VO_2max represents the maximum rate at which an individual can consume oxygen during exhaustive exercise, serving as a quantitative indicator of "exercise tolerance". While VO_2AT is the oxygen consumption during the anaerobic threshold, providing the most direct evidence for rehabilitation prescriptions. Together, these two indicators reflect aerobic capacity and functional endurance, collectively describing the CRF.

The gold standard for measuring VO_2max and VO_2AT is Cardiopulmonary Exercise Testing (CPET) [2, 24], where a patient must wear a mask connected to a flowmeter and a gas analyzer to measure the ventilation, as well as 12-lead electrocardiogram (ECG) to monitor heart activity. During the test, the patient is asked to perform aerobic exercise on a treadmill or stationary bicycle with programmed incremental exercise load, where the ventilation (from the gas analyzer) and heart activity (from the ECG) data are monitored to provide VO_2max and VO_2AT (see Section 2). However, the clinical CPET system is very bulky and expensive (over USD 100,000), which largely limits its widespread use in community clinics and rehabilitation centers. In addition, traditional CPET requires a heavy sensor setup, including wearing a face mask and the 12-lead ECG as shown in Figure 1(a). Thus, a health professional must help the patient put on the sensors, which may take more than 30 minutes. These drawbacks of CPET impede its application and availability in the wild.

To solve these issues, previous work has sought to obtain VO_2max using more convenient methods. Decades ago, formulas were developed to predict oxygen uptake (VO_2) [14, 45], where the subject inputs a few demographic and performance statistics, such as heart rate, and oxygen uptake can be roughly computed. Nevertheless, as pointed out by previous research [34, 51], achieving peak oxygen uptake (VO_2max) is a complex physiological process and cannot be oversimplified by a single formula. Even when formulas are replaced with machine learning (ML) models [1], these methods still rely on static user metrics and, therefore, suffer from low accuracy and generalizability to other populations [51]. Recently, some studies have tried to measure real-time physiological signals during exercise to infer oxygen consumption. Shandhi *et al.* [39] propose a method to predict instantaneous oxygen uptake during exercise with wearable cardio-electromechanical and environmental sensors. However, such methods fail to achieve sufficient accuracy and cannot provide VO_2AT , which is also very important in evaluating CRF, especially for patients.

In this research, we propose CaReFit, a low-cost **CardioRespiratory Fitness** assessment system to solve the above dilemma – achieving both **accurate prediction** and **minimum setup overhead**, so that the CRF assessment is more approachable, such as in community fitness centers, and can be easily administered even by the user themselves. To achieve this objective, we aim to use a minimal set of sensors to measure CRF. We draw inspiration from medical research that there are two most important factors affecting oxygen uptake – cardio activity [39, 52] and muscle metabolic capacity [18, 38]. Therefore, we include only two types of sensors to measure the above physiological signals, ensuring the utmost hardware simplicity. First, we include the single-lead ECG sensor because it is the minimum sensor unit capable of recording cardiac activity. Second, for the muscle

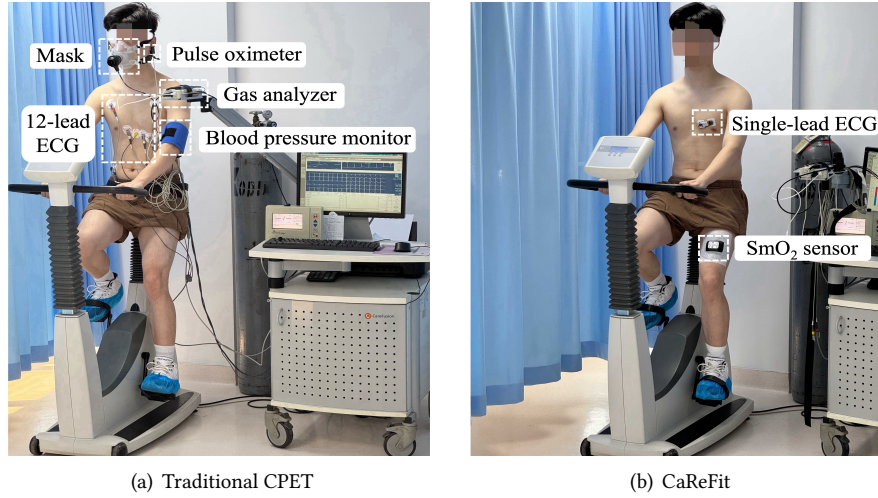


Fig. 1. Assessing CRF (a) Traditional CPET (b) CaReFit.

metabolic capacity, which refers to the ability of muscle tissue to produce energy, it is impossible to estimate it non-invasively. Even so, medical research has suggested that it is correlated with the muscle oxygen level (SmO₂) [20], which can be measured with non-invasive near-infrared spectroscopy. Hence, we include the SmO₂ sensor in the sensor set, together with the single-lead ECG sensor. Our setup is presented in Figure 1(b).

While conceptually feasible, no existing studies have established a model linking ECG and SmO₂ to oxygen uptake. Therefore, the first and biggest challenge is to predict CRF metrics - VO₂max and VO₂AT-from such a simple sensor combination. Specifically, since ECG and SmO₂ are two independent measurements with a natural and nondeterministic delay due to the delayed hemodynamic response [42], how to encode each measurement independently and dynamically fuse these two requires careful design. In addition to this challenge, as discussed in Section 2, a subject must perform cycling exercise during the CRF assessment, which will constantly introduce significant noise (including baseline wander, muscle artifact, as well as electrode motion artifact) to the ECG signals due to stretching of the skin and movement near the electrodes. These artifacts will significantly degrade ECG signal quality. However, unlike multi-lead ECG recordings, which provide spatially redundant information that can facilitate artifact detection and mitigation, a single-lead ECG provides no such cross-lead redundancy, making it impossible to eliminate noise independently. Also, the noise during a cycling exercise is opportunistic and subject-varied since different individuals have different cycling speed and intensity, making traditional filter-based denoising algorithms less effective. Thus, the second challenge faced by CaReFit is to effectively remove exercise-induced noise from the single-lead ECG signal. Notably, since we capture SmO₂ signals utilizing near-infrared reflectance spectroscopy, the influence of the above artifacts on SmO₂ can be negligible.

To address the first challenge, we carefully craft physiology-informed unimodal encoders for ECG and SmO₂, according to their distinct temporal dynamics and physiological roles. Then, we introduce a self-adapted fusion block with a cross-attention mechanism to fuse these two measurements. This fusion design adaptively adjusts the attention weights between the ECG and SmO₂ signals to ensure effective modality fusion, with an optimized loss function. Moreover, considering CRF is also partially associated with the subject's pulmonary function in addition to cardio and muscle activities, we design a pluggable text encoder to process the text-based Electronic Health Record (EHR) including pulmonary function test (PFT) report if it is available and fuse it with the ECG and SmO₂

representations to achieve a holistic CRF assessment¹. To address the second challenge—exercise-induced noise in single-lead ECG, we first construct a paired synthetic dataset by combining clean ECG segments with extracted artifact signals from two public datasets (Section 3.1). Inspired by the prevalence of diffusion-based models [25] in computer vision, which are well-suited to learning structured noise distributions, we train a diffusion-based ECG denoiser on this dataset to suppress cycling-related artifacts. Finally, to preserve the clinically important morphology (e.g. R peak, P wave, and T wave) and avoid oversmoothing, we employ a composite loss function that explicitly encourages faithful waveform reconstruction under diverse noise conditions.

We build and evaluate CaReFit on an eight-hour dataset collected from 50 patients with different diseases in collaboration with a hospital’s testing department. Because CPET data are dense participant-level time series, leakage-free evaluation is critical. We therefore adopt a strict subject-disjoint protocol in which all segments, metadata, and labels from one participant are assigned to a single fold. The evaluation result shows the mean absolute percentage errors (MAPE) for VO_2max and VO_2AT are 9.30% (95% CI: 7.06–11.80) and 9.49% (95% CI: 7.02–12.12), respectively. We further conduct a head-to-head comparison with a commercial portable device during the same session, using the CPET result as ground truth for VO_2max . These errors are similar to those of commercial portable devices, but with a much lower price (around 800 USD, saving more than 80%) and setup overhead. CaReFit could additionally estimate VO_2AT , which is unavailable from the commercial system. To accelerate innovation in accessible health monitoring, we release this multi-sensor CPET dataset with the evaluation code. We summarize the contributions of this work as follows:

- We propose CaReFit, the exploratory low-cost yet accurate CRF assessment system. Under the strict subject-disjoint setting evaluation, CaReFit achieves low prediction error using only a single-lead ECG and a SmO_2 sensor.
- We synthesize a noisy ECG dataset from two public datasets and leverage a diffusion-based denoising model to remove motion artifacts from single-lead ECG signals during exercise.
- We design physiology-based unimodal encoders for ECG and SmO_2 and develop a cross-attention fusion block to combine these two modalities for a holistic CRF prediction.
- We collaborate with a hospital testing center and evaluate our system on real-world patient data, and open-source this multi-sensor dataset as well as evaluation code under the CPET exercises. They are publicly available in the project repository via <https://github.com/ERICXUCHI/CareFit-Dataset.git>.

2 BACKGROUND

Before we dive into the detailed architecture, we will first discuss the background of our research, including how to estimate CRF and its relation with ECG and SmO_2 .

2.1 Cardiorespiratory Fitness and CPET

Cardiorespiratory fitness (CRF) refers to the ability of the circulatory (heart, blood vessels) and respiratory (lungs) systems to supply oxygen to working muscles during sustained physical activity. It is a key indicator of overall health and physical fitness. CRF is often measured by maximal oxygen uptake (VO_2max), which reflects the body’s efficiency in utilizing oxygen, and oxygen uptake at the anaerobic threshold (VO_2AT) where the body shifts from primarily aerobic energy production to increasingly anaerobic metabolism, which is a critical marker of endurance performance. While VO_2max reflects maximal exercise capacity, VO_2AT represents the transition point at which aerobic metabolism becomes insufficient and anaerobic metabolism begins to dominate. This threshold is widely used for exercise prescription, perioperative risk assessment, pulmonary rehabilitation, and chronic disease management because it better reflects sustainable functional capacity under daily-life conditions.

¹It is reasonable to assume that a large number of subjects conducting the CRF assessment already have their PFT report at hand because it is a common and easily available test. In fact, almost all patients recruited in this research already have a PFT report (Section 4).

The golden standard to measure CRF is Cardiopulmonary Exercise Testing (CPET). The test typically involves cycling or treadmill exercise with progressively increasing intensity until the participant cannot persist, spanning from the resting phase, warm-up phase, incremental exercise phase, anaerobic threshold (AT) phase, peak exercise phase (maximal effort), and recovery phase (cool-down) [2]. During the exercise maneuver, the subject must wear a series of sensors to record the real-time physiological response as the exercise intensity increases. To name a few, the subject must wear the 12-lead ECG sensors to measure the cardioactivity, a blood pressure cuff to measure the blood pressure, and a face mask connected to a metabolic analyzer for gas (O_2 and CO_2) exchange analysis. Together, these parameters are analyzed and compute key CRF metrics, such as $VO_2\max$ and VO_2AT , and others. Clinically, CPET helps summarize physiologic competence and diagnose conditions like heart failure, pulmonary diseases, and metabolic disorders.

2.2 ECG and SmO_2 in Relation to Cardiorespiratory Fitness

The electrocardiogram (ECG) is a non-invasive medical test that records the heart's electrical activity, which is a core component of CPET. By tracking the coarse-grained features (e.g. heart rate and rhythm) and fine-grained waveform features (e.g. ST-segment deviations), ECG helps detect cardiovascular abnormalities. In addition, ECG is highly related to oxygen uptake. As the exercise intensity increases, the oxygen demand increases accordingly. Because of that, the ST segment in the ECG waveform will rise more sharply, and the T wave will decrease in height at first but increase with more intense exercises [40]. In other words, the detailed ECG waveform characteristics are predictive of oxygen uptake. Also, the heart rate computed from the ECG signal is a natural predictor of oxygen consumption [6].

In addition to ECG, the muscle function is also one of the essential factors affecting CRF. A valid measurement of blood lactate concentration reveals muscle performance and determines the position of the anaerobic threshold, but it is unrealistic to measure blood lactate concentration because it involves running tests on the blood sample. Even so, according to the muscle oxygen saturation study [19], SmO_2 can help indirectly determine the anaerobic threshold [38], where the high lactate concentration is associated with a low SmO_2 value. Therefore, the SmO_2 sensor, which is noninvasive, can be used to act as a surrogate for the muscle activity measurement during the graded exercise. Moreover, as mentioned before, SmO_2 and ECG signals exhibit temporal delays (lag effects) relative to metabolic and cardiac output during CPET. We would demonstrate and solve this problem in Section 3.3.3 in detail.

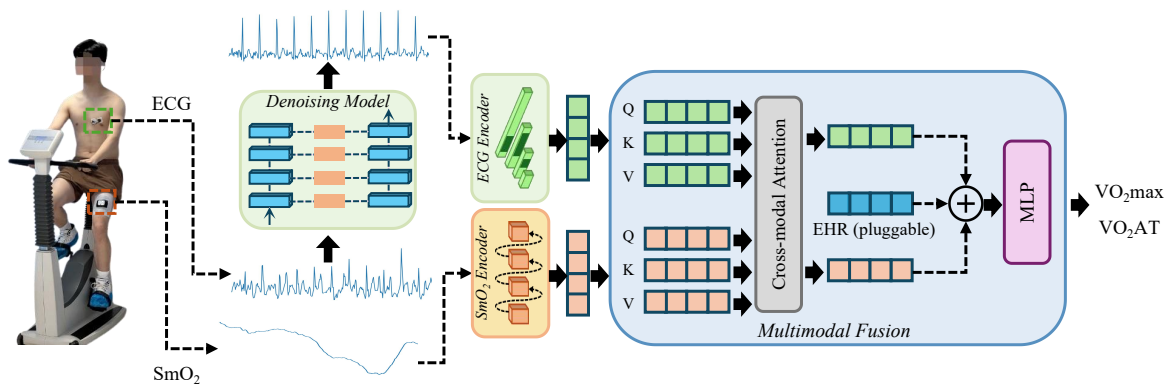


Fig. 2. Overview of CaReFit.

3 SYSTEM DESIGN

In this section, we will elaborate on the details of CaReFit. The general framework is presented in Figure 2. As discussed in Section 1, the ECG signal will be largely affected by the noise induced by the exercises in the CRF assessment. Therefore, the design of CaReFit comprises two main modules, where the first module removes the noise from ECG signals and then the second module predicts oxygen uptake, combining ECG, SmO_2 , and Electronic Health Record (EHR) data. In the first module, since there are no well-established signal-processing algorithms or deep learning models to remove nondeterministic noise from ECG signals, we construct a synthetic dataset and train a diffusion-based noise-removal model [28] from scratch. The synthetic dataset contains a large number of clean ECG signals from the MIT-BIH Normal Sinus Rhythm Database [30], together with noise signal from The MIT-BIH Noise Stress Test Database [31]. We will elaborate on the construction of the noise-ECG dataset as follows. In the second module, since we have multimodal data, we design dedicated encoders tailored to the signals' intrinsic characteristics and fuse the encoded features using a cross-attention mechanism. By leveraging their physiological information, our system performs well in the final prediction.

3.1 Synthetic Noisy Dataset Preparation

As discussed above, the denoising model is trained with a synthetic dataset with clean and noise-corrupted ECG data. Although we can access a large number of clean ECG samples from public datasets, there is no publicly available data on the specific noise generated by CPET. Fortunately, we find that there are public datasets containing various types of noise caused by activities, which resemble CPET exercises. Therefore, we first build a large-scale polluted ECG dataset by combining public clean ECGs and noise signals. Then, we use this dataset to train a diffusion-based denoising model to predict the additive noise and restore the original ECG.

3.1.1 Clean ECG Extraction. The MIT-BIH Normal Sinus Rhythm Database (NSRDB) [30] is a publicly available reference collection, containing 18 long-term ECG recordings from individuals who were found to have no significant arrhythmias. Each recording is approximately 24 hours in duration and includes two ECG signals sampled at 128 Hz. There are two channels in each data, named *ECG1* and *ECG2*, instead of the real leads' names.

3.1.2 Noise Selection. The MIT-BIH Noise Stress Test Database (NSTDB) [31] is a specialized collection that, in addition to the clean ECG signals, includes three types of calibrated noise: baseline wander (BW), electrode motion artifact (EM), and muscle artifact (MA), at varying signal-to-noise ratios. Among typical ECG artifacts, the electrode motion artifact presents the greatest challenge. Its morphology can closely resemble ectopic beats, and unlike other noise types, it resists removal by standard filtering techniques according to the website.

3.1.3 Synthetic Dataset. In our scenario, the participants are required to perform cycling with progressively increasing resistance, where the predominant noise is electrode motion artifact (EM). The reason is that during cycling, participants sweat profusely and violently shake their bodies (especially the upper limbs and trunk), causing sharp, rapid fluctuations in the contact impedance between the electrodes and the skin. Most people do not contract the upper limb muscles significantly, and baseline wander is weak throughout the CPET. Therefore, we only leverage the EM noise from NSTDB. To build the polluted ECG dataset, inspired by the data processing in [28, 55], we select a noise segment of matching length for each clean ECG segment and scale it adaptively by a random factor α (0.2–2.0), determined by the ratio between the noise's peak-to-peak amplitude and the ECG's amplitude range to maintain realistic signal-to-noise variability. This process generated a paired dataset where each noisy observation corresponded to its clean counterpart, enabling supervised learning for denoising tasks across complex noise types at controlled contamination levels. The examples of our synthetic dataset are depicted in Figure 3.

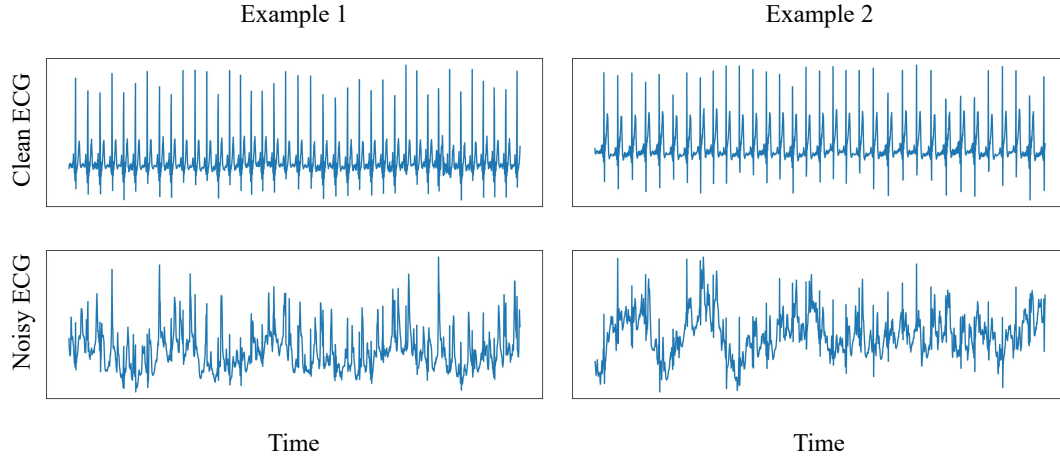


Fig. 3. Examples of Synthetic Dataset.

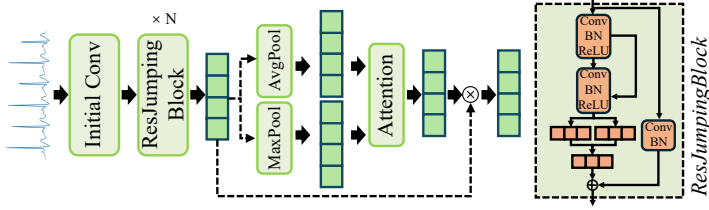


Fig. 4. ECG Encoder.

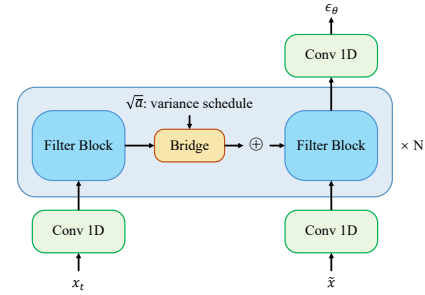


Fig. 5. Pipeline of Denoising Model.

3.2 Diffusion-Based ECG Denoising Model

ECG signals collected under ambulatory or exercise conditions are inevitably contaminated by various types of noise and artifacts. Traditional denoising approaches, such as wavelet filter-based methods [12], often struggle to preserve fine-grained morphological features while effectively suppressing noise, particularly under high-amplitude corruption. To address these challenges, we utilize a diffusion-based denoising framework that fundamentally reformulates noise removal as a conditional generative process. Unlike deterministic mapping approaches, this model learns the underlying conditional distribution of clean ECG signals given noisy observations, enabling high-fidelity reconstruction through iterative refinement from pure Gaussian noise. This subsection provides the theoretical foundation of Denoising Diffusion Probabilistic Models (DDPMs), describes our adapted network architecture with temporal attention mechanisms, and introduces a composite loss function specifically designed to preserve clinically relevant ECG characteristics.

3.2.1 Denoising Model. Denoising Diffusion Probabilistic Models (DDPMs) [25] are a class of generative models that learn to reverse a gradual noising process, allowing them to generate or reconstruct data from pure noise. In the context of ECG denoising, DDPMs are adapted to remove electrode motion artifact and other structured noise by learning to iteratively refine a noisy signal toward a clean version. The mathematical expression for DDPM and its transformation under ECG conditions are formulated in [Section B.1](#).

The framework is mainly based on DeScoD-ECG [28]. We leverage the Dual-Stream Design, where the network has two parallel streams: one processes the noisy observation $\tilde{x}$, and the other processes the current latent variable x_t at diffusion timestep t . The main building block is inspired by multi-kernel filters and half-instance normalization. It uses multi-scale convolutional filters to extract features at different receptive fields, followed by channel concatenation, a convolutional layer, and a half-instance-norm operation (normalizing only half the features to preserve natural statistics). To facilitate communication between the two streams. It conditions the fusion of features on the continuous noise level $\sqrt{\alpha_t}$ using a modified Feature-wise Linear Modulation mechanism. The noise level is encoded via sinusoidal embeddings and a 1×1 convolutional layer to produce scaling and shifting vectors that modulate the features. Meanwhile, to enhance the model's capability to capture critical temporal dependencies in ECG signals—particularly the morphological characteristics of QRS complexes, P and T waves, we integrate a Temporal Attention mechanism within the Half Normalized Filter block. This addition enables the network to adaptively focus on diagnostically significant waveform segments while processing multiscale features. The overview of the model is illustrated in Figure 5.

Thus, the improved Filter Block can be expressed in mathematics as Equation B.7, where input features $F \in \mathbb{R}^{B \times C \times T}$, and B, C, T denote batch size, channels, and time steps.

3.2.2 Loss Function Design. The original loss function is to minimize the difference between real noise and predicted one, as expressed in Equation B.6. However, this single loss has several drawbacks when it deals with ECG denoising process. The main drawback is that it focuses only on the noise discrepancy, not the ECG signal itself. Even though we apply such loss on denoised signals, it inadequately considers sharp morphological features of the original ECG, particularly the rapid QRS complexes with steep slopes, as the loss tends to oversmooth these critical edges while prioritizing average error reduction at the expense of local details. Moreover, the uniform treatment of all frequency components neglects the distinct spectral bands that characterize different ECG waveform components, potentially distorting clinically-relevant frequencies from baseline wander to high-frequency QRS notches. Last, equal weighting across all temporal regions disregards the varying clinical significance between diagnostically critical QRS complexes and less informative baseline segments. Thus, we have designed multiple loss functions from different perspectives to restore the original clean ECG as follows.

- The first improvement is the gradient loss, which captures the temporal sharpness of waveforms by computing first- and second-order derivatives:

$$\mathcal{L}_{\text{grad}} = \sum_{i=1}^{T-1} |\nabla \hat{x}_i - \nabla x_i| + 0.5 \sum_{i=1}^{T-2} |\nabla^2 \hat{x}_i - \nabla^2 x_i| \quad (3.1)$$

where ∇ and ∇^2 denote first and second finite differences. This formulation explicitly enforces preservation of rapid transitions, particularly the steep upstrokes and downstrokes of QRS complexes, while the second-order term helps maintain waveform concavity patterns essential for ST-segment and T-wave morphology.

- Another design is the frequency-domain loss, operating in the spectral domain to ensure proper distribution of signal energy across physiologically relevant frequencies shown in Equation 3.2, where $\mathcal{F}$ denotes the Fourier transform and $|\cdot|$ extracts magnitude spectra. By comparing amplitude spectra directly, this loss prevents spectral leakage of noise into diagnostic bands and maintains the harmonic structure inherent to cardiac electrical activity. This is particularly crucial for suppressing baseline wander (typically $< 0.5\text{Hz}$) while preserving higher-frequency QRS components ($10 - 25\text{Hz}$).

$$\mathcal{L}_{\text{freq}} = |||\mathcal{F}\{\hat{x}\}| - |\mathcal{F}\{x\}|||_1 \quad (3.2)$$

- The last part is the QRS-weighted loss, which introduces an adaptive spatial weighting scheme that emphasizes regions with high gradient magnitude—a reliable indicator of QRS locations:

$$\mathcal{L}_{\text{QRS}} = \sum_{i=1}^T w_i |\hat{x}_i - x_i|, \quad w_i = 1 + 2 \cdot \frac{|\nabla x_i| - \min(|\nabla x|)}{\max(|\nabla x|) - \min(|\nabla x|)} \quad (3.3)$$

The weights w_i are derived from normalized gradient magnitudes of the clean reference signal, automatically amplifying the importance of high-slope regions by a factor without requiring explicit QRS detection algorithms. This data-driven weighting ensures that reconstruction errors in critical diagnostic regions receive proportionally greater penalty during training.

Finally, we combine each loss function with hyperparameters λ_1 and λ_2 . We choose the L1 loss as the base loss because it is less sensitive to extreme values than the L2 loss and provides more stable training by reducing the disproportionate influence of outliers. The total loss is formulated in [Equation 3.4](#).

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{L1}} + \lambda_2 (\mathcal{L}_{\text{grad}} + \mathcal{L}_{\text{freq}} + \mathcal{L}_{\text{QRS}}) \quad (3.4)$$

3.3 Multimodal Signal Encoding and Fusion

As discussed previously, we have two types of signals: ECG and SmO_2 . ECG captures cardiac electrical activity, while muscle oxygen saturation SmO_2 reflects metabolic demand. However, these signals differ in temporal dynamics, noise profiles, and physiological meaning, making simple fusion approaches inadequate. We introduce specialized encoders for each modality—preserving their unique characteristics—followed by a cross-attention fusion mechanism that learns dynamic, non-linear interactions between them. This structured approach enables the model to leverage synergistic information while maintaining interpretability and robustness, ultimately enhancing prediction accuracy for oxygen uptake estimation. For the optional text-based Electronic Health Record (EHR), we use a pluggable multi-layer perceptron to encode text features and improve prediction accuracy.

3.3.1 ECG Encoder. We analyze the characteristics of ECG signals and find that the features can be divided into inter-wave and intra-wave features. Inter-wave features reflect the overall condition changes of the whole body, while intra-wave features reflect the details of cardiac variations. For example, if the S-T segment rises, it indicates that the cardiac output is increasing [40]. Additionally, if the heart rate changes rapidly over a short period, the participants may be approaching the anaerobic threshold. Therefore, our model should recognize both levels of characteristics. The details of the ECG encoder are illustrated in [Figure 4](#).

First, we use multiple ResNet layers with different kernel sizes to capture receptive-field representations of varying sizes. However, a single ResNet cannot capture all features perfectly, because not all signal components are meaningful (e.g., the interval between two waveforms provides little information). We then use the Jumping Knowledge module [54], explicitly save the first convolutional output within a ResNet block, and concatenate it with the second convolutional output. Additionally, we add a sigmoid activation function as a knowledge gate to generate weighted attention. In this way, we can better capture multi-scale features, gradient flow, and feature reuse.

Moreover, to better aggregate the information from different channels, we also use the Channel Attention Mechanism to adaptively adjust the weights. Average pooling preserves global and dominant features, whereas maximum pooling captures the most significant features and highlights high-activation anomalies. After obtaining the weight matrix, we multiply it by the original output to represent ECG signals during exercise, with a focus on diagnostically important ECG segments.

3.3.2 SmO_2 Encoder. Compared to ECG, SmO_2 is not as complicated since fluctuations within small timestamps can be ignored. There are approximately four periods in SmO_2 signals. First, the signal progresses smoothly when participants cycle without any load. During the second period, the curve drops as the cycling load increases. After the participant reaches the anaerobic threshold, the signal drops sharply until the person can no longer persist,

reaching the lowest point at the $VO_2\max$, which is the third phase. The fourth period is called the recovery phase, where the person cycles slowly to a stop.

According to the signal characteristics, we utilize a bidirectional GRU to encode it in the first step to obtain a general representation. Then, we design an attention mechanism to enable the model to identify which timestamps are more important than others. For example, points at the intersection of different periods contain more information. The calculation of attention (α) and output can be expressed in Equation 3.5. With this design, our model can adaptively focus on key periods and enhance its sensitivity to important fragments.

$$\begin{aligned}\alpha_t &= \text{Softmax}(\mathbf{W} \times \tanh(\mathbf{V} \times h_t)) \\ \text{Output} &= \text{MLP}\left(\sum_t \alpha_t \times h_t\right),\end{aligned}\tag{3.5}$$

where $\mathbf{W}$ and $\mathbf{V}$ are parameters.

3.3.3 Fusion Block. Due to hemodynamic response lag [42], the two are not synchronously aligned on the time axis, and this delay is often nonlinear and varies with individual differences. In other words, it is difficult to determine the exact timestamp of the peak oxygen uptake using only the signal's basic shape or features. However, traditional feature fusion (e.g., simple concatenation or weighted sum) typically assumes that features are linearly aligned or in a fixed proportion, and thus is not applicable to complex patterns.

Moreover, EHR consist of table-based data, including personal information and pulmonary function test results. Therefore, we use a multi-layer perceptron to encode them. It is evident that the high-dimensional representation of the first two signals contains more information than the low-dimensional representation of the text. We have designed the following fusion block.

Specifically, we use cross-attention to construct query, key, and value matrices from ECG and SmO_2 encodings, respectively. The attention mechanism will automatically search for the part of the SmO_2 representation that is most relevant to the current ECG feature and vice versa. Despite delays, the model can capture cross-modal nonlinear correlations via the learned weights. Meanwhile, bidirectional crossover and symmetric design ensure that the model can not only address the "who leads and who lags behind" issue but also extract features from both dimensions to form a more robust joint representation, mitigating information loss caused by one-way lag. For implementation, we first use ECG as the query and SmO_2 as the key to compute the initial similarity matrix. We then multiply this matrix by the SmO_2 value to obtain the first fusion output. These operations are performed in reverse with SmO_2 as the query, allowing us to get the second fusion output. This method helps address the lag effect noted earlier, in which changes in ECG and SmO_2 are not linearly correlated.

In detail, we stack the queries, keys, and values as (2, Batch, Dimension) matrices. The calculation formula of *attention_weight* and *fusion* is expressed in Equation B.9 and Equation B.10 using Einstein summation, where m and n can only be 0 or 1, representing two modality indices.

With such calculations, the final fusion includes both self- and cross-modality attention, which is critical for tasks where both modalities provide complementary information.

Moreover, it is unclear whether these two encodings contribute equally to the final output. Thus, we use an equation to adjust their weights through a data-driven method to calculate the final output. If the participant has the EHR, its encoding is fed into the last layer of our fusion block, making the oxygen output prediction more accurate.

Last but not least, an optimized composite loss function is developed to enhance regression performance by jointly enforcing accuracy, trend alignment, and distributional consistency, described in Equation 3.6. The loss integrates three components: (1) a mean-squared error (MSE) term to ensure pointwise numerical precision, (2) a Pearson correlation loss ($1 - r$) to promote strong linear correlation between predictions and targets, and (3) a mean-matching loss that penalizes deviations between the predicted and target means to mitigate systematic

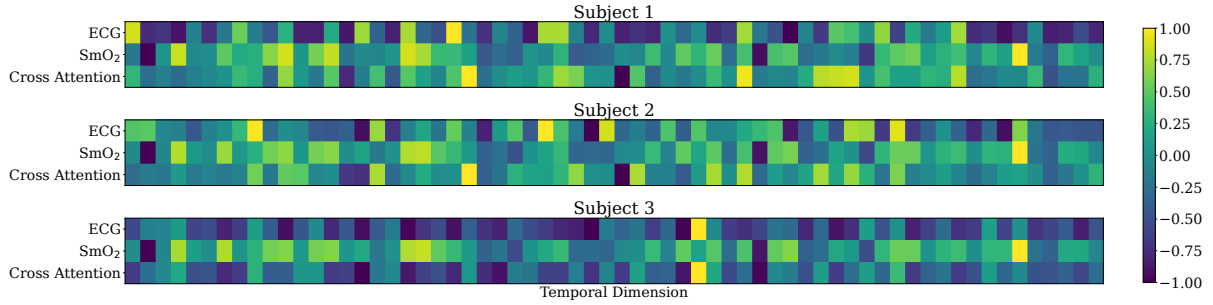


Fig. 6. Feature Representation.

bias. The Pearson component is computed directly from first principles—using centered variables and normalized covariance—to avoid dependency on external libraries. This multi-objective formulation ensures that the model not only minimizes absolute error but also captures the underlying temporal or structural trends while preventing distributional shift.

$$\mathcal{L}_{\text{total}} = \gamma_1 \cdot \text{MSE}(y, \hat{y}) + \gamma_2 \cdot (1 - r(y, \hat{y})) + \gamma_3 \cdot \text{MSE}(\mu_y, \mu_{\hat{y}}), \quad (3.6)$$

We observe that during the ramp phase, ECG activations become denser while SmO₂ representations exhibit a gradual decline, whereas during recovery, SmO₂ recovers smoothly while ECG activity stabilizes, consistent with known physiological dynamics. To further connect the model behavior with physiological reasoning, we visualize the intermediate representations produced by the ECG encoder, the SmO₂ encoder, and the cross-attention fusion module for representative subjects, as shown in Figure 6. The ECG and SmO₂ encoders exhibit distinct activation characteristics that reflect their different physiological roles. Specifically, the ECG encoder produces relatively localized high-response regions corresponding to transient cardiac waveform patterns, whereas the SmO₂ encoder exhibits smoother and more temporally continuous activations, consistent with the slower dynamics of muscle oxygenation during exercise. The fused representation is visually different from either individual modality. It retains prominent localized ECG responses while preserving broader temporal patterns originating from SmO₂, suggesting that the cross-attention module integrates complementary information from both modalities rather than simply averaging or copying their representations. This observation is consistent with our physiological motivation: ECG primarily captures rapid cardiac responses, whereas SmO₂ reflects slower muscle-metabolic changes. Because these two physiological processes are coupled but naturally separated by hemodynamic delay, the cross-attention mechanism enables the model to associate temporally misaligned yet complementary information.

4 DATASET DEVELOPMENT

We recruited a total of 50 participants, including 36 males and 14 females. Among them, 23 are patients with COPD or asthma, 20 are in the perioperative period, and 17 have other diseases (such as hypertension, cancer, hernia, shortness of breath, or weakness in the limbs of unknown cause). Note that some patients may fall into more than one subgroup. All personally identifiable information was removed or de-identified prior to analysis, in accordance with institutional and regulatory guidelines. Participant identifiers were replaced with anonymized study codes. All experiments were approved by the institution's IRB², and doctors from our collaborating hospital supervised the data collection process. We open-source the dataset and evaluation code with the link

²The Hong Kong University of Science and Technology HREP-2025-0298

Table 1. Demographics of the subjects.

Stats	COPD/ asthma	Peri- operative	Others	All
Population*	23	20	17	50
Age (years)	65.8 (8.5)	59.5 (14.1)	53.9 (16.4)	58.9 (14.3)
Height (cm)	165.8 (6.2)	162.3 (7.6)	162.8 (7.2)	163.6 (7.0)
Weight (kg)	67.6 (12.3)	65.4 (11.6)	65.3 (12.7)	65.8 (11.6)
BMI (kg/m ²)	24.5 (3.5)	24.7 (3.0)	24.7 (4.8)	24.5 (3.7)
FVC (L)	2.9 (0.6)	2.9 (0.6)	3.0 (0.7)	3.0 (0.6)
FEV1 (L)	1.6 (0.5)	2.1 (0.7)	2.5 (0.6)	2.1 (0.7)
PEF (L/s)	4.3 (1.6)	5.6 (2.2)	6.5 (1.7)	5.5 (2.0)
FEV1/FVC (%)	56.4 (9.0)	68.9 (14.0)	83.2 (7.6)	70.3 (15.2)
VO ₂ max (ml/kg/min)	20.8 (3.9)	22.9 (5.1)	23.0 (4.1)	22.4 (4.3)
VO ₂ AT (ml/kg/min)	15.2 (3.1)	15.9 (3.7)	14.4 (2.1)	15.2 (3.1)

Format: mean (standard deviation)

*Some patients belong to two subgroups.

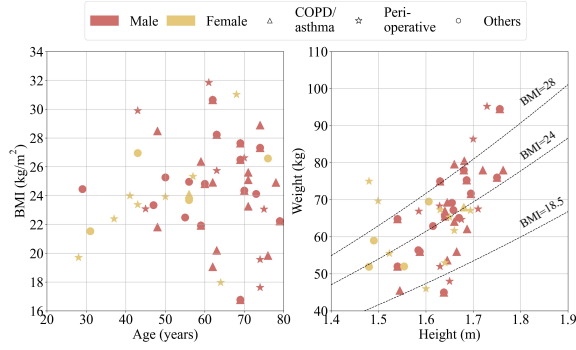


Fig. 7. Details of demographics.

<https://github.com/ERICXUCHI/CareFit-Dataset.git>. The repository includes ECG and SmO₂ signals, labels, patients' metadata, as well as the checkpoints for reproduction.

4.1 Prerequisite

The data were collected in a professional pulmonary function testing room within the hospital's testing department. The doctors assisted with participant recruitment and explained the purpose of the study. After obtaining participants' consent and introducing the study, we first recorded participants' personal information, including age, gender, height, weight, and pulmonary function test results. All subjects were asked to wear a plastic mask connected to a gas analyzer and perform cycling exercises with a gradually increasing workload. In addition to the sensors required for CPET, the user was required to wear our ECG and muscle oxygenation sensors. The ECG patch is attached to the fourth intercostal space on the left edge of the sternum according to the American Heart Journal [9], and the muscle oxygenation patch is attached to the quadriceps femoris muscle, which exerts the most force based on a biomechanical model. The IMU sensor is attached to the feet to record the timestamp of the start of cycling and will not be used during training or inference.

4.2 Collection Procedures

During a patient undergoing a CPET, he/she would sit quietly on the bicycle while resting heart rate, blood pressure, ECG, and resting pulmonary function are recorded, and our system starts to record simultaneously. The first period is an unloaded warm-up for 3 minutes, pedaling against zero or very low resistance to establish a baseline. The second period is the most important, during which the "workload" increases steadily by the minute. The resistance increases in a smooth, linear "ramp" until volitional exhaustion. The goal is to make the patient reach peak exertion from the start of the ramp. This is the "sweet spot" for diagnostic accuracy. The test continues until the patient cannot continue despite strong encouragement ("maximal effort"), or the clinician stops it for safety reasons. It usually lasts about 8 minutes in healthy people (but some severe patients can only persist for 4 minutes). The third period is a 3-minute recovery. After completing the second stage, the workload will be reduced to zero, and the patient will be required to continue cycling to prevent blood pooling and aid recovery while key parameters are still monitored. Thus, each patient contributes about 10 minutes of data, and the total dataset is around 8 hours. The sampling rate is 512 Hz for the ECG sensor, 1 Hz for the SmO₂ sensor, and 50 Hz for the IMU.

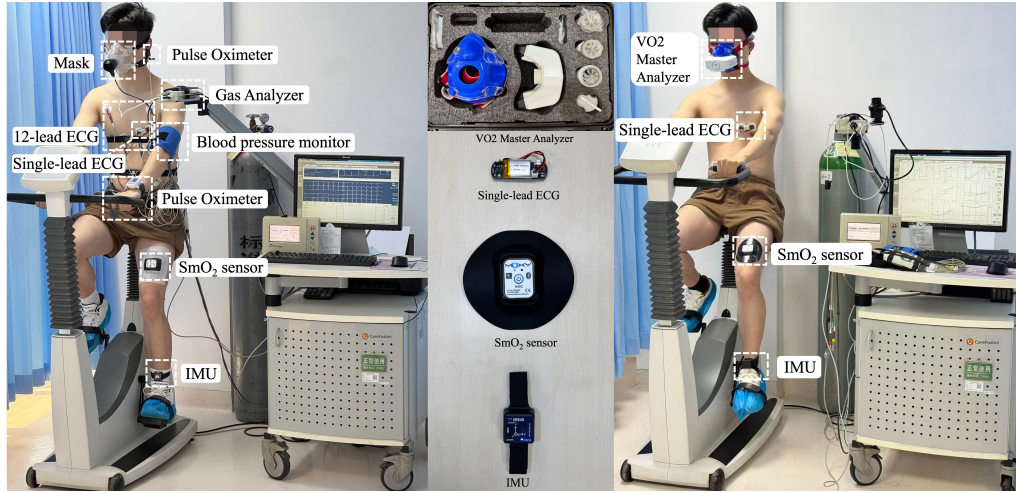


Fig. 8. Data collection setup.

4.3 Data Distribution

The demographic and clinical characteristics of the study cohort are summarized in [Table 1](#) and [Figure 7](#). Key pulmonary function metrics are reported, including Forced Vital Capacity (FVC), Forced Expiratory Volume in one second (FEV1), Peak Expiratory Flow (PEF), and the FEV1/FVC ratio—standard indices for evaluating restrictive patterns, airflow obstruction, and disease severity in conditions such as COPD. Cardiopulmonary fitness is quantified by peak oxygen uptake (VO_2max) and oxygen uptake at the anaerobic threshold (VO_2AT).

Participants were grouped according to their primary clinical indication for CPET: COPD/asthma ($n=23$), peri-operative evaluation ($n=20$), and other diagnostic indications ($n=17$). Some individuals met criteria for more than one subgroup, reflecting overlapping clinical presentations. This distribution aligns with common referral patterns for clinical exercise testing and supports the generalizability of the dataset to real-world patient populations.

5 EVALUATION

This section will give a detailed evaluation of CaReFit, containing the experiment setup, overall performance, and ablation study to show the effectiveness of each module. We also conducted a preliminary qualitative assessment of user experience and clinical workflow. Given the limited sample size, we present these findings as exploratory observations in [Appendix A](#).

5.1 Experiments Setup

5.1.1 Hardware and Software. CaReFit adopts three sensors, a single-lead ECG sensor with BMD101 Chip, an NIRS Moxy Monitor (Fortori Designs LLC, Hutchinson, MN) sensor [32] to monitor the muscle oxygenation signals, and an IM948 IMU attached to the user's feet for time synchronization. The sampling rate is 512 Hz for the ECG sensor, 1 Hz for the SmO_2 sensor, and 50 Hz for the IMU. We build our system based on Python 3.12 and PyTorch 2.2.1 on NVIDIA® GeForce RTX™ 4090. The data collection software is written with the PySerial package, and the Moxy Monitor is connected to VO2 Master APP. The collection scenario is shown in the left part of [Figure 8](#). The ground truth is JAEGER MasterScreen™ CPX metabolic cart. For the hyperparameters of the deep learning pipeline, all dropout rates are 0.3. The optimizer is AdamW, with a learning rate of $1e-4$ and $1e-2$

weight decay. The LR scheduler is StepLR, with a step size of 50 and a gamma of 0.5. Regarding the first diffusion model's loss function, we let $\lambda_1 = 1.0$ and $\lambda_2 = 0.01$. This configuration is adopted because the magnitude of the second loss term is typically substantial; assigning it a smaller weight prevents it from dominating the overall loss and overshadowing the primary objective function. The final loss function combines MSE, a Pearson correlation loss ($1 - r$), and a mean-matching loss. Their weights are set as $\gamma_1 = 0.5$, $\gamma_2 = 0.5$, $\gamma_3 = 0.1$. Such a weighting scheme balances the contributions of each objective, ensuring that the primary emphasis remains on minimizing prediction error (MSE) and maximizing linear correlation, while the auxiliary mean-matching term provides a weaker regularizing influence. During each CPET session, ECG and SmO_2 signals are continuously recorded throughout the incremental cycling protocol. Following preprocessing and synchronization, the signals are taken as an entire temporal window for model input.

5.1.2 Commercial Device Comparison Protocol. To provide a more direct comparison with commercial wearable devices, we conducted an additional controlled study. One widely used device is the VO2 Master Analyzer (VO2 Master Health Sensors Inc., Vernon, British Columbia, CA) [35], so we select it as the comparison device. According to the IRB's constraint that the patients are unable to perform a two-day cycling exercise because of potential health risks, the original clinical cohort could not be reused for external device benchmarking. Therefore, we recruited 10 healthy participants for this comparison. We note that commercial devices are typically validated on healthy populations, whereas our primary study focuses on clinical patients with heterogeneous conditions. This population mismatch limits direct comparability, and the controlled experiment should be interpreted as a reference comparison rather than an equivalent clinical validation.

The study followed a two-day protocol. On Day 1, each participant underwent a standard cardiopulmonary exercise test (CPET) in a clinical setting to obtain ground-truth measurements of VO_2max and VO_2AT . These values serve as the ground truth for subsequent evaluation. On Day 2, participants performed the same cycling exercise session under a controlled and reproducible protocol. During this session, participants simultaneously wore our CaReFit system and the VO2 Master Analyzer. Each person spent five minutes choosing the mask that best suited their face size. To ensure strict comparability between the two sessions, we used a computer-controlled ergometer bicycle to replicate the exact same exercise protocol (i.e., identical workload progression and resistance schedule) as in Day 1. This design minimizes variability due to differences in exercise intensity or pacing across sessions. The comparison is demonstrated in the right part of Figure 8.

For the commercial device, VO_2max estimates were obtained directly from its companion mobile application. For CaReFit, VO_2max was predicted from the recorded ECG and SmO_2 signals using the trained model. We then compared both estimates against the CPET-derived ground truth from Day 1. The results are demonstrated in Section 5.2.3. It is important to note that the commercial device does not provide VO_2AT estimation. Therefore, this comparison is limited to VO_2max . While this experiment enables a head-to-head comparison under identical exercise conditions, we acknowledge that the sample size is limited and consists only of healthy participants. Thus, the results should be interpreted as a preliminary validation rather than a comprehensive clinical benchmarking.

5.1.3 Evaluation Metrics. To evaluate the system performance, we adopt the most widely used metric called Mean Absolute Percentage Error (MAPE),

$$\text{MAPE} = \sum \frac{\text{abs}(\text{Pred} - \text{Label})}{\text{abs}(\text{Label})} / N \times 100\% \quad (5.1)$$

This metric can better measure the overall accuracy of our predictions compared with Mean Absolute Error.

5.2 Performance Study

5.2.1 Overall Performance. To avoid potential data leakage in this clinical time-series setting, we adopt a strict *subject-disjoint* evaluation protocol for all experiments. We report Leave-One-Subject-Out (LOSO) as the primary

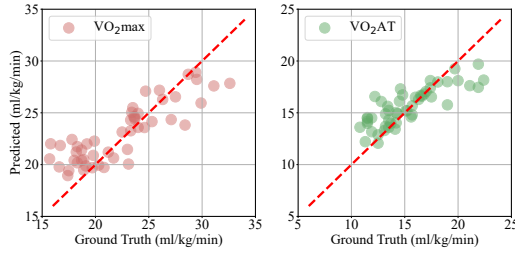


Fig. 9. Scatter diagram.

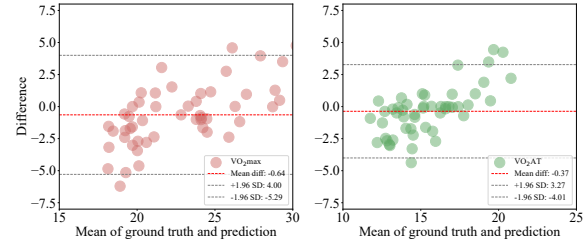


Fig. 10. Bland-Altman plot.

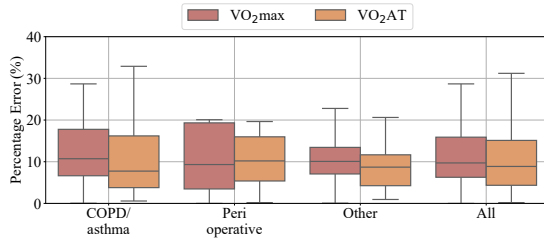


Fig. 11. Demographic study - disease.

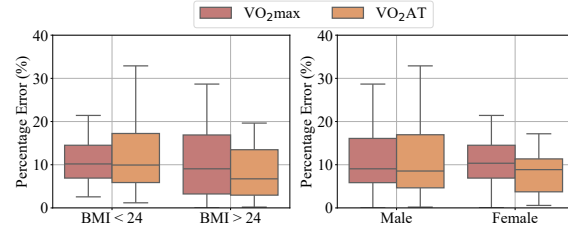


Fig. 12. Demographic study - BMI and gender.

result and 10-fold CV is used only for comparison/ablation. Specifically, all evaluations are performed at the participant level rather than the segment level: all ECG segments, SmO₂ recordings, metadata, and labels from the same participant are assigned to exactly one fold and never appear in both training and testing sets simultaneously. This prevents identity leakage across folds and ensures that the reported performance reflects true cross-subject generalization rather than within-subject memorization.

We conduct a Leave-One-Subject-Out (LOSO) evaluation as a strict validation protocol. In LOSO, each participant is used once as the test subject while the remaining participants are used for training. We compute 95% CIs by bootstrapping participants over LOSO test predictions with 5000 resamples. The overall performance is shown in Figure 9 and Figure 10, where the mean absolute percentage errors (MAPE) for VO₂max and VO₂AT are 9.30% (95% CI: 7.06–11.80) and 9.49% (95% CI: 7.02–12.12), respectively. For mean absolute error (MAE), they are 1.92 (95% CI: 1.51–1.97) and 1.38 (95% CI: 1.02–1.74) for two metrics, demonstrating that CaReFit maintains stable performance even under strict subject-independent evaluation. The model generally preserves the relative ranking of participants, allowing clinically meaningful differences in exercise tolerance to be identified in most cases. According to the medical reference [27, 47], such performance supports our system’s utility as a digital marker of cardiorespiratory fitness evaluation.

For all comparisons, we utilize strict subject-disjoint 10-fold cross-validation (10-fold CV), where all segments from a participant were assigned to exactly one fold to prevent data leakage to ensure fairness and the MAPE for VO₂max and VO₂AT is 10.59% and 9.95% during 10-fold CV. We also report the Mean Absolute Errors (MAE), which are 2.32 and 1.55 ml/kg/min, respectively. Furthermore, we validate our system on various patients, including those with COPD or asthma, other diseases, and postoperative patients. Note that our model does not perform very well when processing extreme values, which may be due to the lack of samples at these extremes. This will be discussed in more detail in the following section. However, this can be improved as the amount of data increases.

Table 2. Comparison with baselines.

Baselines	Akay <i>et al.</i> [1]		Shandhi <i>et al.</i> [39]		Vaishnavi <i>et al.</i> [46]		CaReFit
Dataset*	Original	Ours	Original	Ours	Original	Ours	Ours
VO ₂ max (ml/kg/min)	6.24	11.75	3.68	3.45	5.17	4.34	2.32
VO ₂ AT (ml/kg/min)	- [†]	- [†]	- [†]	- [†]	- [†]	- [†]	1.55

* "Original" denotes the performance reported in the respective publications using their own datasets. These values are included only for contextual reference and are not directly comparable to the re-implemented results ("Ours"), which are evaluated on our dataset under a unified protocol. Since the baselines report Mean Absolute Error, we use this metric as a comparison.

[†] Not provided in the paper.

Table 3. Demographic characteristics of the 10 healthy participants used for commercial-device comparison.

Gender	Age	Height (cm)	Weight (kg)	VO ₂ max (ml/kg/min)
F	24	160	48.0	41.0
F	25	167	50.0	46.3
F	27	166	65.0	35.0
F	22	160	60.0	49.2
F	32	153	44.5	25.4
M	26	178	71.0	34.1
M	28	180	70.0	29.4
M	25	171	77.0	30.8
M	22	173	66.0	33.4
M	23	172	65.0	27.8
Mean (SD)	25.4 (3.06)	168.0 (8.51)	61.65 (10.81)	35.24 (7.89)

5.2.2 Demographic Factor. The demographics of our dataset are presented in [Section 4.3](#). To demonstrate our system's performance across different demographic groups, we evaluate CaReFit in terms of various diseases, BMI, and gender.

We first evaluate our system's performance across three clinical categories: COPD/asthma, peri-operative, and other diseases (including hypertension, cancer, hernia, shortness of breath, or weakness in the limbs of unknown cause). As shown in [Figure 11](#), prediction errors for VO₂max and VO₂AT are similar across all groups, indicating consistent system performance regardless of the underlying clinical condition.

Second, we analyze performance stratified by body mass index (BMI) and gender (see [Figure 12](#)). This analysis yields two key observations: (1) Model accuracy remains relatively stable across BMI categories, with a slight increase in error observed among lower-BMI patients; and (2) No large gender-related difference was observed in this cohort. Although prediction intervals are marginally wider for male participants in some cases, overall accuracy is comparable between genders.

We attribute these observed variations primarily to limitations in the dataset composition rather than to inherent algorithmic flaws. For instance, the cohort contains fewer patients with a BMI below 24, which likely contributes to the marginally higher error rate in that subgroup. Importantly, the model's performance across diverse BMI and gender groups supports its generalizability and underscores its robustness. Future work will focus on refining the model to enhance precision within specific subpopulations. This will involve exploring hybrid modeling techniques and expanding training datasets to better represent demographic and clinical heterogeneity, which is anticipated to reduce current prediction errors.

5.2.3 Comparisons with the Commercial Device. To conduct a head-to-head empirical comparison, we used a commercial VO_2max estimation device under the same CPET protocol described in [Section 5.1.2](#). We recruited 10 healthy participants and collected data from both the commercial device and our sensing setup during the same exercise sessions. The demographic characteristics of these participants are summarized in [Table 3](#). The ground-truth VO_2max values were obtained from the clinical metabolic cart used in our main study. We emphasize that this experiment is intended as a controlled supplementary comparison for VO_2max estimation only, rather than a full replacement for clinical validation. Unlike CaReFit, the tested commercial device does not provide VO_2AT estimation, and therefore the comparison is limited to VO_2max . In addition, this supplementary cohort consists of healthy participants, while our main evaluation focuses on clinical patients.

The commercial device achieves a MAPE of 14.56% for VO_2max prediction, whereas CaReFit achieves a MAPE of 15.09% on the same participants under the same protocol. This result suggests that, in this supplementary healthy-subject comparison, our lightweight ECG-SmO₂ sensing approach provides similar VO_2max estimation to the tested commercial device. More importantly, this direct comparison indicates CaReFit shows promising accuracy relative to a commercial VO_2max estimation device in a small controlled cohort, while further large-scale and multi-site validation is still needed before making stronger clinical claims. Thus, this preliminary comparison provides a direct reference under the same exercise protocol, but we do not claim equivalence with commercial devices pending further validation. Our reported errors correspond to prediction errors against CPET-derived ground truth, whereas commercial devices report measurement errors relative to gas analysis systems; these two should not be directly conflated.

5.2.4 Comparisons with Baselines. We select three representative studies that estimate VO_2max based on electronic health record (EHR) variables or physiological features as reference baselines. Since the original datasets are not publicly available, we re-implement these methods following the descriptions in the respective papers and evaluate them on our dataset. The comparison results are summarized in [Table 2](#). However, we emphasize that these methods were originally designed under different sensing modalities and problem formulations, and therefore are not directly comparable to our setting. To enable approximate comparison, we re-implemented or adapted these methods using the available modalities in our dataset.

- Akay *et al.* [1]. This research derives a multivariate regression equation to predict VO_2max by six predictors—gender, age, height, weight, maximal heart rate HR_{max} , and test duration (TT). The equation is given in [Equation B.8](#). Note that this regression equation was developed using only 18 healthy sports-major university students. Our cohort, however, consists predominantly of middle-aged and older clinical patients undergoing CPET for cardiopulmonary assessment, covering a substantially different distribution of age, fitness level, and disease status. Since the regression coefficients were optimized for a young healthy population, they do not generalize well to our clinical cohort.
- Shandhi *et al.* [39]. This study focuses on using a combination of seismocardiogram (SCG), ECG, and atmospheric pressure (AP) signals to estimate instantaneous VO_2 . Due to the lack of SCG and AP, we re-implement this method using only ECG data. Because ECG and SCG are both utilized to extract heartbeat information, using only ECG is feasible. The predicted result of our ECG demonstrates similar performance of this study.
- Vaishnavi *et al.* [46]. This paper employed four features—gender, body mass index (BMI), aerobic duration, and anaerobic duration—to predict VO_2max . It compared the performance of multiple machine learning models, including Linear Regression, Ridge, Lasso, etc. Among these methods, SVR exhibited lower prediction errors. We adopt the same parameters as the paper and run the algorithm on our dataset.

The results in [Table 2](#) should be interpreted as a coarse reference rather than a strict state-of-the-art comparison. It demonstrates that CaReFit achieves lower prediction error on VO_2max compared to these reference baselines under the same evaluation setting. This suggests that leveraging multimodal physiological signals (ECG and

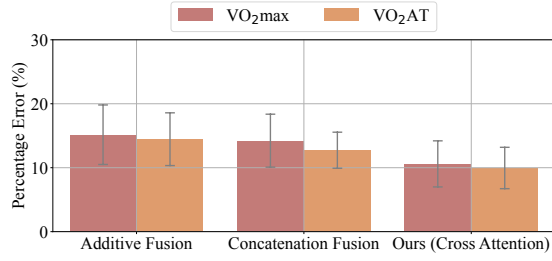


Fig. 13. Different fusion methods.

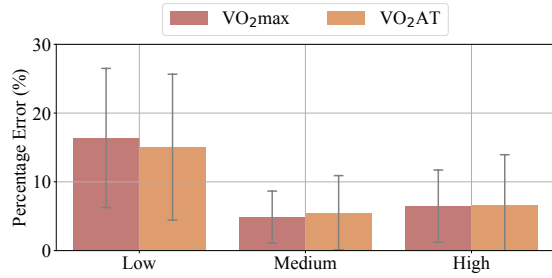


Fig. 14. Stratified performance.

Table 4. System delay.

Device	Denoising	Encoding and Fusion		
		94.53 ms		
CPU	52.42 s	ECG	SmO₂	Fusion
		8.51 ms	84.93 ms	1.09 ms
GPU	6.26 s	11.86 ms		
		ECG	SmO₂	Fusion
		3.11 ms	8.34 ms	0.41 ms

Table 5. Stratified performance details.

Metric	Group	N	GT Range	MAPE (%)	MAE
VO ₂ max	Low	17	15.7–19.2	16.38	2.84
VO ₂ max	Medium	17	19.7–23.7	4.87	1.08
VO ₂ max	High	16	24.0–32.6	6.47	1.84
VO ₂ AT	Low	19	10.8–13.5	15.04	1.81
VO ₂ AT	Medium	14	13.7–16.0	5.51	0.81
VO ₂ AT	High	17	16.3–22.4	6.57	1.35

SmO₂) together with learned representations can provide improved estimation performance compared to formula-based or tabular-feature-based approaches. We emphasize that these results should be interpreted in light of the differences in modality and task formulation discussed above.

5.2.5 Power Consumption and System Latency. The ECG power consumption is around 70 mW, and the muscle oxygenation device is 1-2 watts. It is hard to get the exact value of the NIRS Moxy Monitor because it is a commercial product. Both batteries can last more than 4 hours, enough for all kinds of CPET. For inference, we split the pipeline into two stages: (1) Denoising and (2) Encoding and Fusion. Each component is tested on both CPU "Intel® Core™ i9-13900KF Processor" and GPU "NVIDIA® GeForce RTX™ 4090".

As presented in Table 4, the denoising process takes most of the time, while the prediction latency can be negligible. The reason is that DDPMs perform generation by simulating a lengthy, step-by-step physical process (reverse diffusion), where each of the 50 predefined steps requires a full forward pass through a large neural network.

To mitigate this latency and accelerate the inference in future implementations, several strategies can be explored. Architectural optimizations, such as using distilled or consistency models, can significantly reduce the number of diffusion steps required without a substantial loss in output quality. Alternatively, implementing more efficient sampling schedulers or leveraging latent diffusion techniques could streamline the process, like Denoising Diffusion Implicit Models (DDIM) [43]. These improvements would enhance the practicality of CaReFit for lower-latency deployment. At its current stage, the system is more appropriately positioned for post-session or near-offline analysis, where the denoising latency has limited impact on the overall assessment workflow.

5.2.6 Stratified Performance. To better evaluate our system, we further report stratified performance across clinically relevant fitness ranges. Instead of stratifying participants by predicted values, we divide participants into low, medium, and high groups according to the ground-truth VO₂max and VO₂AT values, respectively. Participants were stratified independently for VO₂max and VO₂AT according to tertiles of the corresponding ground-truth values. Specifically, the bottom, middle, and top thirds of the ground-truth distribution were labeled

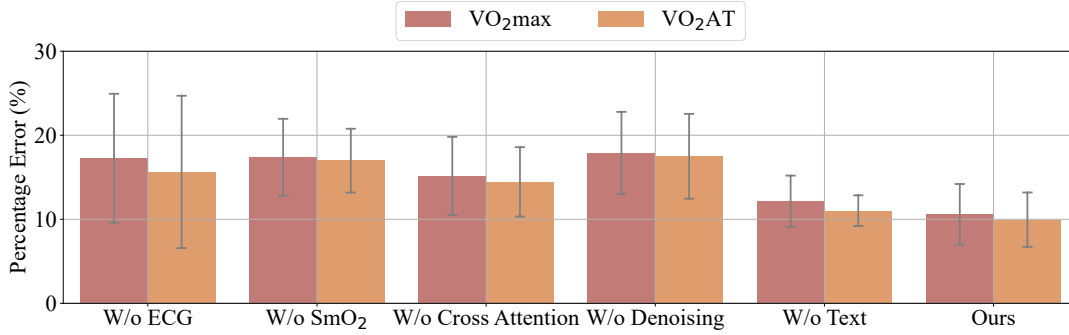


Fig. 15. Ablation study.

as low, medium, and high fitness groups, respectively. This protocol avoids arbitrary threshold selection and yields approximately balanced subgroup sizes, avoiding bias from model predictions, and allowing us to examine whether CaReFit performs consistently across different fitness levels. As shown in Figure 14, CaReFit achieves lower errors in the medium and high fitness groups, while the low-fitness group exhibits higher prediction errors for both VO₂max and VO₂AT. For VO₂max, the MAPE values are 16.38%, 4.87%, and 6.47% for the low, medium, and high groups, respectively. For VO₂AT, the corresponding MAPE values are 15.04%, 5.51%, and 6.57%, as depicted in Table 5. The higher error in the low-fitness group is consistent with our observation that extreme or clinically severe cases are under-represented in the current dataset and often exhibit larger physiological variability during CPET.

These results suggest that CaReFit maintains stable performance for medium- and high-range fitness levels, while further data collection is needed to improve calibration for low-fitness participants. This limitation is important because low-fitness subjects are clinically meaningful in rehabilitation and disease-management scenarios. In future work, we will expand the dataset with more low-fitness and severe clinical cases to improve subgroup robustness.

5.3 Ablation Study and Analysis

We present a comprehensive ablation study to validate the contribution of each module in our system. The technical components are divided into five parts. First, we need to verify the function of each modality, which means we use only ECG or SmO₂ separately to predict oxygen uptake. Second, since we claim our fusion block is effective, we replace it with a basic matrix addition operation to demonstrate the difference. Third, we fuse the ECG without MA removal and SmO₂ to verify the impact of denoising methods, and the effectiveness of the EHR. Each ablation result is displayed in Figure 15.

5.3.1 Impact of ECG. When we remove the ECG signal from our system, the error increases rapidly. This is because SmO₂ data can only provide the information about the stage the patient is performing, lacking physiological details. As we mentioned previously, both coarse-grained and fine-grained waveforms indicate cardiac activity, helping to provide an accurate prediction at the maximal exercise stage.

5.3.2 Impact of SmO₂. Muscle oxygenation divides the exercise into four parts clearly, including warm-up, increasing workload, anaerobic threshold, and maximal exercise stage. Without SmO₂, even though we preserve the cardiac information, our system lacks the hints of which stage the patient is in. Specifically, it is hard to identify the anaerobic threshold, let alone predict VO₂AT.

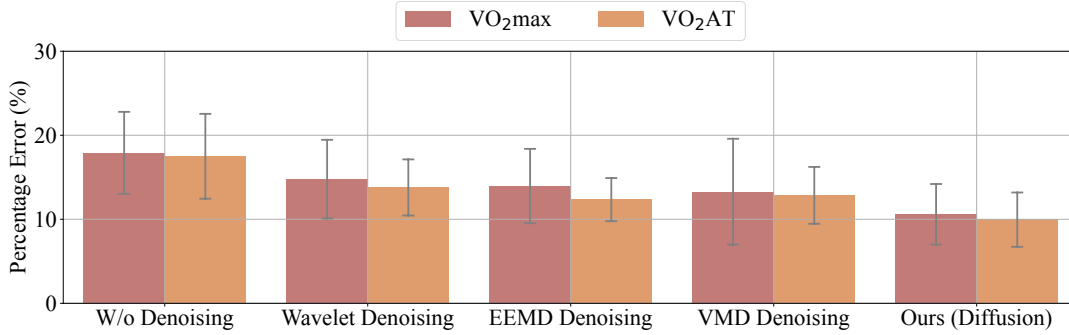


Fig. 16. Comparison of different denoising methods.

5.3.3 Impact of Fusion Strategy. We conduct an additional ablation study by replacing the proposed cross-attention fusion block with two commonly used lightweight fusion strategies: *additive fusion* and *concatenation fusion*.

In the additive fusion setting, the modality-specific representations from the ECG encoder, the SmO₂ encoder and text embedding are directly summed before the final prediction head, assuming equal and fixed contributions from both modalities. In the concatenation fusion setting, the three representations are concatenated and then projected through a multilayer perceptron for prediction, allowing higher-dimensional feature aggregation but still without explicit modeling of cross-modal interactions. In contrast, our proposed cross-attention fusion constructs bidirectional query-key-value relationships between ECG and SmO₂ representations, enabling each modality to dynamically attend to the most relevant information from the other modality. This design is particularly important because ECG and SmO₂ are physiologically coupled but temporally misaligned due to hemodynamic response lag during CPET.

The results are shown in Figure 13. Both additive fusion and concatenation fusion produce noticeably higher prediction errors than our cross-attention design. Specifically, additive fusion achieves approximately 15.16% and 14.46% MAPE for VO₂max and VO₂AT, while concatenation fusion reduces the errors slightly to 14.23% and 12.73%, respectively. Our cross-attention fusion further improves the performance to 10.59% for VO₂max and 9.95% for VO₂AT, showing the lowest error across both tasks. These results indicate that simple feature combination strategies are insufficient for accurate CRF estimation. Although concatenation preserves more modality-specific information than direct addition, it still treats the two modalities as statically aligned features and cannot effectively handle the nonlinear and subject-dependent lag between cardiac response and muscle oxygen dynamics. By explicitly modeling bidirectional cross-modal dependencies, the proposed fusion block better captures the physiological relationship between ECG and SmO₂, leading to more robust and accurate prediction of both VO₂max and VO₂AT.

5.3.4 Impact of Diffusion Denoising Methods. During cardiopulmonary exercise testing (CPET), ECG signals are strongly contaminated by motion artifacts, muscle activity, electrode displacement, baseline drift, and high-frequency noise—especially near peak exercise (VO₂max), which is also the most clinically informative stage. In severe cases, even R-peak detection becomes unreliable. Because our downstream task focuses on subtle morphology changes related to oxygen uptake, preserving physiological structures is more critical than merely reducing noise variance. Therefore, an effective denoising method must both suppress noise and retain clinically meaningful waveform morphology. To validate our diffusion-based approach, we compared it with conventional methods as follows. This experiment aims to quantify the contribution of the denoising module and assess

whether diffusion-based denoising offers unique advantages for exercise-phase ECG signals. Implementation details are provided in [Section B.5](#).

Wavelet denoising was implemented based on the classical soft-threshold wavelet shrinkage framework proposed in [15]. Although wavelet denoising is effective for suppressing baseline drift and high-frequency interference, motion artifacts during peak exercise frequently overlap with the frequency band of the QRS complex, making simple thresholding prone to removing diagnostically important morphology together with noise. As a result, although wavelet denoising improves signal quality, it often fails to preserve the fine-grained physiological information required for accurate oxygen uptake prediction, as shown in [Figure 16](#).

For the EEMD [44] method, we manually analyze the various types of noise present in the raw sensor signals to recover the clean ECG. Although it is effective, EEMD introduces randomness through noise injection and requires the manual selection of the relevant IMF. This means professionals need to check all processed signals, which prevents full automation and renders the method unsuitable for large-scale datasets. Moreover, the algorithm remains sensitive to the choice of noise amplitude and ensemble size. It is unrealistic to automatically remove noise with EEMD.

The Variational Mode Decomposition (VMD) [16] algorithm is a non-recursive, adaptive signal decomposition method that decomposes a real-valued input signal into a discrete number of band-limited intrinsic mode functions (IMFs). However, the result is still less than satisfactory. Under severe motion artifacts near peak exercise, inaccurate parameter settings often lead to mode mixing or over-smoothing, which weakens clinically relevant morphology.

Finally, we compared these conventional approaches with our diffusion-based denoising model. Unlike decomposition-based methods, diffusion denoising learns the mapping from noisy exercise ECG to clean physiological morphology directly from data rather than relying on predefined frequency assumptions. This enables the model to preserve subtle waveform characteristics that are strongly associated with cardiopulmonary function, especially near peak exercise, where traditional filtering methods often fail. Experimental results are summarized in [Figure 16](#). All denoising strategies outperform the setting without denoising, confirming that denoising is an indispensable component for exercise-phase ECG modeling. Our diffusion-based denoising model achieves the best overall performance, representing absolute improvements of around 5%. This improvement is clinically meaningful because prediction accuracy near peak exercise directly affects cardiopulmonary functional assessment.

5.3.5 Impact of Text-Based EHR. We also evaluate CaReFit without electronic health records (EHR). The results for VO_2max and VO_2AT are 12.13% and 11.02%, respectively. The plot is also shown in [Figure 15](#). This indicates that personal information provides complementary information that improves prediction accuracy, although its contribution is smaller than that of the dynamic physiological signals. There are two reasons for this. First, compared to time-series signals, texts contain low-dimensional information, so it provides only a small gain over these two modalities. Second, they only reflect static examinations and cannot assess the overall body condition under stress. Thus, they may miss exercise-induced abnormalities, limiting their correlation with actual functional capacity. In other words, two people with similar static examinations may have different CRFs after exhaustive exercise due to abnormalities. This also explains why dynamic assessment is so significant.

6 RELATED WORK

In this section, we review the studies related to this paper from two perspectives: formula-based oxygen uptake prediction and mobile oxygen uptake prediction systems.

6.1 Formula-Based Oxygen Uptake Prediction

There are some works attempting to predict VO_2max only with steady personal information, like height, weight, sex, age, and perceived functional ability questionnaire [8]. Bruce protocol is a classic method to predict VO_2max , where the participants keep running with progressively increasing the speed and incline of the treadmill. Griffith *et al.* [23] propose a formula with the time of the duration to predict Peak Oxygen Uptake for Pediatric Patients. Another method is the Cooper test, for which Ramirez *et al.* [4] propose an equation based primarily on the distance covered during the 12-minute run. A similar approach is the 20-meter multistage shuttle run test, in which speed increases between cycles. Arsa *et al.* [5] leverage the linear multiple regression only with the maximum speed during the test and age in the predictive equation. However, according to [50], the primary limitation of formula-based prediction is its low accuracy.

6.2 Mobile Oxygen Uptake Prediction System

In addition to formula-based estimation, several studies aim to estimate oxygen uptake using wearable sensors. Altini *et al.* [3] present a system for estimating oxygen consumption during daily activities using a combined accelerometer and heart-rate sensor. Beltrame *et al.* [10] develop the methods to predict and investigate oxygen uptake during activities of daily living based on a random forest method. They utilize a hip accelerometer, an ECG sensor, and a respiration band integrated in a shirt. Similarly, Shandhi *et al.* [39] design a system that combines seismocardiogram (SCG), electrocardiogram (ECG), and atmospheric pressure (AP) signals and extracts features to estimate the VO_2 data using a regression model. Recently, Sheridan *et al.* [41] verify the feasibility of using machine learning models to estimate oxygen uptake via inertial measurement units (IMU), heart rate monitors, and breathing rate sensors. Bangaru *et al.* [7] demonstrate a strong correlation between IMU and EMG features and oxygen uptake, providing an easy way to continuously monitor physiological status. However, most of these works fail to achieve predictions with less than 10% error, which is the criterion for current commercial VO_2 devices. There are some works that attempt to monitor the respiration volume [29, 36, 53], which can be further used to predict the oxygen uptake. And even if they could predict the maximal oxygen uptake VO_2max , it is almost impossible to detect the VO_2AT because they lack the corresponding information.

7 DISCUSSION AND FUTURE WORK

While our method demonstrates competitive accuracy in a clinical cohort, several important limitations and promising avenues for future research warrant discussion. In this section, we outline the limitations of CaReFit and propose directions for future research.

Commercial Comparison and Clinical Utility. The comparison with the commercial VO_2max Master Analyzer provides an important reference for understanding the practical capability of CaReFit. On the healthy-participant cohort, CaReFit achieved a VO_2max MAPE of 15.09%, which is slightly higher than the 14.56% obtained by the commercial device under the same protocol. Moreover, the performance difference between the clinical cohort and the healthy cohort is also noteworthy. Several factors may contribute to this discrepancy. First, the healthy participants span a considerably wider VO_2max range than the clinical cohort, making regression more challenging. Second, the healthy comparison cohort contains only ten participants, making the estimated error more sensitive to individual variation. In addition, the healthy cohort has a different gender composition compared with the clinical cohort, which may further limit direct quantitative comparison. Although CaReFit demonstrates encouraging overall performance, prediction accuracy decreases for participants with relatively low cardiorespiratory fitness. This subgroup is clinically the most important population because individuals with impaired exercise capacity are often the primary targets of rehabilitation and chronic disease management. Consequently, improving prediction robustness for low-fitness populations represents one of the principal limitations of the current system and an important direction for future work.

Generative Modeling for Physiological Data Augmentation. Although we conducted a substantial validation of our system using real-world patient data, our dataset may still lack representation of extreme cases due to insufficient examples. A significant challenge in collecting data for CRF is the requirement for high patient compliance, which can be difficult to achieve. Additionally, the limited number of professionals makes it more challenging to collect authentic data on a large scale. To mitigate data scarcity and enhance model robustness, future work will explore the application of advanced generative artificial intelligence (AI). Specifically, we will investigate synthesizing high-fidelity, multimodal physiological signals—including ECG and muscle oxygenation time series—conditioned on exercise intensity and individual metadata [11, 22, 48, 49]. Such synthetic data, grounded in the biophysical relationships learned from real CPET recordings, could augment training sets, improve model generalization to underrepresented populations, and facilitate the simulation of rare clinical scenarios without additional patient burden.

Prediction from Submaximal Exercise Protocols. The requirement for maximal exertion limits the applicability of both CPET and our current system CaReFit for certain populations, such as the elderly, individuals with disabilities, or patients with cardiopulmonary conditions who may not safely tolerate peak exercise testing. Emerging clinical evidence [13, 17] increasingly supports that cardiorespiratory fitness can be accurately estimated from physiological responses during submaximal exercise. A critical next research step is to extend CaReFit to such protocols by modeling the dynamic, low-intensity relationships between ECG, SmO_2 , and other potentially informative modalities, along with workload or perceived exertion, to extrapolate or directly predict $\text{VO}_{2\text{max}}$ and $\text{VO}_{2\text{AT}}$. Key challenges in this adaptation include accounting for higher inter-individual variability in submaximal responses and ensuring model robustness across heterogeneous populations and exercise protocols. Successfully implementing this approach could significantly broaden the clinical utility and real-world accessibility of CRF assessment, enabling safe and feasible monitoring in a wider range of settings, including primary care, rehabilitation, and remote patient management.

Treadmill Testing Verification. As we mentioned in Section 2.1, CPET typically involves cycling or treadmill exercise with progressively increasing intensity until the participant's maximal capacity. However, our system is only validated on cycling, and future work needs to explore the performance on a treadmill. The biomechanics of running involve greater whole-body vibration than the relatively stable seated pedaling action. This will elevate the baseline level of motion artifact (MA) contaminating physiological signals, particularly for the electrocardiography (ECG) leads. Consequently, the denoising pipeline would require retraining or algorithmic adaptation to handle a more aggressive and high-frequency noise profile. For muscle oxygenation monitoring, since we stick the sensor patch to the quadriceps femoris muscle, the signals should be as stable as in the current scenario. Nevertheless, cycling predominantly engages the quadriceps and gluteal muscles, whereas treadmill running involves more complex coordination of the lower limbs, core, and upper body for balance. This could affect the interpretation of localized muscle oxygenation (SmO_2) signals from the quadriceps. Although we adhere the sensor to the vastus lateralis, its role and metabolic demand profile during running may differ, potentially altering the signal dynamics and their relationship to whole-body $\text{VO}_{2\text{max}}$ and $\text{VO}_{2\text{AT}}$. Validation is needed to confirm if the current regression models generalize or require modality-specific calibration.

Implications for Ubiquitous Sensing Beyond CPET. While CaReFit is developed and validated under a structured CPET protocol, an important question for the mobile and ubiquitous computing community is its applicability in less controlled, real-world environments. The current system relies on a standardized incremental exercise protocol, where workload progression is precisely controlled and physiological responses follow relatively predictable dynamics. In contrast to CPET, daily-life activities are unstructured, with variable intensity, interruptions, and heterogeneous motion patterns. Nevertheless, the core principle of our approach—leveraging the coupled dynamics between cardiac activity (ECG) and muscle metabolic response (SmO_2)—may still hold during sustained moderate-to-high intensity activities (e.g., running or cycling) in free-living settings. This suggests a potential pathway toward opportunistic CRF estimation by identifying suitable activity segments

rather than requiring dedicated testing sessions. However, several challenges remain. First, the absence of controlled workload progression makes it difficult to standardize physiological responses across individuals. Second, real-world sensing introduces more severe motion artifacts and signal quality variability, requiring more robust denoising and quality assessment mechanisms. Third, commodity wearable devices may not provide direct SmO_2 measurements, necessitating the use of surrogate modalities (e.g., PPG), which could reduce estimation accuracy. Finally, real-world data are often sparse and fragmented, requiring models that can operate under incomplete observations. Therefore, we position CaReFit as a step toward bridging clinical-grade CRF assessment and ubiquitous sensing, with future work focusing on relaxing protocol constraints, improving robustness to in-the-wild noise, and exploring modality substitutions for deployment on widely available wearable platforms.

8 CONCLUSION

This paper introduces CaReFit, a novel and exploratory mobile system for the assessment of cardiorespiratory fitness (CRF) that leverages convenient, low-cost wearable sensors. Our methodology is predicated on the acquisition of electrocardiogram (ECG) and muscle oxygen saturation (SmO_2) signals—two key physiological modalities with established correlations to CRF metrics. To address motion artifacts and electrode noise in mobile ECG, we adopt a diffusion-based deep learning model for robust denoising, trained on a composite dataset without manual feature engineering. To capture the nonlinear relationship between multimodal signals and oxygen consumption, we design a neural network with modality-specific encoders and an attention-based fusion module, enabling rich cross-modal representations and improved predictive accuracy without additional hardware. We evaluate CaReFit in a clinical study with 50 patients, achieving mean absolute percentage errors of 9.30% (95% CI: 7.06–11.80) for VO_2max and 9.49% (95% CI: 7.02–12.12) for VO_2AT under a subject-disjoint protocol. We open-source the dataset of synchronized ECG and SmO_2 signals collected during CPET.

Acknowledgments

The authors would like to thank the anonymous editors and reviewers for the valuable comments and suggestions. This research is supported in part by RGC under Contract CERG 16205824, 16204523, SZSTI25EG20, SRFS2425-6S05, AoE/E-601/22-R, and HB016. Dr. Qian Zhang and Dr. Shifang Yang are the co-corresponding authors.

References

- [1] Mehmet Akay, Fatih Abut, Kıymet Kaya, Ebru Cetin, and Imdat Yarim. 2016. New regression equations for estimating the maximal oxygen uptake of college of physical education and sports students in Turkey. *New Trends and Issues Proceedings on Humanities and Social Sciences* (2016).
- [2] Khaled Albouaini, Mohaned Egred, Albert Alahmar, and David Justin Wright. 2007. Cardiopulmonary exercise testing and its application. *Postgraduate medical journal* 83, 985 (2007), 675–682.
- [3] Marco Altini, Julien Penders, and Oliver Amft. 2015. Estimating oxygen uptake during nonsteady-state activities and transitions using wearable sensors. *IEEE journal of biomedical and health informatics* 20, 2 (2015), 469–475.
- [4] José Alvarez-Ramirez and Eduardo Rodriguez. 2021. Theoretical analysis of the 12 min Cooper’s test to estimate the maximal oxygen uptake rate. *Biomedical Signal Processing and Control* 69 (2021), 102885.
- [5] Gisela Arsa, Fernanda Cordoba Lanza, Lucieli Teresa Cambri, Ednei Luiz Antonio, Neif Murad, Marco Tulio de Mello, Alexandra Alberto Santos, Paulo de Tarso Camillo de Carvalho, Jose Antonio Silva-Junior, Danilo Sales Bocalini, et al. 2018. Predicted equation for VO_2 based on a 20-meter multistage shuttle run test for children. *International journal of sports medicine* 39, 14 (2018), 1049–1054.
- [6] Atiqah Ashfaq, Neil Cronin, and Philipp Müller. 2022. Recent advances in machine learning for maximal oxygen uptake (VO_2max) prediction: A review. *Informatics in Medicine Unlocked* 28 (2022), 100863.
- [7] Srikanth Sagar Bangaru, Chao Wang, Fereydoun Aghazadeh, Shashank Muley, and Suede Willoughby. 2025. Oxygen Uptake Prediction for Timely Construction Worker Fatigue Monitoring Through Wearable Sensing Data Fusion. *Sensors* 25, 10 (2025), 3204.
- [8] Alan Carlos Brisola Barbosa, Evandro Fornias Sperandio, Barbara de Barros Gonze, Giovanna Domingues Spina, Rodolfo Leite Arantes, Antônio Ricardo de Toledo Gagliardi, Marcelo Romiti, and Victor Zuniga Dourado. 2018. A new equation to predict peak VO_2 in obese patients during cardiopulmonary exercise testing. *Clinical Physiology and Functional Imaging* 38, 3 (2018), 462–467.

- [9] Arlie R. Barnes, Harold Ensign Bennett Pardee, Paul Dudley White, Frank N. Wilson, and Charles Christian Wolferth. 1938. Standardization of precordial leads. *American Heart Journal* 15 (1938), 235–239. <https://api.semanticscholar.org/CorpusID:56682176>
- [10] Thomas Beltrame, Robert Amelard, Alexander Wong, and Richard L Hughson. 2017. Prediction of oxygen uptake dynamics by machine learning analysis of wearable sensors during activities of daily living. *Scientific reports* 7, 1 (2017), 45738.
- [11] Laurenz Berger, Max Haberbush, and Francesco Moscato. 2023. Generative adversarial networks in electrocardiogram synthesis: Recent developments and challenges. *Artificial Intelligence in Medicine* 143 (2023), 102632.
- [12] Yetong Cao, Fan Li, Huijie Chen, Xiaochen Liu, Li Zhang, and Yu Wang. 2022. Guard your heart silently: Continuous electrocardiogram waveform monitoring with wrist-worn motion sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–29.
- [13] Jeff Coombes and Chris Askew. 2014. Submaximal tests for cardiorespiratory fitness. *ESSA's student manual for health, exercise and sport assessment* (2014), 219–235.
- [14] Kenneth H Cooper. 1968. A means of assessing maximal oxygen intake: correlation between field and treadmill testing. *Jama* 203, 3 (1968), 201–204.
- [15] David L Donoho. 1995. De-noising by soft-thresholding. *IEEE transactions on information theory* 41, 3 (1995), 613–627.
- [16] Konstantin Dragomiretskiy and Dominique Zosso. 2013. Variational mode decomposition. *IEEE transactions on signal processing* 62, 3 (2013), 531–544.
- [17] Marc-Olivier Dugas, Pénlope Paradis-Deschênes, Laurie Simard, Tommy Chevrete, Patricia Blackburn, and Martin Lavallière. 2023. Comparison of VO₂max estimations for maximal and submaximal exercise tests in apparently healthy adults. *Sports* 11, 12 (2023), 235.
- [18] Andri Feldmann, Linda Ammann, Flurin Gächter, Marc Zibung, and Daniel Erlacher. 2022. Muscle oxygen saturation breakpoints reflect ventilatory thresholds in both cycling and running. *Journal of human kinetics* 83 (2022), 87.
- [19] Andri Feldmann, Roger Schmitz, and Daniel Erlacher. 2019. Near-infrared spectroscopy-derived muscle oxygen saturation on a 0% to 100% scale: reliability and validity of the Moxy Monitor. *Journal of biomedical optics* 24, 11 (2019), 115001–115001.
- [20] Marco Ferrari, Tiziano Binzoni, and Valentina Quaresima. 1997. Oxidative metabolism in muscle. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 352, 1354 (1997), 677–683.
- [21] Catarina Goncalves, Armando Raimundo, Ana Abreu, and Jorge Bravo. 2021. Exercise intensity in patients with cardiovascular diseases: systematic review with meta-analysis. *International journal of environmental research and public health* 18, 7 (2021), 3574.
- [22] Aldren Gonzales, Guruprabha Guruswamy, and Scott R Smith. 2023. Synthetic data in health care: A narrative review. *PLOS Digital Health* 2, 1 (2023), e0000082.
- [23] Garrett J Griffith, Alan P Wang, Robert I Liem, Michael R Carr, Tyler Corson, and Kendra Ward. 2024. A reference equation for peak oxygen uptake for pediatric patients who undergo treadmill cardiopulmonary exercise testing. *The American Journal of Cardiology* 212 (2024), 41–47.
- [24] Artur Haddad Herdy, Luiz Eduardo Fonteles Ritt, Ricardo Stein, Claudio Gil Soares de Araújo, Mauricio Milani, Romeu Sérgio Meneghelo, Almir Sérgio Ferraz, Carlos Hossri, Antonio Eduardo Monteiro de Almeida, Miguel Morita Fernandes-Silva, et al. 2016. Cardiopulmonary exercise test: background, applicability and interpretation. *Arquivos brasileiros de cardiologia* 107, 5 (2016), 467–481.
- [25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [26] Leonard A Kaminsky, Ross Arena, Øyvind Ellingsen, Matthew P Harber, Jonathan Myers, Cemal Ozemek, and Robert Ross. 2019. Cardiorespiratory fitness and cardiovascular disease—the past, present, and future. *Progress in cardiovascular diseases* 62, 2 (2019), 86–93.
- [27] Elke Kothmann, Gerard Danjoux, SJ Owen, A Parry, AJ Turley, and AM Batterham. 2009. Reliability of the anaerobic threshold in cardiopulmonary exercise testing of patients with abdominal aortic aneurysms. *Anaesthesia* 64, 1 (2009), 9–13.
- [28] Huayu Li, Gregory Ditzler, Janet Roveda, and Ao Li. 2023. Descod-ecg: Deep score-based diffusion model for ecg baseline wander and noise removal. *IEEE Journal of Biomedical and Health Informatics* 28, 9 (2023), 5081–5091.
- [29] Yang Liu, Qiang Yang, Kayla-Jade Butkow, Jake Stuchbury-Wass, Dong Ma, and Cecilia Mascolo. 2025. EarMeter: Continuous Respiration Volume Monitoring with Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 4 (2025), 1–29.
- [30] George Moody. 1999. MIT-BIH Normal Sinus Rhythm Database. <https://www.physionet.org/content/nsrdb/1.0.0/> Accessed: 2026-01-27.
- [31] George Moody and Roger Mark. 1999. MIT-BIH Noise Stress Test Database. <https://www.physionet.org/content/nstadb/1.0.0/> Accessed: 2026-01-27.
- [32] C Myers. 2020. Muscle oxygenation applications to endurance training. *Anatomy & Physiology: Open Access Journal* 1, 1 (2020).
- [33] Antonio Pelliccia, Sanjay Sharma, Sabiha Gati, Maria Bäck, Mats Börjesson, Stefano Caselli, Jean-Philippe Collet, Domenico Corrado, Jonathan A Drezner, Martin Halle, et al. 2021. 2020 ESC Guidelines on sports cardiology and exercise in patients with cardiovascular disease: The Task Force on sports cardiology and exercise in patients with cardiovascular disease of the European Society of Cardiology (ESC). *European heart journal* 42, 1 (2021), 17–96.
- [34] James E Peterman, Matthew P Harber, Mary T Imboden, Mitchell H Whaley, Bradley S Fleenor, Jonathan Myers, Ross Arena, W Holmes Finch, and Leonard A Kaminsky. 2020. Accuracy of nonexercise prediction equations for assessing longitudinal changes to cardiorespiratory fitness in apparently healthy adults: BALL ST cohort. *Journal of the American Heart Association* 9, 11 (2020), e015117.

- [35] José Riesco-Villar, Aarón Escribano-Pascual, Alejandro González-Fernández, Héctor Pascual-Redondo, Marina Gil-Calvo, and Daniel Boullosa. 2025. Neuromuscular, cardiometabolic, and perceptual responses to a short sprint interval training session with different cycling torques. *European Journal of Applied Physiology* 125, 9 (2025), 2461–2475.
- [36] Julia Romero, Andrea Ferlini, Dimitris Spathis, Ting Dang, Katayoun Farrahi, Fahim Kawsar, and Alessandro Montanari. 2024. OptiBreathe: An earable-based PPG system for continuous respiration rate, breathing phase, and tidal volume monitoring. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*. 99–106.
- [37] Robert Ross, Steven N Blair, Ross Arena, Timothy S Church, Jean-Pierre Després, Barry A Franklin, William L Haskell, Leonard A Kaminsky, Benjamin D Levine, Carl J Lavie, et al. 2016. Importance of assessing cardiorespiratory fitness in clinical practice: a case for fitness as a clinical vital sign: a scientific statement from the American Heart Association. *Circulation* 134, 24 (2016), e653–e699.
- [38] Carlos Sendra-Perez, Jose Luis Sanchez-Jimenez, Joaquin Martin Marzano-Felisatti, Alberto Encarnacion-Martinez, Rosario Salvador-Palmer, and Jose I Priego-Quesada. 2023. Reliability of threshold determination using portable muscle oxygenation monitors during exercise testing: a systematic review and meta-analysis. *Scientific Reports* 13, 1 (2023), 12649.
- [39] Md Mobashir Hasan Shandhi, William H Bartlett, James Alex Heller, Mozziyar Etemadi, Aaron Young, Thomas Plötz, and Omer T Inan. 2020. Estimation of instantaneous oxygen uptake during exercise and daily activities using a wearable cardio-electromechanical and environmental sensor. *IEEE journal of biomedical and health informatics* 25, 3 (2020), 634–646.
- [40] Salaheddin Sharif and Stephen E Alway. 2016. The diagnostic value of exercise stress testing for cardiovascular disease is more than just st segment changes: A review. *J Integr Cardiol* 2, 4 (2016), 341–355.
- [41] Dermot Sheridan, Arne Jaspers, Dinh Viet Cuong, Tim Op De Beëck, Niall M Moyna, Toon T de Beukelaar, and Mark Roantree. 2025. Estimating oxygen uptake in simulated team sports using machine learning models and wearable sensor data: A pilot study. *Plos One* 20, 4 (2025), e0319760.
- [42] Ken-ichi Shibuya, Junya Tanaka, and Tetsuro Ogaki. 2004. Muscle oxygenation kinetics at the onset of exercise do not depend on exercise intensity. *European journal of applied physiology* 91 (2004), 712–715.
- [43] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. <https://openreview.net/forum?id=St1giarCHLP>
- [44] Maria E Torres, Marcelo A Colominas, Gaston Schlotthauer, and Patrick Flandrin. 2011. A complete ensemble empirical mode decomposition with adaptive noise. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4144–4147.
- [45] Niels Uth, Henrik Sørensen, Kristian Overgaard, and Preben K Pedersen. 2004. Estimation of VO₂max from the ratio between HR max and HR rest—the Heart Rate Ratio Method. *European journal of applied physiology* 91 (2004), 111–115.
- [46] CK Vaishnavi, V Sricharan, G Sri Gayathri, N Danush Adhithya, SP Preejith, et al. 2024. Optimizing VO₂ max Prediction in Gamified Cardiac Assessment: Leveraging Effective Feature Selection and Refined Protocols for Robust Models. In *2024 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE, 266–271.
- [47] Bas Van Hooren, Tjeu Souren, and Bart C Bongers. 2024. Accuracy of respiratory gas variables, substrate, and energy use from 15 CPET systems during simulated and human exercise. *Scandinavian Journal of Medicine & Science in Sports* 34, 1 (2024), e14490.
- [48] Lei Wang, Xingwei Wang, Dalin Zhang, Xiaolei Ma, Yong Zhang, Haipeng Dai, Chenren Xu, Zhijun Li, and Tao Gu. 2023. Knowing your heart condition anytime: User-independent ECG measurement using commercial mobile phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–28.
- [49] Zhi Wang, Beihong Jin, Siheng Li, Fusang Zhang, and Wenbo Zhang. 2023. ECG-grained cardiac monitoring using uwb signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–25.
- [50] John R Wicks and Neil B Oldridge. 2016. How accurate is the prediction of maximal oxygen uptake with treadmill testing? *PLoS One* 11, 11 (2016), e0166608.
- [51] Szczepan Wiecha, Przemysław Seweryn Kasiak, Igor Cieśliński, Tim Takken, Tomasz Palka, Beat Knechtle, Pantelis T Nikolaidis, Łukasz A Małek, Marek Postuła, Artur Mamcarz, et al. 2023. External validation of VO₂max prediction models based on recreational and elite endurance athletes. *PLoS One* 18, 1 (2023), e0280897.
- [52] Liangliang Xiang, Kaili Deng, Qichang Mei, Zixiang Gao, Tao Yang, Alan Wang, Justin Fernandez, and Yaodong Gu. 2022. Population and age-based cardiorespiratory fitness level investigation and automatic prediction. *Frontiers in Cardiovascular Medicine* 8 (2022), 758589.
- [53] Wentao Xie, Chi Xu, Yanbin Gong, Yu Wang, Yuxin Liu, Jin Zhang, Qian Zhang, Zeguang Zheng, and Shifang Yang. 2024. DeepBreath: Breathing Exercise Assessment with a Depth Camera. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–26.
- [54] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. 2018. Representation learning on graphs with jumping knowledge networks. In *International conference on machine learning*. 5453–5462.
- [55] Jhang Yu-Syuan. 2022. MIT-BIH NSRDB with NSTDB EM Noise. <https://www.kaggle.com/datasets/yusyuanjhang/mitbih-nsrdb-with-nstadb-em-noise> Accessed: 2026-01-27.

Appendix

A PRELIMINARY SUBJECTIVE FEEDBACK

In addition to the technical evaluation, we conducted a preliminary qualitative assessment to understand user perceptions of CaReFit compared to conventional CPET equipment. We interviewed 5 patients who had previously undergone standard CPET and 3 physicians specializing in cardiopulmonary rehabilitation. The interviews were semi-structured and focused on comfort, usability, and workflow. The questions are listed below.

For Patients:

- Q1: How do you compare the physical comfort and convenience of the CaReFit system to the traditional CPET setup?
- Q2: To what extent does the equipment affect your ability to perform exercise naturally?
- Q3: How does the overall testing experience influence your willingness to undergo future fitness assessments?

For Physicians:

- Q1: How does the CaReFit system affect the clinical workflow, particularly in terms of preparation time and procedural complexity?
- Q2: To what degree does the system allow you to focus on patient performance and physiological data during testing?
- Q3: How does the system compare to traditional CPET in terms of practical utility in clinical or rehabilitation settings?

Overall, responses consistently highlight two main themes: improved patient comfort and reduced clinical workload. Patients report that the lightweight sensing setup significantly lowers physical burden and enables more natural movement during exercise. For example, one participant notes: *"It is much easier to wear and almost without any burden during the test"* (P1), while another reports: *"After removing all heavy equipment, especially the 12-lead ECG during cycling, I feel I can perform better throughout the test"* (P3). These responses suggest that reducing instrumentation complexity may improve compliance and willingness to undergo repeated assessments.

From the clinicians' perspective, the primary benefit lies in workflow efficiency and reduced preparation time. Compared to conventional CPET setups, which typically require over 20 minutes of preparation, CaReFit can be deployed in less than 5 minutes. Physicians also report that the simplified setup allows greater focus on patient monitoring rather than equipment management. As one clinician describes, *"It removes complicated steps in the preparatory phase and lets me focus more on the patient's key physiological indicators during cycling"* (P2).

We emphasize that these findings are preliminary and based on a small sample size without formal qualitative coding or quantitative measures. Therefore, they should be interpreted as indicative rather than conclusive. A more rigorous human-centered evaluation, including larger cohorts and systematic analysis (e.g., thematic coding or standardized usability metrics), is left for future work.

B EQUATIONS AND IMPLEMENTATION

B.1 Mathematical Expression for DDPM and Its Transformation under ECG Conditions

- Forward/Diffusion Process. The forward process is a predefined Markov chain that transforms a clean ECG signal x_0 into pure Gaussian noise over T discrete time steps. At each step t , Gaussian noise with variance β_t is added to the signal, where β_t follows a monotonically increasing schedule from $\beta_1 = 0.0001$ to $\beta_T = 0.5$. This carefully designed progression ensures that, after a sufficient number of steps, the signal distribution converges to an isotropic Gaussian distribution while maintaining mathematical tractability.

$$q(x_t|x_{t-1}) := \mathcal{N}\left(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}\right) \quad (\text{B.1})$$

where β_t is a variance schedule increasing from 0.0001 to 0.5, ensuring x_T approximates isotropic Gaussian noise.

A crucial mathematical convenience of this formulation is that one can sample x_t directly from x_0 without iterating through intermediate steps:

$$x_t := \sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \quad (\text{B.2})$$

with $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$.

- Reverse/Denoising process.

The reverse process is a learned Markov chain that aims to reconstruct the original signal x_0 starting from Gaussian noise X_T . At each step t , the model predicts the distribution of x_{t-1} given x_t starting from Gaussian noise $x_t \sim \mathcal{N}(0, I)$:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)\mathbf{I}) \quad (\text{B.3})$$

where the mean μ_θ is learned by a neural network ϵ_θ that predicts the added noise as below:

$$\mu_\theta(x_t, t) := \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) \quad (\text{B.4})$$

- Conditional extension for ECG denoising. For ECG noise removal, the process is conditioned on the noisy observation $\tilde{x}$:

$$p_\theta(x_{t-1}|x_t, \tilde{x}) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t|\tilde{x}), \sigma_\theta(x_t, t|\tilde{x})\mathbf{I}) \quad (\text{B.5})$$

This allows the model to reconstruct clean signals x_0 from random noise while being guided by the noisy input.

- Basic training goal. The model is trained to minimize the discrepancy between the true noise added during the forward process and the noise predicted by the network. The simplified training loss is:

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t, x_0, \epsilon} [\|\epsilon - \epsilon_\theta(\sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon, t)\|] \quad (\text{B.6})$$

t is uniformly sampled from $1, \dots, T$, x_0 is a clean ECG segment from the training set, and ϵ is standard Gaussian noise. The network ϵ_θ learns to predict the noise component present in x_t , which indirectly enables accurate estimation of μ_θ . This objective is widely adopted due to its simplicity, stability, and strong empirical performance in both unconditional and conditional generation tasks.

B.2 Filter Block Expression

$$\begin{aligned} \mathbf{F}_k &= \text{Conv1D}_k(\mathbf{F}) \\ \mathbf{F}_{\text{concat}} &= [\mathbf{F}_0, \mathbf{F}_1, \dots] \in \mathbb{R}^{B \times C' \times T} \\ \mathbf{F}_{\text{fused}} &= \text{Conv1D}_1(\mathbf{F}_{\text{concat}}) \in \mathbb{R}^{B \times C \times T} \\ \mathbf{Q}, \mathbf{K}, \mathbf{V} &= \mathbf{F}_{\text{fused}}^T W_q, \mathbf{F}_{\text{fused}}^T W_k, \mathbf{F}_{\text{fused}}^T W_v \\ \mathbf{Attention} &= \text{Softmax}(\mathbf{QK}^T / \sqrt{d_k}) \in \mathbb{R}^{B \times T \times T} \\ \mathbf{F}_{\text{attn}} &= \mathbf{F}_{\text{fused}} + (\eta \mathbf{Attention} \times \mathbf{V})^T \in \mathbb{R}^{B \times C \times T} \\ \mathbf{F}_{\text{out}} &= \mathbf{F} + \text{LayerNorm}(\mathbf{F}_{\text{attn}}) \end{aligned} \quad (\text{B.7})$$

B.3 Akay *et al.* Multivariate Regression

$$\begin{aligned} VO_2max = & -12.331 \times \text{gender} - 0.805 \times \text{age} \\ & + 0.883 \times \text{height} - 1.167 \times \text{weight} \\ & - 0.052 \times \text{HR}_{\max} - 0.158 \times \text{TT} \\ & + 6.473. \end{aligned} \quad (\text{B.8})$$

B.4 Cross Attention Calculation

$$\text{attn_scores}[b, m, n] = \sum_{d=1}^D Q[m, b, d] \cdot K[n, b, d] \quad (\text{B.9})$$

$$\text{attention_weight} = \text{Softmax}(\text{attn_scores} / \sqrt{D_model})$$

$$\text{fusion}[m, b, d] = \sum_{n=0}^1 \text{attention_weight}[b, m, n] \cdot V[n, b, d] \quad (\text{B.10})$$

B.5 Denoising

B.5.1 Wavelet Denoising. Given a one-dimensional ECG signal $x \in \mathbb{R}^N$, wavelet denoising was applied to suppress high-frequency noise while preserving the main morphological characteristics of the ECG waveform. In this study, the signal was decomposed using the Daubechies 6 (db6) wavelet, which provides good time-frequency localization and is suitable for representing transient ECG components such as the QRS complex. The decomposition level L was automatically selected according to the signal length and the maximum allowable wavelet decomposition level:

$$x \rightarrow \{c_A^{(L)}, c_D^{(1)}, c_D^{(2)}, \dots, c_D^{(L)}\}, \quad (\text{B.11})$$

where $c_A^{(L)}$ denotes the approximation coefficient at level L , and $c_D^{(l)}$ denotes the detail coefficient at level l .

The approximation coefficient mainly contains the low-frequency components of the ECG signal, including baseline trends and slow physiological variations. Therefore, it was preserved during the denoising process to avoid excessive distortion of the original ECG morphology. For each set of detail coefficients $c_D^{(l)}$, the noise level was estimated using the median absolute deviation (MAD):

$$\sigma_l = \frac{\text{median}(|c_D^{(l)}|)}{0.6745}. \quad (\text{B.12})$$

Based on the estimated noise level, a level-dependent threshold was calculated as:

$$T_l = \alpha \sigma_l \sqrt{2 \log(n_l)}, \quad (\text{B.13})$$

where α is the threshold scaling factor, and n_l is the number of coefficients at level l . A larger value of α results in stronger denoising, while a smaller value preserves more signal details.

Soft-thresholding was then applied to the detail coefficients:

$$\hat{c}_D^{(l)} = \text{sign}(c_D^{(l)}) \max(|c_D^{(l)}| - T_l, 0). \quad (\text{B.14})$$

This operation reduces small-amplitude coefficients that are likely caused by noise, while retaining larger coefficients associated with important ECG structures.

Finally, the denoised ECG signal $\hat{x}$ was reconstructed using the inverse discrete wavelet transform:

$$\hat{x} = \text{IDWT} \left(c_A^{(L)}, \hat{c}_D^{(1)}, \hat{c}_D^{(2)}, \dots, \hat{c}_D^{(L)} \right). \quad (\text{B.15})$$

The reconstructed signal was truncated to the original signal length to ensure that the output had the same dimension as the input. This wavelet-based approach effectively reduces noise while preserving clinically relevant ECG features such as P waves, QRS complexes, and T waves.

B.5.2 EEMD. Some noise components can be attenuated using conventional signal-processing techniques. For example, baseline drift and high-frequency interference are removed by applying a 0.5–40 Hz band-pass filter. Spike outliers arising from circuit anomalies are detected with a Hampel filter and replaced by the local median. Minor fluctuations are smoothed with a mean filter. Nevertheless, these simple processing techniques can hardly remove all the noise. We therefore apply *ensemble empirical mode decomposition* (EEMD) [44] to separate noise from the exercise-phase ECG, which offers superior decomposition for the highly non-stationary, non-linear signals encountered during vigorous exercise. For each of N noisy versions of the original signal, we add random zero-mean Gaussian noise $n_i(t)$ to the pre-filtered trace $S(t)$ as expressed in Equation B.16:

$$x_i(t) = S(t) + n_i(t). \quad i = 1, \dots, N \quad (\text{B.16})$$

Then, we decompose each noise-added signal $x_i(t)$ via the standard Empirical Mode Decomposition (EMD) sifting process, yielding a set of intrinsic mode functions (IMFs), $\text{IMF}_{i,j}(t)$, and a residual component $r_i(t)$, as shown in Equation B.17:

$$x_i(t) = \sum_{j=1}^{K_i} \text{IMF}_{i,j}(t) + r_i(t), \quad (\text{B.17})$$

where K_i is the total number of IMFs obtained for each surrogate.

Next, we ensemble-average the corresponding IMFs across all N realizations to obtain the final averaged mode, $\text{IMF}_j(t)$, represented by Equation B.18:

$$\text{IMF}_j(t) = \frac{1}{N} \sum_{i=1}^N \text{IMF}_{i,j}(t). \quad (\text{B.18})$$

Instead of recombining several modes, we retain a single $\text{IMF}_{k_*}(t)$, whose morphology preserves the dominant P, Q, R, and T peaks; this mode is regarded as the clean ECG as follows in Equation B.19,

$$S_{\text{clean}}(t) = \text{IMF}_{k_*}(t). \quad (\text{B.19})$$

B.5.3 VMD. First, the dominant frequency of each input signal was estimated via a real-valued Fast Fourier Transform (`dominant_freq_hz`) to identify the primary oscillatory component while excluding the DC offset. The signal was mean-centered prior to spectral analysis, and the frequency corresponding to the maximum peak in the amplitude spectrum (excluding the zero-frequency component) was selected as the dominant frequency. This step provided an initial characterization of the signal's oscillatory behavior and served as a reference for subsequent decomposition.

Subsequently, Variational Mode Decomposition (VMD) was applied to decompose the signal into K intrinsic mode functions (IMFs), each representing a narrow-band oscillatory component with an adaptive center frequency. Unlike traditional empirical mode decomposition, VMD formulates the decomposition as a constrained variational optimization problem, which improves robustness against mode mixing and endpoint effects. The parameter α controls the bandwidth constraint of each mode, where larger values enforce narrower-band components, while τ regulates the noise-tolerance of the reconstruction fidelity term. Additional parameters, including initialization strategy and convergence tolerance, were set to ensure numerical stability and reproducibility. After decomposition, the final center frequencies of all IMFs were obtained from the converged VMD solution. These normalized frequencies were converted into physical frequency units (Hz) according to the sampling frequency f_s . Only IMFs whose center frequencies were above a predefined physiological threshold were retained, while

low-frequency components associated with baseline drift, motion artifacts, and slow nonstationary trends were discarded. The denoised signal was then reconstructed by summing the selected IMFs.

This frequency-selective reconstruction strategy preserves physiologically meaningful oscillatory patterns, particularly those associated with cardiovascular and respiratory dynamics, while effectively suppressing irrelevant low-frequency interference and high-frequency noise. As a result, VMD-based denoising improves signal quality and enhances the reliability of subsequent feature extraction and predictive modeling stages.