From Lab to Wrist: Bridging Metabolic Monitoring and Consumer Wearables for Heart Rate and Oxygen Consumption Modeling

Barak Gahtan
Computer Science
Technion, Israel Institute of Technology
Haifa, Israel
barakgahtan@cs.technion.ac.il

Sanketh Vedula
Computer Science
Technion - Israel Institute of Technology
Haifa, Israel
sanketh@campus.technion.ac.il

Gil Samuelly Leichtag
Physical Therapy University of Haifa
Haifa, Israel
gilsamuelly@gmail.com

Einat Kodesh
Physical Therapy University of Haifa
Haifa, Israel
ekodesh@univ.haifa.ac.il

Alex Bronstein
Computer Science
Technion
Haifa, Haifa, Israel
ISTA Institute of Science and Technology
Vienna, Austria
bron@cs.technion.ac.il

Abstract

Understanding physiological responses during running is critical for performance optimization, tailored training prescriptions, and athlete health management. We introduce a comprehensive framework—what we believe to be the first capable of predicting instantaneous oxygen consumption (VO_2) trajectories exclusively from consumer-grade wearable data. Our approach employs two complementary physiological models: (1) accurate modeling of heart rate (HR) dynamics via a physiologically constrained ordinary differential equation (ODE) and neural Kalman filter, trained on over 3 million HR observations, achieving 1-second interval predictions with mean absolute errors as low as 2.81 bpm (correlation 0.87); and (2) leveraging the principles of precise HR modeling, a novel VO_2 prediction architecture requiring only the initial second of VO_2 data for calibration, enabling robust, sequence-to-sequence metabolic demand estimation. Despite relying solely on smartwatch and chest-strap data, our method achieves mean absolute percentage errors of approximately 13%, effectively capturing rapid physiological transitions and steady-state conditions across diverse running intensities. Our synchronized dataset, complemented by blood lactate measurements, further lays the foundation for future noninvasive metabolic zone identification. By embedding physiological constraints within modern machine learning, this framework democratizes advanced metabolic monitoring, bridging laboratory-grade accuracy and everyday accessibility, thus empowering both elite athletes and recreational fitness enthusiasts.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Neural networks**; • **Human-centered computing** → *Human computer interaction (HCI)*; • **Applied computing** → **Health informatics**; **Consumer health**.

Keywords

Modeling Physiological Responses during Exercise, Machine Learning, Oxygen Consumption Prediction, Heart Rate Prediction, Multi-Modal Dataset, Neural Kalman Filter, Neural ODE, Instantaneous VO2 Prediction

ACM Reference Format:
Barak Gahtan, Sanketh Vedula, Gil Samuelly Leichtag, Einat Kodesh, and Alex Bronstein. 2025. From Lab to Wrist: Bridging Metabolic Monitoring and Consumer Wearables for Heart Rate and Oxygen Consumption Modeling. In *Proceedings of the 27th International Conference on Multimodal Interaction (ICMI '25), October 13–17, 2025, Canberra, ACT, Australia.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3716553.3750815

1 Introduction

Heart rate (HR) and oxygen consumption (VO_2) jointly reveal cardiovascular performance and metabolic demand, making them essential for optimizing athletic performance, preventing overtraining, and safeguarding health for both elite athletes and recreational fitness enthusiasts [24, 29, 43]. While HR is readily accessible via consumer-grade wearables, VO_2 measurement remains confined to specialized laboratory equipment (costing over $30 000) and expert supervision [41], limiting broad access to critical metabolic insights.

To bridge this gap, we propose two complementary physiological models that rely exclusively on data from consumer-grade wearables. First, we introduce an advanced HR dynamics model based on neural ordinary differential equations (ODEs) [6] and with neural Kalman filtering [20] to capture cardiac responses to exercise intensity. This approach precisely infers HR from external parameters, compensates for photoplethysmography (PPG) signal

ICMI '25, Canberra, ACT, Australia

ACM ISBN 979-8-4007-1499-3/25/10
https://doi.org/10.1145/3716553.3750815

loss, and reproduces cardiac behavior during varied exercise efforts. Second, and more significantly, we present, to the best of our knowledge, the first approach for direct estimation of instantaneous VO_2 trajectories from smartwatch data using a Kalman-inspired deep learning architecture [18]. By requiring only the initial second of VO_2 data for calibration, our framework enables sophisticated metabolic monitoring without specialized laboratory equipment, a capability previously unavailable through consumer devices.

Our framework advances physiological monitoring in several key ways. We address three main barriers limiting consumer wearables: (i) continuous HR availability despite PPG drop-outs during high-intensity motion, (ii) power-efficient "generative sequencing" that synthesizes an entire HR stream from a single-second ECG seed, and (iii) the first laboratory-grade, second-by-second VO_2 trajectory estimation without the $30 000 equipment barrier. We achieve a mean absolute error (MAE) of 2.81 bpm (correlation 0.87) in 1-second HR predictions across diverse running conditions, demonstrating robustness to wearable signal dropouts. Building on this foundation, our VO_2 model generates complete metabolic trajectories with a mean absolute percentage error (MAPE) of approximately 13%, accurately capturing both rapid transitions and steady-state conditions. We validate these results through leave-one-runner-out cross-validation against gold-standard portable Cosmed K5 measurements, showing strong generalization across individuals. To achieve this, we collected a first-of-its-kind rich synchronized dataset that simultaneously acquires data from consumer smartwatch (Garmin 965), chest-strap HR, portable Cosmed K5 metabolic system, video capture every 200m, and blood lactate measurements, which also lays the groundwork for future noninvasive metabolic zone classification.

To support HR modeling, we leveraged a comprehensive dataset of 831 running sessions from 20 participants (Section A), representing over 52,825 minutes (approximately 880 hours) of activity and 10,222.90km of cumulative distance; HR models were evaluated via leave-three-runner-out cross-validation against concurrently recorded smartwatch and chest-strap data. For VO_2 prediction, we conducted an IRB-approved clinical trial with ten highly trained runners, each completing two structured track sessions; VO_2 models were validated using leave-one-runner-out cross-validation against gold-standard Cosmed K5 measurements. All device streams including Garmin 965, chest HR strap, portable metabolic system, video captures every 200m, and blood lactate samples—were time-synchronized to ensure robust multimodal evaluation across individuals and exercise intensities.

By embedding physiology-informed constraints within modern machine learning (ML), our approach extracts laboratory-quality insights from everyday wearables. The generalizable principles emerging from this work include the importance of synchronized multimodal data collection for robust cross-modality modeling and a transferable architecture that democratizes advanced metabolic monitoring on consumer devices.

The rest of this paper is organized as follows: Section 2 reviews related work; Section 3 describes our data and protocol; Section 4 details HR dynamics modeling; Section 5 presents VO_2 prediction; and Section 6 concludes with limitations and future directions.

2 Related Work

Physiological parameter estimation during exercise spans exercise physiology and computer science. We review literature across three interconnected domains—physiological modeling, wearable technology, and ML, before highlighting our contributions.

Physiological Modeling of Exercise Response VO_2 kinetics modeling began with Hill and Lupton's work linking exercise intensity to oxygen uptake [29], extended through compartmental gas-exchange models [23] and the two-component framework (fast/slow VO_2 responses) [32], with refinements adding time delays [8] and intensity-dependent parameters [4]. Recent studies emphasize inter-individual variability in VO_2 kinetics across age, training status, and health conditions [12, 13, 26], but remain constrained by controlled lab protocols that fail to capture real-world variability in intensity, environment, and motivation [1, 34]. VO_2 responses also differ markedly between laboratory and free-living contexts [10], influenced by altitude, humidity, fatigue, nutrition, and emotional state [2, 15]. Although tools like `VO2FITTING` aim to streamline kinetics analysis [45], high-precision models still rely on expensive stationary gas analyzers.

Wearable Technology and Exercise Monitoring Consumer wearables now provide continuous PPG based HR and inertial sensing outside the laboratory [7, 40]. Although recent devices have improved HR measurement accuracy [35], motion artifacts and signal dropouts remain significant challenges during high-intensity activities [28, 42]. Many commercial platforms include proprietary VO_2max estimators, yet these algorithms often lack rigorous validation across diverse populations and exercise conditions [21, 25]. To bridge the gap between consumer-grade sensor outputs and laboratory-grade cardiorespiratory metrics, cross-modal inference techniques have been proposed to translate wearable data into VO_2 and other physiological parameters [3, 5], but they frequently rely on steady-state assumptions or are evaluated on narrow cohorts. Emerging integrations of SpO_2 and ECG tracking show promise for reducing estimation errors [9], yet advanced algorithmic approaches remain essential to overcome the inherent limitations of wearable hardware in real-world exercise scenarios.

ML for Physiological Parameter Estimation Recent years have witnessed a surge in applying ML techniques to physiological data analysis. Traditional approaches used statistical models to relate HR to VO_2 [36], but these models typically assume steady-state conditions and struggle with dynamic exercise.

Deep learning approaches have shown promise in processing multivariate physiological signals [14], with recurrent and convolutional architectures demonstrating particular efficacy for temporal data. Attention mechanisms further boost performance by focusing on relevant signal patterns across various time scales [17], enabling systems to better detect meaningful physiological events and limit false alarms.

Several recent studies have applied hybrid modeling to HR dynamics during exercise [11, 27], demonstrating improved accuracy compared to purely mechanistic or purely data-driven approaches. However, the extension of these methods to VO_2 modeling from wearable data remains largely unexplored, particularly in settings involving variable-intensity exercise and diverse subject populations [39]. **Notably absent from the literature is any method**

capable of predicting instantaneous VO_2 trajectories using only consumer wearables—a gap our work directly addresses.
Contributions and Distinctions from Prior Work. Our research addresses these gaps through two major contributions. First, we advance HR prediction during exercise with novel modeling approaches that differ from recent systems in several significant ways. Second, we introduce the first sequence-to-sequence approach for VO_2 estimation that uses only consumer-wearable data, bridging laboratory precision and real-world accessibility.

Our HR model delivers 1-second predictions versus Nazaret et al.'s 10-second averages [27], capturing rapid cardiovascular transitions during variable-intensity exercise for responsive feedback with greater personalization flexibility. Unlike Nazaret et al., which requires extensive historical data per user, our approach generalizes effectively for new users with no historical data—critical for deployments where prior data is unavailable. This is achieved via our neural ODE framework or Kalman filter architecture, capturing physiological responses while adapting to individual cardiovascular characteristics without extensive calibration.

We enhance the practical applicability of our HR modeling through dual evaluation protocols that reflect realistic usage scenarios. Specifically, we utilize non-overlapping windows to evaluate our model under two distinct conditions: (1) a continuous monitoring condition where the initial HR measurement is known for each window, and (2) a minimal-data condition where only the first second's measurement (the first sample of the first window) is available for the entire exercise session (up to 120 minutes). This second protocol represents a significant challenge as it requires the model to maintain accurate predictions over extended durations with extremely limited initialization data. This approach transforms our sequence-to-sequence prediction into a **generative forecasting model**, demonstrating robust long-term predictive capability with minimal initialization data, albeit with an expected accuracy trade-off that we quantify and analyze.

Building on our HR modeling advances, we introduce the first sequence-to-sequence modeling approach for VO_2 estimation using only data available from consumer wearables. Unlike previous approaches that either (1) rely on steady-state assumptions [36] which fail during variable intensity exercise, or (2) require specialized equipment for direct measurement [41], our method captures the dynamic nature of VO_2 during real-world variable-intensity exercise using only consumer-grade devices. Our VO_2 estimation method employs a Kalman filter architecture that learns shared physiological parameters among runners. Our approach demonstrates significant improvements over baseline methods, achieving mean absolute percentage errors of approximately 13% for most participants across diverse running intensities. The model explicitly accounts for individual differences in physiological parameters through a learnable state-space representation, allowing personalization while maintaining the interpretability necessary for sports science applications.

Together, these advances in HR dynamics modeling and VO_2 estimation form an integrated system that bridges the gap between laboratory-grade physiological assessment and consumer-accessible wearable technology, representing a significant step toward democratizing advanced exercise science for both research and practical applications.

Figure 1: Runners during experimental sessions

3 Experimental Protocol and Data Collection

Our research methodology required comprehensive physiological data capturing both laboratory-grade measurements and consumer wearable outputs. We used two complementary datasets: (1) a synchronized, multimodal dataset collected specifically for VO_2 predictions, providing aligned measurements from both laboratory and consumer devices, and (2) a more extensive historical HR dataset for training robust HR dynamics models, offering the breadth and diversity needed for developing generalizable models. Below, we describe the collection protocols and key characteristics of each dataset.

Synchronized Multimodal Dataset: We implemented a comprehensive protocol to gather synchronized physiological and biomechanical data from experienced runners during controlled exercise tests. Conducted at two athletic facilities, the study involved 10 participants who each completed two structured sessions. We recruited runners aged 20–50 with demonstrated high-level performance (top 1% 10 km race times for their age groups). This focus on highly trained athletes was intended to reduce physiological variance, facilitating more precise modeling of exercise responses. Before participation, each runner obtained medical clearance from a certified sports physician to confirm eligibility for high-intensity testing.

The protocol comprised two sessions targeting different aspects of physiological performance. In Session One (Maximal Capacity Assessment), participants performed a 1500-meter maximal-effort run to determine VO_2max. We measured metabolic data with a portable Cosmed K5 system, capturing O_2 and CO_2 exchange rates, while simultaneously recording HR and biomechanical data via a Garmin 965 smartwatch. VO_2max was calculated as the 30-second peak average VO_2, and mean running speed was computed for the entire distance. In Session Two (Incremental Testing), participants followed an individualized protocol based on Session One results, running repeated 1200-meter sets at progressively higher speeds (approximately 5% speed increase per set) until exhaustion. To preserve physiological response patterns vital to our modeling, blood lactate samples were obtained between sets with brief (10-second) interruptions, minimizing disruptions to the overall protocol.

Data collection employed a multi-device synchronized measurement system consisting of: (1) a wrist-mounted Garmin 965 smartwatch for continuous activity monitoring; (2) a chest-mounted Garmin HR monitor capturing both HR and running dynamics (pace, cadence, vertical oscillation, altitude, stance time, vertical ratio, and step length); (3) a portable Cosmed K5 metabolic system (300g) with dedicated harness and face mask for breath-by-breath gas exchange measurement; and (4) a two-to-four-camera video system capturing biomechanical data at 200-meter intervals from dual angles throughout the athletic stadium. This multi-sensor approach ensured comprehensive data capture across both consumer-grade and laboratory-grade measurement systems, creating a synchronized dataset with precise alignment between laboratory and wearable measurements. Figure 1 illustrates participants during the protocol while wearing the integrated measurement equipment.

HR Modeling Dataset: While the synchronized dataset supported VO_2 modeling, we developed our HR dynamics models using a separate, substantially more extensive dataset. This comprehensive resource comprises 831 running sessions from 20 distinct runners collected over a seven-year period (2018-2025), totaling 52,824 minutes (approximately 880 hours) and 10,222 km of running activity, with more than 3.1 million HR data points. The scale and diversity of this dataset provided a robust foundation for capturing complex cardiovascular dynamics across various running conditions, intensities, and individual physiological profiles. Table 3 in Appendix A presents detailed statistics for each participant, including session counts, accumulated training time, and total distance covered.

We deliberately chose chest strap HR monitors over wrist-worn devices for our HR data collection due to their superior measurement accuracy during exercise. Unlike wrist-based smartwatches that use photoplethysmography (PPG)—an optical method measuring blood flow through optical sensors—chest straps capture electrical signals directly from the heart via electrocardiogram (ECG), providing precise beat-to-beat measurements even during intense physical activity. The limitations of wrist-worn PPG sensors are well-documented: movement artifacts, sweat interference, and reduced peripheral blood flow during high-intensity exercise can compromise optical sensors' signal integrity [19, 31, 37]. In contrast, ECG-based chest straps maintain consistent accuracy across varying exercise intensities, making them the preferred choice for physiological research requiring high temporal resolution and reliability.

4 Modeling Heart Rate Dynamics

Accurately modeling HR dynamics during physical activity is fundamental for understanding physiological exercise responses. Precise HR estimation addresses three critical challenges: (1) exploring how biomechanical features influence physiological responses, revealing movement-cardiovascular relationships; (2) solving signal interruption/interpolation in wearables during high-intensity activities; (3) leveraging HR-power-VO_2 correlations for advanced training insights.

Wearables receive HR via three pathways: (i) chest-strap ECG, (ii) wrist-based PPG, or (iii) our "generative sequencing" module. Since consumer smartwatches cannot simultaneously log ECG/PPG during exercise, we use chest-strap ECG as the gold standard for training, while architecting the pipeline to accept PPG segments or generated HR streams at inference without retraining. This design enables sensor burden/battery life vs. accuracy tradeoffs: ECG achieves 2.81 bpm MAE, PPG falls between ECG and synthetic streams, and generated sequences maintain 3-5 bpm accuracy even during vigorous motion, ensuring continuous HR availability when PPG fails.

In this section, we introduce two complementary approaches for HR prediction during running, each offering unique advantages: a physiologically-constrained ordinary differential equation (ODE) approach that mathematically models cardiac adaptation rates, and a Kalman filtering framework that optimally balances prior physiological estimates with new observations.

Problem Formulation. Our goal is to predict HR during running, given a multivariate time series of biomechanical and environmental features. Let $X = \{x_1, x_2, \ldots, x_T\}$ be the sequence of external parameters (e.g., pace, cadence, vertical oscillation, altitude, stance time, vertical ratio, step length), where each $x_t \in \mathbb{R}^d$ describes running dynamics at second t. We aim to produce the corresponding HR sequence $Y = \{y_1, y_2, \ldots, y_T\}$ at 1-second intervals. The main challenge is capturing both rapid cardiac responses to intensity changes and slower physiological adaptations, all within biologically plausible limits.

Data Preprocessing. We used the HR modeling dataset (Section 3), comprising over 800 running sessions from 20 runners (2018–2025). Raw FIT files were segmented into 60-second, non-overlapping windows, with domain-specific transformations applied to maintain physiological relevance. Pace (m/s) was converted to sec/km, vertical oscillation was normalized by each runner's height, and stance time (percent) was converted to a fraction. Step length was scaled from millimeters, altitude data was split into absolute values and relative gains to capture cardiovascular demand influences, and cadence was doubled to reflect full running cycles. These transformations preserved physiological interpretability and ensured proper input scaling.

Model Architecture. Our framework combines neural feature extraction with domain-specific models of cardiac dynamics as described in Figure 2. We denote learnable functions by f and their parameters by $\boldsymbol{\vartheta}$, with subscripts indicating the specific part of the system.

The model receives a workout window $\mathbf{x}_t$ spanning the interval $[t - T, t]$. The window is encoded into a latent space by an encoder, $\mathbf{e}_t = f_{\text{enc}}(\mathbf{x}_t; \boldsymbol{\vartheta}_{\text{enc}})$, followed by an auto-regressive model $\mathbf{s}_t = f_{\text{temp}}(\mathbf{e}_t, \mathbf{s}_{t-T}; \boldsymbol{\vartheta}_{\text{temp}})$ with $\mathbf{s}_0 = \mathbf{0}$. Two dynamic models predict the latent timeseries g_t, from which the predicted HR is decoded using $\hat{h}_t = f_{\text{dec}}(g_t, \mathbf{s}_t; \boldsymbol{\vartheta}_{\text{dec}})$. This backbone is implemented with a fully connected feature encoder with leaky ReLU activation for f_{enc}, and a gated recurrent unit (GRU) for f_{temp}. For decoding HR from g_t, we use an affine mapping $\hat{h}_t = \hat{\sigma}_t g_t + \hat{\mu}_t$, where μ_t and σ_t are the HR mean and standard deviation in the interval $[t - T, t]$ estimated from the latent state by fully-connected models $\hat{\mu}_t = \mu(\mathbf{s}_t; \boldsymbol{\vartheta}_{\text{nrm}})$ and $\hat{\sigma}_t = \sigma_t(\mathbf{s}_t; \boldsymbol{\vartheta}_{\text{nrm}})$. For additional details, refer to Appendix B.1.

We now describe our two alternative approaches for dynamic HR modeling: **ODE-based dynamic model.** Our first model for the latent dynamics is based on a non-linear first-order continuous-time ODE, $\frac{\partial g_t}{\partial t} = f_{\text{ODE}}(g_t, \mathbf{s}_t; \boldsymbol{\vartheta}_{\text{ODE}})$, which we integrate with a fourth-order Runge-Kutta (RK4) method. As the right-hand side of the ODE, we used $f_{\text{ODE}} = d_t - g_t$, where $d_t = d(\mathbf{s}_t; \boldsymbol{\vartheta}_{\text{dec}})$ is a fully-connected network estimating the demand from the latent state $\mathbf{s}_t$.

Kalman Filter Approach. Our second model for the latent dynamics is a neural version of an enhanced Kalman filter [16]. We define a two-dimensional state vector $\mathbf{z}_t = (g_t, \dot{g}_t)$ comprising the

Figure 2: HR prediction: Wearable data encoded to latent states $\mathbf{s}_t$, processed via: (1) neural ODE solver ($\frac{dg_t}{dt} = d_t - g_t$) or (2) learnable Kalman filter. Both use moments $\hat{\mu}_t$, $\hat{\sigma}_t$ for denormalization to predict HR $\hat{h}_t$. Gray: learnable parameters $\boldsymbol{\vartheta}_*$.

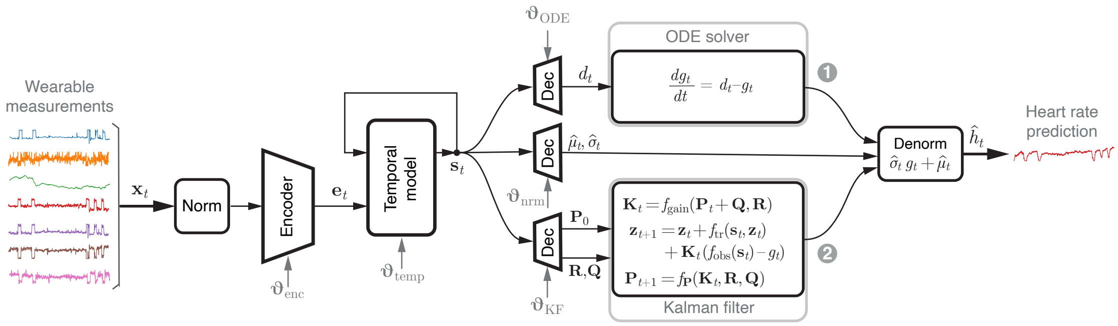

Wearable measurements
$\mathbf{x}_t$
Norm
Encoder
ϑ_{enc}
$\mathbf{e}_t$
Temporal model
ϑ_{temp}
$\mathbf{s}_t$
ϑ_{ODE}
Dec
d_t
ODE solver
$\frac{dg_t}{dt} = d_t - g_t$
1
Dec
$\hat{\mu}_t, \hat{\sigma}_t$
ϑ_{nrm}
Dec
$\mathbf{P}_0$
$\mathbf{R}, \mathbf{Q}$
ϑ_{KF}
$\mathbf{K}_t = f_{\text{gain}}(\mathbf{P}_t + \mathbf{Q}, \mathbf{R})$
$\mathbf{z}_{t+1} = \mathbf{z}_t + f_{\text{tr}}(\mathbf{s}_t, \mathbf{z}_t) + \mathbf{K}_t(f_{\text{obs}}(\mathbf{s}_t) - g_t)$
$\mathbf{P}_{t+1} = f_{\mathbf{P}}(\mathbf{K}_t, \mathbf{R}, \mathbf{Q})$
Kalman filter
2
Denorm
$\hat{\sigma}_t\, g_t + \hat{\mu}_t$
$\hat{h}_t$
Heart rate prediction

Table 1: Performance comparison of HR prediction models using Standard/Generative approaches. Values shown as "Standard/Generative" for four models.

Metric	Kalman (128,2)	Kalman (64,3)	ODE (128,2)	ODE (64,3)
Overall Performance				
MAE (bpm)	**2.81**/11.70	3.01/11.12	2.84/12.79	2.91/12.52
RMSE (bpm)	**4.60**/13.98	4.96/13.45	4.70/15.22	4.87/14.75
MAPE (%)	**2.17**/8.49	2.39/8.35	2.18/9.18	2.24/8.84
Correlation	**0.87**/0.46	0.86/0.45	0.87/0.47	0.86/0.47
R^2	**0.73**/-2.51	0.71/-1.59	0.72/-2.41	0.70/-1.95
Mean Diff. (bpm)	0.02/-2.41	0.05/-1.69	0.01/-2.32	**-0.02**/-6.60
StdDev Diff. (bpm)	**4.55**/10.13	4.92/10.42	4.67/10.78	4.83/9.81
Performance by HR Zone (MAE in bpm)				
Low HR	**10.78**/15.12	13.24/18.82	10.91/14.65	12.22/13.92
Medium HR	**3.05**/9.82	3.22/9.65	3.13/10.74	3.22/9.69
High HR	**1.99**/12.22	2.09/11.87	2.02/13.24	2.09/15.20
Performance by HR Stability (MAE in bpm)				
Transitions	**10.50**/15.05	11.93/16.52	11.44/16.36	11.64/14.75
Steady-State	**2.78**/11.68	2.99/11.11	2.81/12.78	2.89/12.51
Avg. sessions/split	125/125	113/113	122/122	122/122

latent HR and its velocity. A scalar Kalman gain is calculated using a neural network $k_t = f_{\text{gain}}(\mathbf{P}_t + \mathbf{Q}, \mathbf{R})$ and is used for the *a posteriori* state update $\mathbf{z}_{t+1} = \mathbf{z}_t + f_{\text{tr}}(\mathbf{s}_t, \mathbf{z}_t; \boldsymbol{\vartheta}_{\text{tr}}) + k_t\,(f_{\text{obs}}(\mathbf{s}_t; \boldsymbol{\vartheta}_{\text{obs}}) - g_t)\,(1, \gamma_t)$, mimicking the standard Kalman filter.

The scaling factor γ_t for velocity updates is set to 0.5, implementing a differential relationship between HR and velocity corrections. This reduced influence on velocity updates allows the model to maintain smoother trajectory changes while still responding to new observations, effectively damping oscillations that might occur from measurement noise.

The covariance update $\mathbf{P}_{t+1}$ follows a modified Kalman formulation: $\mathbf{P}_{t+1} = (\mathbf{P}_t + \mathbf{Q}) \odot \begin{pmatrix} 1 - k_t & 0 \\ 0 & 1 - \gamma_t k_t \end{pmatrix}$, where the diagonal structure preserves computational efficiency while allowing independent uncertainty tracking for both HR value and velocity components.

The decoder neural networks provide the complete set of Kalman filter parameters: initial state $\mathbf{z}_0 = (g_0, \dot{g}_0)$, initial covariance matrix $\mathbf{P}_0 = \text{diag}(\sigma^2_{g_0}, \sigma^2_{\dot{g}_0})$, process noise covariance $\mathbf{Q} = \text{diag}(\sigma^2_{\text{proc},g}, \sigma^2_{\text{proc},\dot{g}})$, and measurement noise variance $\mathbf{R} = \sigma^2_{\text{meas}}$. These parameters are computed from the final hidden state of the GRU encoder, allowing the model to adapt its filtering behavior to different individuals and physiological conditions.

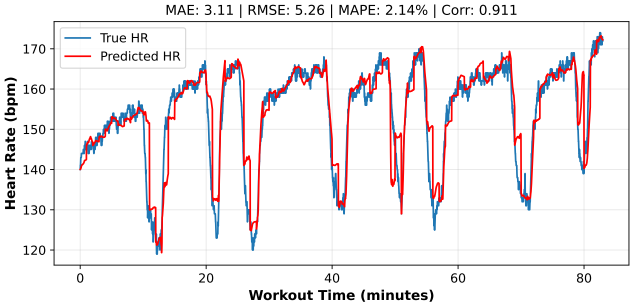
We used fully-connected models for f_{enc}, f_{tr}, f_{obs}, and f_{gain}. The gain function captures prediction errors made by the observation function, assigning greater weight to new measurements when errors are large. Additional details are provided in Appendix B.2.

Training Methodology. Both models are trained fully-supervised to minimize the Mean Absolute Error (MAE) between predicted and true HR. To ensure physiological realism, we added coarse-scale regularization terms supervising the HR first- and second-order statistics, $\mathcal{L} = \mathbb{E}_t|\hat{h}_t - h_t| + \lambda\mathbb{E}_t|\hat{\mu}_t - \mu_t| + \lambda\mathbb{E}_t|\hat{\sigma}_t - \sigma_t|$, Here, $\mathbb{E}_t$ denotes temporal expectation (in practice, finite-sample average on the training set), h_t is the ground-truth HR, and $\mu_t = \int_{t-T}^{t} g_\tau d\tau$ and $\sigma_t^2 = \int_{t-T}^{t} (g_\tau - \mu_\tau)^2 d\tau$ are the ground-truth first- and second-order moments used for the supervision. The parameter λ controls the influence of regularization by temporal statistics and was set to $\lambda = 0.1$ following an ablation study. For the ODE model, we backpropagate through the neural ODE solver using the adjoint method as proposed [6].

We tested generalizability via runner-based leave-three-out cross-validation, ensuring no overlap between training and evaluation participants. Models were assessed in eight splits, each with more than 110 sessions (20–140 minutes per session). We compared two network architectures—128 neurons (2 layers) vs. 64 neurons (3 layers)—under consistent settings (batch size = 64). An adaptive learning rate scheduler and early stopping reduced overfitting. Appendix B.3 discusses additional information.

Results. Table 1 presents a comprehensive evaluation of our HR prediction models across different architectures and inference settings. The table compares Standard inference (where the first second of each window is known) versus Generative Sequencing (where only the first second of the entire session is provided) across Kalman and ODE models with varying architectures. Our approach offers advantages over prior work: Apple's approach [27] used 10-second intervals with proprietary data, whereas we provide fine-grained 1-second predictions. Our open-source implementations enhance reproducibility and enable direct comparisons. Crucially,

Figure 3: Example HR prediction from ODE-based model (128, 2) for high-intensity workout. Blue: true HR; red: predictions.



our approach requires no historical data from previous sessions, making it suitable for new runners without a calibration phase.

Under standard inference, the Kalman (128,2) model achieved the best overall performance (MAE: 2.81 bpm, RMSE: 4.60 bpm, MAPE: 2.17%, correlation: 0.87, R^2=0.73). The ODE (128,2) model followed closely at 2.84 bpm MAE. This accuracy is particularly notable given the diversity of running intensities in our dataset. **Under Generative Sequencing**, using only the first second of the first window to drive predictions across an entire session—all models demonstrated consistent performance over extended durations, though with an expected decrease without periodic recalibration. Analysis by HR zone and stability state revealed strong results in steady-state and high-intensity scenarios, with MAE as low as 1.99 bpm in high HR zones. This outcome has significant practical importance for wearable devices prone to signal disruptions during exercise. Figure 3 illustrates ODE-based model (128,2) performance for a high-intensity interval session (HR: 120-170 bpm). The model achieves exceptional accuracy (MAE: 1.68 bpm, correlation: 0.933) despite challenging physiological transitions, effectively capturing both general trends and fine-grained HR dynamics. Additional visualizations of different workout intensities and prediction modes appear in Appendix C.1.

These visualizations confirm our models effectively capture both general trends and fine-grained HR dynamics across diverse running conditions. Intermittent recalibration (e.g., at window boundaries) could harness Standard inference accuracy while addressing real-world challenges like PPG signal loss in consumer wearables.

5 Predicting Instantaneous Oxygen Consumption from Consumer Wearables

Accurately estimating VO_2 during exercise is essential for understanding energy expenditure, quantifying training adaptation, and assessing overall fitness [30]. However, conventional VO_2 measurement methods rely on specialized metabolic carts and face masks for respiratory gas collection—equipment that typically costs over 30,000$ and requires expert operation. These technical and financial constraints fundamentally restrict practical use in real-world, continuous-monitoring scenarios where most exercise actually occurs.

To overcome this critical methodological limitation, we introduce what we believe to be the first method for predicting instantaneous VO_2 trajectories solely from consumer-grade wearable sensor data. Our approach eliminates the need for costly respiratory gas analysis while maintaining clinically acceptable accuracy, effectively democratizing access to sophisticated metabolic insights previously confined to laboratory settings.

Problem Formulation. We aim to predict VO_2 during running using a multivariate time series of biomechanical and environmental features. Let $X = \{x_1, x_2, \ldots, x_T\}$ represent a sequence of external parameters including pace, cadence, vertical oscillation, altitude, stance time, vertical ratio, and step length, where each $x_t \in \mathbb{R}^d$ captures running dynamics at second t. Our objective is to generate a corresponding VO_2 sequence $Y = \{y_1, y_2, \ldots, y_T\}$ at 1-second resolution. The primary challenge lies in accurately modeling both rapid metabolic responses to intensity changes and slower physiological adaptations while maintaining biologically plausible constraints.

Data Preprocessing. Our methodology leverages a precisely synchronized multimodal dataset (Section 3) establishing direct relationships between laboratory-grade metabolic measurements and wearable device outputs. We paired breath-by-breath data from a Cosmed K5 metabolic analyzer—the gold standard for VO_2 measurement [33]—with physiological and biomechanical metrics from a Garmin 965 smartwatch and chest-mounted HR monitor. Temporal alignment was achieved through a zero-order hold on Cosmed signals and session synchronization via GPS coordinates and event markers.

Sessions were segmented into 60-second windows coupling Cosmed and smartwatch data, creating a comprehensive dataset where metabolic parameters (VO_2, VCO_2, ventilation, RER) sampled at 0.3–0.5 Hz precisely align with smartwatch features (HR, cadence, vertical oscillation, altitude) at 1 Hz. To mitigate breath-by-breath variability, we applied a Savitzky-Golay filter (15-sample window, polynomial order 3) [38] that preserves rapid physiological transitions while removing measurement artifacts.

From the refined dataset, we derived a comprehensive feature set capturing the multifaceted nature of exercise physiology. The features include cardiac measurements like HR, biomechanical indicators such as cadence, vertical oscillation ratio, step length, and stance time, contextual variables including pace, grade, and cumulative distance, positional encodings of session window index and total elapsed time, and anthropometric data like age, gender, height, and weight. This diverse feature set enables our model to account for immediate cardiorespiratory responses to exercise intensity transitions and individual physiological characteristics, establishing a robust foundation for accurate VO_2 estimation using only consumer-grade wearable data.

Model Architecture. Predicting VO_2 during exercise presents unique challenges due to the complex interplay between physiological systems and rapid metabolic adaptations. We model VO_2 dynamics as a state–estimation problem subject to physiologically informed constraints, denoting the latent oxygen-consumption state at time t by v_t (and $\hat{v}_t, \tilde{v}_{t^-}$ for its predicted and trend-extrapolated versions, respectively). Figure 4 summarizes the framework. In what follows, learnable functions are written $f(\cdot\,; \boldsymbol{\vartheta})$, where the subscript on $\boldsymbol{\vartheta}$ indicates the module.

The model processes a workout window $\mathbf{x}_t$ spanning $[t - T, t]$ containing ten biomechanical, HR, and temporal features. A feature

Figure 4: The VO_2 prediction framework normalizes and encodes wearable sensor inputs, then uses a dual-stream architecture (neural Kalman filter + direct estimation) adaptively blended for final VO_2 predictions.

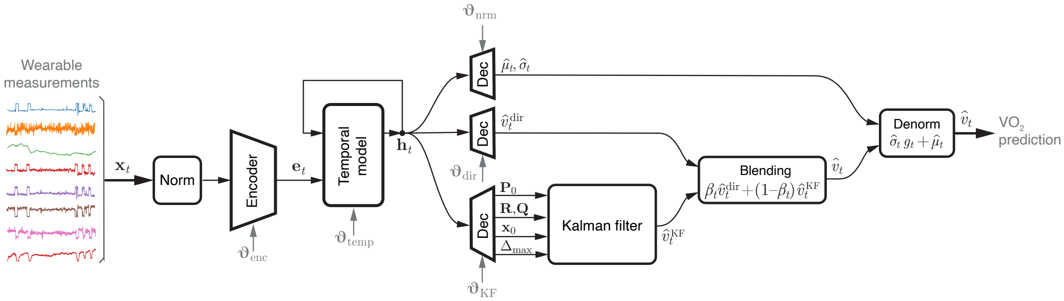

Wearable measurements
$\mathbf{x}_t$
Norm
Encoder
ϑ_{enc}
$\mathbf{e}_t$
Temporal model
$\vartheta_{\mathrm{temp}}$
$\mathbf{h}_t$
ϑ_{nrm}
Dec
$\hat{\mu}_t, \hat{\sigma}_t$
Dec
$\hat{v}_t^{\mathrm{dir}}$
ϑ_{dir}
Dec
$\mathbf{P}_0$
$\mathbf{R},\mathbf{Q}$
$\mathbf{x}_0$
$\Delta_{\max}$
ϑ_{KF}
Kalman filter
$\hat{v}_t^{\mathrm{KF}}$
Blending
$\beta_t \hat{v}_t^{\mathrm{dir}} + (1-\beta_t)\,\hat{v}_t^{\mathrm{KF}}$
$\hat{v}_t$
Denorm
$\hat{\sigma}_t\, g_t + \hat{\mu}_t$
$\hat{v}_t$
VO_2 prediction

Table 2: Seq-to-Seq VO_2 Prediction for two models using Leave-One-Runner-Out Cross-Validation, with real HR.

	Model 256-2			Model 128-4		
Runner	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
Runner-1	520.20	582.88	17.17	**325.40**	**399.68**	**11.59**
Runner-2	1326.20	1361.69	35.40	**453.74**	**493.50**	**12.81**
Runner-3	218.41	473.09	11.47	212.02	477.44	11.59
Runner-4	127.55	198.96	7.85	129.08	207.55	7.95
Runner-5	606.88	636.92	22.08	451.26	488.13	**16.42**
Runner-6	261.49	328.66	13.26	190.94	270.33	**10.02**
Runner-7	159.01	201.17	6.19	131.81	192.59	5.35
Runner-8	321.80	382.35	13.54	235.29	**298.0**	**10.45**
Runner-9	321.01	398.18	25.05	267.14	344.00	**21.26**
Runner-10	108.75	155.70	6.24	112.37	160.21	6.54
Aggregate	397.13	471.96	15.83	**251.00**	**333.20**	**11.40**

encoder and bidirectional GRU generate a hidden representation $\mathbf{h}_t = f_{\mathrm{temp}}\big(f_{\mathrm{enc}}(\mathbf{x}_t; \boldsymbol{\vartheta}_{\mathrm{enc}}); \boldsymbol{\vartheta}_{\mathrm{temp}}\big)$.

Neural–Kalman update. Using $\mathbf{h}_t$, the neural Kalman filter generates a Kalman filter-based VO_2 estimate $\hat{v}_t^{\mathrm{KF}}$ through an adaptive mechanism that balances new observations with prior estimates:

$$\hat{v}_t^{\mathrm{KF}} = \hat{v}_{t-1} + f_{\mathrm{gain}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\mathrm{KF}}) \left[f_{\mathrm{obs}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\mathrm{obs}}) - \hat{v}_{t-1} \right]_{f_{\Delta_{\min}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\Delta})}^{f_{\Delta_{\max}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\Delta})},$$

where f_{gain} modulates observation trust and $f_{\Delta_{\min}}, f_{\Delta_{\max}}$ enforce physiologically constrained rate-of-change limits.

Trend/direct dual pathway. To enhance robustness, we employ a dual-pathway approach blending trend-extrapolated and direct neural estimates:

$$\tilde{v}_{t^-} = \hat{v}_{t-1} + f_{\mathrm{trend}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\mathrm{trend}})\,(\hat{v}_{t-1} - \hat{v}_{t-2}),$$
$$\hat{v}_t^{\mathrm{dir}} = f_{\mathrm{direct}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\mathrm{dir}}),\ \hat{v}_t = \beta_t \hat{v}_t^{\mathrm{dir}} + (1 - \beta_t)\hat{v}_t^{\mathrm{KF}},$$

where $\beta_t = f_{\mathrm{blend}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\mathrm{blend}})$ determines the blending proportions. All modules $f_{\mathrm{enc}}, f_{\mathrm{gain}}, f_{\mathrm{obs}}, f_{\Delta_{\min}}, f_{\Delta_{\max}}, f_{\mathrm{trend}}, f_{\mathrm{blend}}$, and f_{direct} are implemented as multilayer perceptrons (MLPs) with corresponding parameters $\boldsymbol{\vartheta}_{\mathrm{enc}}, \boldsymbol{\vartheta}_{\mathrm{KF}}, \boldsymbol{\vartheta}_{\mathrm{obs}}, \boldsymbol{\vartheta}_{\Delta}, \boldsymbol{\vartheta}_{\mathrm{trend}}, \boldsymbol{\vartheta}_{\mathrm{blend}}$, and $\boldsymbol{\vartheta}_{\mathrm{dir}}$, respectively. The temporal model f_{temp} is a bidirectional GRU (biGRU) with parameters $\boldsymbol{\vartheta}_{\mathrm{temp}}$ that captures temporal dynamics. The final output is normalized using parameters $\boldsymbol{\vartheta}_{\mathrm{nrm}}$. Detailed layer configurations are provided in Appendix B.4.

Training Methodology. Our model is trained fully-supervised to minimize a composite loss function that balances immediate prediction accuracy with physiological plausibility,

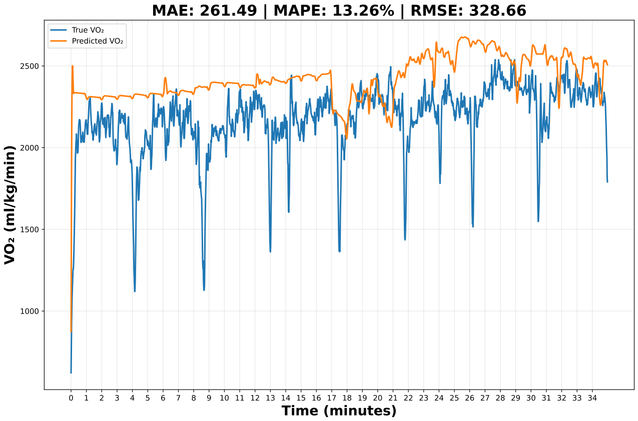
$$\mathcal{L} = \mathbb{E}_t|\hat{v}_t - v_t| + \lambda_{\mathrm{dynamic}}\mathbb{E}_t\left(\alpha\left|\frac{d\hat{v}_t}{dt} - \frac{dv_t}{dt}\right| + \beta\left|\frac{d^2\hat{v}_t}{dt^2} - \frac{d^2v_t}{dt^2}\right|\right)$$
$$+ \lambda_{\mathrm{aux}}\Big(\mathbb{E}_t|\hat{\mu}_t - \mu_t| + \mathbb{E}_t|\hat{\sigma}_t - \sigma_t| + \mathbb{E}_t|\hat{\Delta}_t - \Delta_t|\Big),$$

where $\hat{v}_t$ represents the blended estimate $\beta_t\hat{v}_t^{\mathrm{dir}} + (1-\beta_t)\hat{v}_t^{\mathrm{KF}}$. Here $\mathbb{E}_t$ denotes temporal expectation, v_t is the ground-truth VO_2 value, $\frac{dv_t}{dt}$ and $\frac{d^2v_t}{dt^2}$ represent the first and second derivatives of the VO_2 signal capturing velocity and acceleration dynamics. The coefficients α and β weight the relative importance of the derivatives. For the statistical terms, μ_t is the temporal mean, σ_t is the standard deviation, and $\Delta_t = Q_{0.95}(|v_\tau - v_{\tau-1}|)$ represents the 95th percentile of absolute differences in consecutive VO_2 values. The parameters $\lambda_{\mathrm{dynamic}}$ and λ_{aux} follow a curriculum-based schedule, with $\lambda_{\mathrm{dynamic}}$ increasing from 0 to 0.7 and λ_{aux} increasing from 0.1 to 0.3 over the course of training to gradually emphasize physiological realism.

We complement this curriculum learning with adaptive gradient clipping that permits larger parameter updates during early training stages while enforcing stability as the model converges. This progressive optimization strategy enables the model to establish foundational prediction capabilities before refining its representation of the nuanced transition dynamics that characterize VO_2 response during variable-intensity exercise. Appendix B.5 contains additional information.

Results. We evaluated the model's generalizability using a leave-one-runner-out approach, training on data from all but one participant and testing on the remaining runner. Due to the dataset's limited size, we used overlapping windows during training but non-overlapping windows for left-out runner evaluation, maintaining session integrity. A unique aspect of our method is generating complete sequence predictions using only **the first second of VO_2 data** for initialization—mimicking real-world scenarios where continuous metabolic measurement is impractical. This minimal conditioning approach, potentially derived from domain knowledge or brief calibration, removes the need for ongoing respiratory gas analysis. Our sequence-to-sequence VO_2 predictions were assessed using multiple metrics, detailed in Table 2.

Figure 5: VO_2 predictions from first-second data during an incremental testing session with periodic lactate measurements. Blue: ground truth; orange: model predictions.



Contextualizing Prediction Errors. Our model's performance aligns with established benchmarks for instantaneous VO_2 estimation. At high intensities (VO_2 ~3000 ml/min), gold-standard metabolic equipment typically accepts errors of 300–600 ml/min (10–20%). Our consumer-grade smartwatch method achieves an aggregate Mean Absolute Percentage Error (MAPE) of 11.40%, compared to the 256-2 model's 15.83%, demonstrating remarkable accuracy given the limited input data. This error range corresponds to approximately 4.3–8.6 ml$\cdot$kg$^{-1}\cdot$min^{-1} for a 70 kg athlete and aligns with discrepancies observed between research-grade devices.

Several physiological factors inherently complicate VO_2 prediction: **breath-by-breath variability** [22] introduces inherent (±5-10%) fluctuations, **nonlinear VO_2 kinetics** [46] challenge modeling during transitional phases, **delayed cardiorespiratory equilibrium** [44] causes temporal misalignments, and **individual differences** in body composition, training status, and metabolic efficiency add significant variability.

Despite these challenges, our optimized neural kalman based model (128-4) achieved an aggregate MAPE of 11.4%, with most runners exhibiting errors below 20%. This performance is particularly notable given that our method relies solely on consumer-grade wearable data and constructs the entire VO_2 trajectory from only the first second of metabolic data, without requiring exhaustive laboratory calibration. For these results, we used ground truth HR measurements rather than predicted values to isolate the performance of the VO_2 estimation component. Appendix C.2 contains complementary results using predicted HR values from our trained models defined in previous sections (Section 4).

Figure 5 demonstrates the generative capability of the model by plotting ground truth VO_2 measurements (blue) alongside predictions (orange) during an incremental testing protocol. The model reconstructs the entire VO_2 trajectory using only the first second of metabolic data, maintaining high fidelity over extended periods despite varying exercise intensities. Appendix E contains additional visualizations, including a maximal capacity assessment.

The incremental testing protocol presents a complex scenario, with 35 minutes of progressively faster 1200-meter sets until exhaustion. Despite periodic dips in the ground truth data caused by brief interruptions for blood lactate sampling, the model maintains a MAPE of 13.26%, successfully tracking the overall increasing trend in VO_2 while effectively filtering out these transient artifacts.

Performance Analysis and Limitations. The 128-4 model reveals significant inter-subject variation, with MAPE ranging from 5.35% to 21.26%. This variation underscores the complexity of individual metabolic responses. The deeper 128-4 architecture consistently outperforms the shallower 256-2 model, suggesting that additional neural layers more effectively capture the intricate dynamics of VO_2 time-series.

With an MAE of 251 ml/min, we achieve approximately 9% error at 3000 ml/min—a remarkable accomplishment using only one second of initial metabolic data. Our Kalman-inspired neural model successfully balances pointwise accuracy with realistic temporal dynamics, though challenges persist in modeling rapid high-intensity transitions and individual metabolic variability.

This sequence-to-sequence generation approach represents, to the best of our knowledge, the first method to predict complete instantaneous VO_2 trajectories using minimal initial data and smartwatch signals. Unlike existing approaches that either provide single-point VO_2max estimates or require expensive continuous respiratory measurements, our method offers a practical solution for real-world metabolic monitoring. **The performance of approximately 13% MAPE for most participants—establishes a new benchmark in wearable-based VO_2 prediction**. Future research will explore incorporating additional physiological signals, refining temporal representations, and developing adaptive calibration methods to account for individual fitness changes over time.

6 Discussion and Conclusion

Our work bridges the gap between laboratory-grade physiological assessment and consumer wearables by introducing what we believe is the first framework capable of predicting instantaneous VO_2 trajectories directly from consumer devices. By combining physiologically constrained modeling with modern ML techniques, we demonstrate how sophisticated metabolic metrics—previously confined to specialized laboratories—can be reliably estimated using everyday technology, democratizing advanced exercise physiology.

VO_2 Prediction and HR Dynamics. Our sequence-to-sequence VO_2 framework generates complete metabolic trajectories from just one second of calibration data, achieving ~13% MAPE without specialized lab equipment, delivering sophisticated metabolic insights previously requiring 30,000$+ gas analysis systems. Our HR dynamics models achieve exceptional accuracy (MAE: 2.81 bpm, correlation: 0.87) across diverse running conditions and maintain physiological plausibility during high-intensity exercise (MAE: 1.99 bpm controlled). These models support VO_2 prediction while solving PPG signal interruptions-a persistent consumer device challenge.

Toward Metabolic Zone Classification. Our results establish a foundation for predicting metabolic thresholds from non-invasive data. Our synchronized dataset with blood lactate measurements enables future classification of aerobic, threshold, and anaerobic zones, potentially transforming training prescription by making advanced zone-based guidance accessible beyond elite athletes.

Limitations and Future Directions. Despite promising results, limitations include: (1) participant cohort primarily comprises highly trained runners, potentially limiting generalizability; (2) models exclude environmental factors (temperature, humidity). Future work will fuse on-device thermistors and weather-API data to compensate; (3) approach ignores longitudinal training adaptations. Future work: broaden demographics, integrate environmental sensors, and develop longitudinal models for evolving fitness.

Safe and Responsible Innovation Statement. Our prediction framework prioritizes privacy through local data processing without external transmission, while acknowledging potential bias in our high performing athletes-focused models despite evaluated cross-runner generalizability. While democratizing metabolic monitoring offers significant health benefits, we recognize misinterpretation risks without professional guidance. Future work will incorporate broader demographics, transparent confidence metrics, and clear guidelines for responsible fitness and healthcare applications.

References

[1] Rob Argent, Antonio Bevilacqua, Alison Keogh, Ailish Daly, and Brian Caulfield. 2021. The importance of real-world validation of machine learning systems in wearable exercise biofeedback platforms: A case study. *Sensors* 21, 7 (2021), 2346.

[2] Polly Aylwin, George Havenith, Marco Cardinale, Alexander Lloyd, Mohammed Ihsan, Lee Taylor, Paolo Emilio Adami, Marine Alhammoud, Juan-Manuel Alonso, Nicolas Bouscaren, et al. 2023. Thermoregulatory responses during road races in hot-humid conditions at the 2019 athletics world championships. *Journal of Applied Physiology* 134, 5 (2023), 1300–1311.

[3] Mridula Badrinarayanan, V. Sricharan, Rohan Jais, N. Danush Adhithya, G. Sri Gayathri, S. Preejith, and M. Sivaprakasam. 2024. Evaluating the Utility of a Cardiac Health Assessment Test in Predicting VO2max Obtained from CPET: A Pilot Study. *2024 IEEE 12th International Conference on Serious Games and Applications for Health (SeGAH)* 1 (2024), 1–8. https://doi.org/10.1109/SeGAH61285.2024.10639555

[4] T. Barstow and P. Molé. 1991. Linear and nonlinear characteristics of oxygen uptake kinetics during heavy exercise. *Journal of Applied Physiology* 71 (1991), 2099–2106. https://doi.org/10.1152/JAPPL.1991.71.6.2099

[5] Bryson Carrier, Macy M. Helm, Kyle Cruz, Brenna Barrios, and J. Navalta. 2023. Validation of Aerobic Capacity (VO2max) and Lactate Threshold in Wearable Technology for Athletic Populations. *Technologies* 1 (2023), 1. https://doi.org/10.3390/technologies11030071

[6] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. 2018. Neural ordinary differential equations. *Advances in neural information processing systems* 31 (2018), 1.

[7] Hsueh-Wen Chow and Chao-Ching Yang. 2020. Accuracy of Optical Heart Rate Sensing Technology in Wearable Fitness Trackers for Young and Older Adults: Validation and Comparison Study. *JMIR mHealth and uHealth* 8 (2020), 1. https://doi.org/10.2196/14707

[8] C. Cooper and A. Garfinkel. 2022. A novel geometric method for determining the time constant for oxygen uptake kinetics. *Journal of Applied Physiology* 1 (2022), 1. https://doi.org/10.1152/japplphysiol.00049.2022

[9] Muhammad Etiwy, Zade Akhrass, Lauren Gillinov, A. Alashi, Robert Z. Wang, G. Blackburn, S. Gillinov, D. Phelan, A. Gillinov, P. Houghtaling, Hoda Javadikasgari, and M. Desai. 2019. Accuracy of wearable heart rate monitors in cardiac rehabilitation. *Cardiovascular Diagnosis and Therapy* 9, 3 (2019), 262–271. https://doi.org/10.21037/CDT.2019.04.08

[10] Diana Ferreira, J. Beckert, Ricardo Minhalma, A. Prata, and Marcos Miranda. 2016. Different oxygen consumption kinetics elicits the same oxygen deficit in normalized intensity exercise. *British Journal of Sports Medicine* 50 (2016), A82–A83. https://doi.org/10.1136/bjsports-2016-097120.145

[11] A. Gentilin. 2023. The informative power of heart rate along with machine learning regression models to predict maximal oxygen consumption and maximal workload capacity. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology* 1 (2023), 1. https://doi.org/10.1177/17543371231213904

[12] Tyler M. Grey, M. D. Spencer, G. Belfry, J. Kowalchuk, D. Paterson, and J. Murias. 2015. Effects of age and long-term endurance training on VO2 kinetics. *Medicine and Science in Sports and Exercise* 47, 2 (2015), 289–298. https://doi.org/10.1249/MSS.0000000000000398

[13] A. Hecksteden, W. Pitsch, F. Rosenberger, and T. Meyer. 2018. Repeated testing for the assessment of individual response to exercise training. *Journal of Applied Physiology* 124, 6 (2018), 1567–1579. https://doi.org/10.1152/japplphysiol.00896.2017

[14] E. Hedge, R. Hughson, and R. Amelard. 2021. Deep Learning Models Predict Dynamic Oxygen Uptake Responses from Wearable Sensor Data during Moderate- and Heavy-Intensity Exercise. *The FASEB Journal* 35 (2021), 1. https://doi.org/10.1096/FASEBJ.2021.35.S1.02796

[15] Carl A James, Ashley GB Willmott, CW Daniel Lee, TK Gabriel Pun, Ray Tai, and Oliver R Gibson. 2024. Mixed-method heat acclimation induces heat adaptations in international triathletes without training modification. *Journal of Science in Sport and Exercise* 6, 3 (2024), 253–264.

[16] Rudolph Emil Kalman. 1960. A new approach to linear filtering and prediction problems. *J* 1 (1960), 1.

[17] S. Khurshid, T. Churchill, N. Diamant, P. Di Achille, C. Reeder, P. Singh, S. Friedman, M. M. Wasfy, G. A. Alba, B. A. Maron, D. Systrom, B. M. Wertheim, P. Ellinor, J. E. Ho, A. Baggish, P. Batra, S. Lubitz, and J. S. Guseh. 2023. Deep learned representations of the resting 12-lead electrocardiogram to predict VO2 at peak exercise. *European Journal of Preventive Cardiology* 1 (2023), 1. https://doi.org/10.1093/eurjpc/zwad321

[18] Rahul G Krishnan, Uri Shalit, and David Sontag. 2015. Deep kalman filters. *arXiv preprint arXiv:1511.05121* 1 (2015), 1.

[19] Pilar Martín-Escudero, Ana María Cabanas, María Luisa Dotor-Castilla, Mercedes Galindo-Canales, Francisco Miguel-Tobal, Cristina Fernández-Pérez, Manuel Fuentes-Ferrer, and Romano Giannetti. 2023. Are activity wrist-worn devices accurate for determining heart rate during intense exercise? *Bioengineering* 10, 2 (2023), 254.

[20] Peter S Maybeck. 1982. *Stochastic models, estimation, and control*. Vol. 3. Academic press, None. 1 pages.

[21] Luke D. McCormick. 2018. Predictability Of VO2max From Three Commercially Available Devices: 1130 May 31 9. *Medicine Science in Sports Exercise* 1 (2018), 1. https://doi.org/10.1249/01.mss.0000535959.71788.dc

[22] Craig Ryan McNulty and Robert Andrew Robergs. 2019. Repeat trial and breath averaging: Recommendations for research of VO2 kinetics of exercise transitions to steady-state. *Movement & Sport Sciences-Science & Motricité* 106, 4 (2019), 37–44.

[23] Grégoire P Millet, J Burtscher, N Bourdillon, Giorgio Manferdelli, M Burtscher, and Ø Sandbakk. 2023. The V˙O2max Legacy of Hill and Lupton (1923)-100 Years On. *International Journal of Sports Physiology and Performance* 1 (2023), 1–4. https://doi.org/10.1123/ijspp.2023-0229

[24] Devajit Mohajan and H. Mohajan. 2023. Long-Term Regular Exercise Increases vo2max for Cardiorespiratory Fitness. *Innovation in Science and Technology* 1 (2023), 1. https://doi.org/10.56397/ist.2023.03.07

[25] Pablo Molina-Garcia, Hannah L Notbohm, Moritz Schumann, Rob Argent, Megan Hetherington-Rauth, Julie Stang, Wilhelm Bloch, Sulin Cheng, Ulf Ekelund, Luis B Sardinha, et al. 2022. Validity of estimating the maximal oxygen consumption by consumer wearables: a systematic review with meta-analysis and expert statement of the INTERLIVE network. *Sports Medicine* 52, 7 (2022), 1577–1597.

[26] J. Murias and D. Paterson. 2015. Slower VO2 kinetics in older individuals: Is it inevitable? *Medicine and Science in Sports and Exercise* 47, 11 (2015), 2308–2318. https://doi.org/10.1249/MSS.0000000000000686

[27] Achille Nazaret, Sana Tonekaboni, Gregory Darnell, Shirley You Ren, Guillermo Sapiro, and Andrew C Miller. 2023. Modeling personalized heart rate response to exercise and environmental factors with wearables data. *NPJ Digital Medicine* 6, 1 (2023), 207.

[28] Benjamin W. Nelson and N. Allen. 2019. Accuracy of Consumer Wearable Heart Rate Measurement During an Ecologically Valid 24-Hour Period: Intraindividual Validation Study. *JMIR mHealth and uHealth* 7, 1 (2019), 1. https://doi.org/10.2196/10828

[29] T. Noakes. 2008. How did A V Hill understand the VO2max and the "plateau phenomenon"? Still no clarity? *British Journal of Sports Medicine* 42 (2008), 574–580. https://doi.org/10.1136/bjsm.2008.046771

[30] American College of Sports Medicine et al. 2013. *ACSM's guidelines for exercise testing and prescription*. Lippincott williams & wilkins, None.

[31] Jakub Parak, Mikko Salonen, Tero Myllymäki, and Ilkka Korhonen. 2021. Comparison of heart rate monitoring accuracy between chest strap and vest during physical training and implications on training decisions. *Sensors* 21, 24 (2021), 8411.

[32] D. Paterson and B. Whipp. 1991. Asymmetries of oxygen uptake transients at the on- and offset of heavy exercise in humans. *The Journal of Physiology* 443 (1991), 1. https://doi.org/10.1113/jphysiol.1991.sp018852

[33] Ismael Perez-Suarez, Marcos Martin-Rincon, Juan José Gonzalez-Henriquez, Chiara Fezzardi, Sergio Perez-Regalado, Victor Galvan-Alvarez, Julian W Juan-Habib, David Morales-Alamo, and Jose AL Calbet. 2018. Accuracy and precision of the COSMED K5 portable analyser. *Frontiers in physiology* 9 (2018), 1764.

[34] Peter Piil, T. S. Jørgensen, J. Egelund, N. Rytter, L. Gliemann, J. Bangsbo, Y. Hellsten, and M. Nyberg. 2018. Effects of aging and exercise training on leg hemodynamics and oxidative metabolism in the transition from rest to steady-state exercise: role of cGMP signaling. *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology* 315, 2 (2018), R274–R283. https://doi.org/10.1152/ajpregu.00446.2017

[35] Joel D. Reece, J. Bunn, Minsoo Choi, and J. Navalta. 2021. Assessing Heart Rate Using Consumer Technology Association Standards. *Technologies* 1 (2021), 1. https://doi.org/10.3390/technologies9030046

[36] Siti Sabrena B. Safari, Saaveethya Sivakumar, K. Lim, and Terence Peng Lian Tan. 2022. A Review on Oxygen Consumption and Heart Rate Monitoring Methods Applications. *2022 International Conference on Green Energy, Computing and Sustainable Technology (GECOST)* 462 (2022), 462–469. https://doi.org/10.1109/GECOST55694.2022.10010466

[37] Francesco Sartor, Jos Gelissen, Ralph Van Dinther, David Roovers, Gabriele B Papini, and Giuseppe Coppola. 2018. Wrist-worn optical and chest strap heart rate comparison in a heterogeneous sample of healthy individuals and in coronary artery disease patients. *BMC Sports Science, Medicine and Rehabilitation* 10 (2018), 1–10.

[38] Abraham Savitzky and Marcel JE Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry* 36, 8 (1964), 1627–1639.

[39] Benjamin T Schumacher, Michael J LaMonte, Andrea Z LaCroix, Eleanor M Simonsick, Steven P Hooker, Humberto Parada Jr, John Bellettiere, and Arun Kumar. 2024. Development, validation, and transportability of several machine-learned, non-exercise-based VO2max prediction models for older adults. *Journal of Sport and Health Science* 13, 5 (2024), 611–620.

[40] Dimitris Spathis, I. Perez-Pozuelo, T. Gonzales, S. Brage, N. Wareham, and Cecilia Mascolo. 2022. Longitudinal cardio-respiratory fitness prediction through free-living wearable sensors. *ArXiv* abs/2205.03116, 1 (2022), 1. https://doi.org/10.48550/arXiv.2205.03116

[41] Barbara Strasser. 2018. Survival of the fittest: VO2max, a key predictor of longevity? *Frontiers in Bioscience* 23 (2018), 1505–1516. https://doi.org/10.2741/4657

[42] M. P. Støve, R. Holm, Anne Kjaersgaard, Kristian Duncker, Mie Ravn Jensen, and B. Larsen. 2020. Measurement latency significantly contributes to reduced heart rate measurement accuracy in wearable devices. *Journal of Medical Engineering Technology* 44 (2020), 125 – 132. https://doi.org/10.1080/03091902.2020.1753836

[43] S. Wiecha, P. S. Kasiak, P. Szwed, T. Kowalski, I. Cieśliński, M. Postuła, and A. Klusiewicz. 2023. VO2max prediction based on submaximal cardiorespiratory relationships and body composition in male runners and cyclists: a population study. *bioRxiv* 1 (2023), 46. https://doi.org/10.1101/2023.02.03.527004

[44] Fan Xu and Edward C Rhodes. 1999. Oxygen uptake kinetics during exercise. *Sports medicine* 27 (1999), 313–327.

[45] Rodrigo Zacca, Rui Azevedo, P. Figueiredo, J. Vilas-Boas, F. Castro, D. Pyne, and R. Fernandes. 2019. VO2 FITTING: A Free and Open-Source Software for Modelling Oxygen Uptake Kinetics in Swimming and other Exercise Modalities. *Sports* 7, 2 (2019), 2. https://doi.org/10.3390/sports7020031

[46] Jerzy A Zoladz, AC Rademaker, and Anthony J Sargeant. 1995. Non-linear relationship between O2 uptake and power output at high intensities of exercise in humans. *The Journal of physiology* 488, 1 (1995), 211–217.

Table 3: Comprehensive running dataset statistics showing number of sessions, total time, and total distance for each participant.

Runner	# Sessions	Time		Dis. (km)
		s	min	
runner-1	74	91,232	1,521	305.01
runner-2	19	26,569	443	90.23
runner-3	29	32,739	546	117.42
runner-4	25	25,734	429	102.80
runner-5	3	4,389	73	18.26
runner-6	53	50,908	848	190.51
runner-7	124	610,995	10,183	2,040.84
runner-8	41	142,175	2,370	471.17
runner-9	25	94,731	1,579	314.23
runner-10	15	14,971	250	56.24
runner-11	33	186,187	3,103	662.15
runner-12	48	232,667	3,878	812.43
runner-13	22	22,711	379	66.07
runner-14	68	351,054	5,851	1,240.50
runner-15	34	86,637	1,444	321.47
runner-16	10	60,356	1,006	211.03
runner-17	32	154,292	2,572	485.47
runner-19	76	418,703	6,978	1,161.24
runner-20	100	562,421	9,374	1,555.86
Total	**831**	**3,169,471**	**52,824**	**10,222.90**

7 Appendices

This appendix provides additional implementation details, model architectures, and performance visualizations that complement the main text.

A Comprehensive Running Dataset Statistics

Table 3 presents statistics for the HR modeling dataset, detailing the number of sessions, accumulated training time (in both seconds and minutes), and total distance covered by each participant. The dataset contaisn 831 running sessions from 20 participants, representing over 52,825 minutes (approximately 880 hours) of running activity and covering a cumulative distance of 10,222.90 kilometers. This extensive collection provides a robust foundation for developing and validating our HR prediction models across diverse runners and training conditions. Notable variance exists in individual contributions, with participant data ranging from 3 sessions (runner-5) to 124 sessions (runner-7), highlighting the dataset's capacity to capture both intra- and inter-subject variability in cardiovascular responses to running.

B Detailed Mathematical Formulation of HR Models, Oxygen Consumption Model and Training Methodology

We provide detailed mathematical formulations for our two HR prediction approaches: the ODE-based model and the Kalman filter model. Both models share a common feature processing pipeline but diverge in how they model cardiac dynamics.

B.1 ODE-based Heart Rate Model

We model HR as a nonlinear dynamical system whose evolution depends on workout-specific inputs and remains within physiologically plausible bounds. Learnable maps are denoted $f_\bullet$ with parameters $\vartheta_\bullet$.

Inputs and latent backbone. Given a multivariate input window $\mathbf{x}_t \in \mathbb{R}^{B\times T\times D_{\text{in}}}$ (batch size B, length T, feature dimension D_{in}), a feature encoder produces latent embeddings

$$\mathbf{e}_t = f_{\text{enc}}(\mathbf{x}_t; \vartheta_{\text{enc}}),$$

followed by a recurrent backbone

$$\mathbf{s}_t = f_{\text{temp}}(\mathbf{e}_t, \mathbf{s}_{t-T}; \vartheta_{\text{temp}}), \qquad \mathbf{s}_0 = \mathbf{0},$$

that captures long-range temporal context.

Latent ODE dynamics. We introduce a latent HR state $g_t \in \mathbb{R}^{B\times T\times D_{\text{latent}}}$ governed by a first-order ODE

$$\frac{\partial g_t}{\partial t} = f_{\text{ODE}}(g_t, \mathbf{s}_t; \vartheta_{\text{ODE}}),$$

whose right-hand side embodies cardiac demand

$$f_{\text{ODE}}(g_t, \mathbf{s}_t; \vartheta_{\text{ODE}}) = d(\mathbf{s}_t; \vartheta_{\text{dec}}) - g_t,$$

with $d(\cdot)$ a small MLP that estimates the HR level required by the current exercise intensity.

Numerical integration and learning. The ODE is integrated with a fourth-order Runge–Kutta solver (RK4), and gradients are obtained by the adjoint method of Neural ODEs, enabling end-to-end training.

Physiological normalisation and decoding. From the backbone state we predict per-batch normalisation parameters

$$\hat{\mu}_t = f_\mu(\mathbf{s}_t; \vartheta_{\text{nrm}}) \in \mathbb{R}^B,$$
$$\hat{\sigma}_t = f_\sigma(\mathbf{s}_t; \vartheta_{\text{nrm}}) \in \mathbb{R}^B,$$

then decode observable HR by an affine map

$$\hat{h}_t = f_{\text{dec}}(g_t, \mathbf{s}_t; \vartheta_{\text{dec}}) = \hat{\sigma}_t\, g_t + \hat{\mu}_t.$$

Interpretability and advantages. Our ODE model offers several advantages for HR prediction: (1) Bounded parameters enforce physiological plausibility; (2) The ODE captures gradual HR adaptation to changing intensity; (3) The model remains interpretable, with parameters corresponding to meaningful physiological quantities.

B.2 Kalman Filter Heart Rate Model

We cast HR prediction as a state-estimation problem solved by an enhanced Kalman filter whose parameters are supplied by neural networks. As before, learnable maps are denoted $f_\bullet$ with parameters $\vartheta_\bullet$.

Inputs and latent backbone. A multivariate input window $\mathbf{x}_t \in \mathbb{R}^{B\times T\times D_{\text{in}}}$ is encoded

$$\mathbf{e}_t = f_{\text{enc}}(\mathbf{x}_t; \vartheta_{\text{enc}}), \qquad \mathbf{s}_t = f_{\text{temp}}(\mathbf{e}_t, \mathbf{s}_{t-T}; \vartheta_{\text{temp}}),\ \mathbf{s}_0 = 0,$$

yielding a workout embedding $\mathbf{s}_t \in \mathbb{R}^{B\times D_{\text{hid}}}$ that conditions all filter parameters.

State definition. We track latent HR and its velocity via the two-dimensional state

$$\mathbf{z}_t = \begin{bmatrix} g_t \\ \dot{g}_t \end{bmatrix} \in \mathbb{R}^{B\times 2}, \qquad \mathbf{P}_t \in \mathbb{R}^{B\times 2\times 2}$$

with process noise covariance $\mathbf{Q}_t = f_{\text{noise}}(\mathbf{s}_t; \boldsymbol{\vartheta}_{\text{noise}})$ and measurement noise variance $\mathbf{R}_t = f_{\text{meas}}(\mathbf{s}_t; \boldsymbol{\vartheta}_{\text{meas}})$.

Kalman recursion. Prediction step

$$\begin{aligned} \mathbf{z}_t^{+} &= \mathbf{z}_t + f_{\text{tr}}(\mathbf{s}_t, \mathbf{z}_t; \boldsymbol{\vartheta}_{\text{tr}}), \\ \mathbf{P}_t^{+} &= \mathbf{P}_t + \mathbf{Q}_t. \end{aligned}$$

Update step

$$\begin{aligned} \hat{h}_t &= f_{\text{obs}}(\mathbf{s}_t; \boldsymbol{\vartheta}_{\text{obs}}), \\ \nu_t &= \hat{h}_t - g_t^{+}, \\ k_t &= f_{\text{gain}}\big(\mathbf{P}_t^{+}, \mathbf{R}_t; \boldsymbol{\vartheta}_{\text{gain}}\big), \quad \text{where } \mathbf{P}_t^{+} = \mathbf{P}_t + \mathbf{Q}_t. \end{aligned}$$

State and covariance are corrected by the scalar gain k_t:

$$\begin{aligned} g_{t+1} &= g_t^{+} + k_t \nu_t, \\ \dot{g}_{t+1} &= \dot{g}_t^{+} + \gamma_t \, k_t \, \nu_t, \quad \text{with } \gamma_t = 0.5, \\ \mathbf{P}_{t+1} &= \mathbf{P}_t^{+} \odot \begin{bmatrix} 1 - k_t & 0 \\ 0 & 1 - \gamma_t k_t \end{bmatrix}. \end{aligned}$$

The fixed 0.5 factor damps velocity corrections, yielding smoother trajectories.

Physiological bounds. After each update we clip to a biologically plausible range:

$$g_t \;\leftarrow\; \min\big(\text{HR}_{\max},\ \max(\text{HR}_{\min},\, g_t)\big).$$

Network architecture. All auxiliary maps f_{enc}, f_{tr}, f_{obs}, f_{noise}, f_{meas} and f_{gain} are small MLPs, while f_{temp} is a GRU.

Advantages. Our Kalman filter approach offers several advantages: (1) The two-dimensional state vector captures both HR and its rate of change, enabling better tracking of cardiac dynamics; (2) The adaptive Kalman gain responds to prediction errors, increasing when errors are high to give more weight to new measurements; (3) Neural network prediction of filter parameters allows adaptation to different exercise contexts; (4) The reduced gain for velocity updates creates smoother trajectories while maintaining responsiveness to intensity changes.

B.3 Heart Rate Models Training and Implementation

Both models were implemented in PyTorch and trained using similar procedures. We trained two GRU backbones that differ only in width and depth:

Variant	Hidden dim.	GRU layers
LARGE	128	2
SMALL	64	3

All feed-forward heads (encoders, decoders, gain/demand networks, etc.) are two-layer MLPs with LeakyReLU activations and dropout 0.1. We optimized the model parameters using the Adam optimizer (weight-decay 10^{-5}, gradient clipping $\|g\|_2 \leq 1.0$) with an initial learning rate of 0.001, reducing the learning rate by a factor of 0.5 when validation loss plateaued for 10 epochs. To prevent overfitting, we employed early stopping with a patience of 20 epochs. For the **Kalman** model we used `ReduceLROnPlateau` with `factor`=0.5 and `patience`=10, and an extended early-stopping window of 100 epochs. The **ODE** model employed the same scheduler but with `factor`=0.2, a cooldown of 3 epochs, a minimum learning rate of 10^{-6}, and the 20-epoch patience already noted above. We trained with a batch size of 32 and the default Adam momentum parameters $\beta_1 = 0.9,\ \beta_2 = 0.999$.

For both models, we used a masked mean absolute error (MAE) loss function to handle variable-length sequences:

$$\mathcal{L}_{\text{MAE}} = \frac{\sum_{b=1}^{B} \sum_{t=1}^{T} |h_{b,t}^{\text{pred}} - h_{b,t}^{\text{true}}| \cdot \text{mask}_{b,t}}{\sum_{b=1}^{B} \sum_{t=1}^{T} \text{mask}_{b,t}} \tag{1}$$

Additionally, we incorporated an auxiliary loss to encourage accurate parameter prediction:

$$\mathcal{L}_{\text{aux}} = ||\hat{\mu} - \mu_{\text{obs}}||_1 + ||\hat{\sigma} - \sigma_{\text{obs}}||_1 \tag{2}$$

where μ_{obs} and σ_{obs} are the observed mean and standard deviation of HR in each sequence, and $\hat{\mu}$ and $\hat{\sigma}$ are the predicted parameters.

The total loss was a weighted combination:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MAE}} + \lambda_{\text{aux}} \cdot \mathcal{L}_{\text{aux}} \tag{3}$$

with $\lambda_{\text{aux}} = 0.1$.

Both models achieved comparable performance, with the ODE model excelling in capturing long-term physiological trends and the Kalman filter model showing advantages in responding to rapid intensity changes. The models' complementary strengths suggest potential benefits in ensemble approaches for future work.

B.4 Kalman-based Oxygen Consumption Model

We frame oxygen-consumption dynamics as a state–estimation problem with physiologically informed constraints. Let $v_t \in \mathbb{R}$ denote the latent VO$_2$ state at time t, with $\hat{v}_t$ its filtered estimate and $\tilde{v}_{t^-}$ the trend-extrapolated prediction. As throughout the paper, learnable functions are written $f(\,\cdot\,; \boldsymbol{\vartheta})$, where the subscript on $\boldsymbol{\vartheta}$ indicates the module.

Temporal feature extraction. The network processes a window $\mathbf{x}_t \in \mathbb{R}^{B\times T\times D_{\text{in}}}$ (containing biomechanical features, HR, and temporal context):

$$\mathbf{h}_t = f_{\text{temp}}\big(f_{\text{enc}}(\mathbf{x}_t; \boldsymbol{\vartheta}_{\text{enc}})\,; \boldsymbol{\vartheta}_{\text{temp}}\big),$$

where f_{temp} is a bidirectional GRU and $\mathbf{h}_t \in \mathbb{R}^{B\times D_h}$ is the hidden representation at time t.

Process and measurement-noise prediction. Neural heads produce time-varying covariance surrogates from the hidden state:

$$\begin{aligned} \mathbf{P}_{\text{process},t} &= f_{\text{process}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\text{process}}), \\ \mathbf{R}_{\text{meas},t} &= f_{\text{measurement}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\text{measurement}}). \end{aligned}$$

Neural–Kalman update. Given $\mathbf{h}_t$, the VO$_2$ estimate is updated via

$$\hat{v}_t^{\text{KF}} = \hat{v}_{t-1} + f_{\text{gain}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\text{KF}}) \big[f_{\text{obs}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\text{obs}}) - \hat{v}_{t-1}\big]_{f_{\Delta_{\min}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\Delta})}^{f_{\Delta_{\max}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\Delta})},$$

where the adaptive gain f_{gain} and clamping limits $f_{\Delta_{\min}}, f_{\Delta_{\max}}$ are themselves MLPs, allowing data-driven confidence weighting and physiologically plausible rate-of-change bounds.

Trend/direct dual pathway. To improve robustness,

$$\begin{aligned}\tilde{v}_{t^-} &= \hat{v}_{t-1} + f_{\text{trend}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\text{trend}})\big(\hat{v}_{t-1} - \hat{v}_{t-2}\big),\\ \hat{v}_t^{\text{dir}} &= f_{\text{direct}}(\mathbf{h}_t; \boldsymbol{\vartheta}_{\text{dir}}),\\ \hat{v}_t &= \beta_t \hat{v}_t^{\text{dir}} + (1-\beta_t)\hat{v}_t^{\text{KF}},\end{aligned}$$

Implementation details. All heads f_{enc}, f_{gain}, f_{obs}, $f_{\Delta_{\min}}$, $f_{\Delta_{\max}}$, f_{process}, $f_{\text{measurement}}$, f_{trend}, f_{blend}, and f_{direct} are two-layer MLPs with LeakyReLU activations and dropout 0.1, parameterized by $\boldsymbol{\vartheta}_{\text{enc}}$, $\boldsymbol{\vartheta}_{\text{KF}}$, $\boldsymbol{\vartheta}_{\text{obs}}$, $\boldsymbol{\vartheta}_{\Delta}$, $\boldsymbol{\vartheta}_{\text{trend}}$, $\boldsymbol{\vartheta}_{\text{blend}}$, and $\boldsymbol{\vartheta}_{\text{dir}}$, respectively. The temporal module f_{temp} is a biGRU (two layers, hidden size $D_h = 128$). All outputs are trained end-to-end with an MSE loss on breath-by-breath VO_2 targets; clamping reduces gradient explosion when rapid transients exceed physiological bounds.

B.5 Training and Implementation Details

We employ the composite loss (main text) with base, dynamic, and auxiliary components:

$$\begin{aligned}\mathcal{L} = \underbrace{\mathbb{E}_t\big|\hat{v}_t - v_t\big|}_{\text{MAE}} &+ \lambda_{\text{dynamic}}\underbrace{\mathbb{E}_t\Big(\alpha\Big|\frac{d\hat{v}_t}{dt} - \frac{dv_t}{dt}\Big| + \beta\Big|\frac{d^2\hat{v}_t}{dt^2} - \frac{d^2v_t}{dt^2}\Big|\Big)}_{\text{derivative loss}}\\ &+ \lambda_{\text{aux}}\underbrace{\mathbb{E}_t\big(|\hat{\mu}_t - \mu_t| + |\hat{\sigma}_t - \sigma_t| + |\hat{\Delta}_t - \Delta_t|\big)}_{\text{moment loss}},\end{aligned}$$

The dynamic loss captures temporal patterns using both first and second derivatives:

$$\mathbb{E}_t\left(\alpha\left|\frac{d\hat{v}_t}{dt} - \frac{dv_t}{dt}\right| + \beta\left|\frac{d^2\hat{v}_t}{dt^2} - \frac{d^2v_t}{dt^2}\right|\right).$$

The auxiliary loss enforces statistical consistency between predicted and observed distributions:

$$\begin{aligned}\mathbb{E}_t|\hat{\mu}_t - \mu_t| &= |\hat{\mu} - \mu|\\ \mathbb{E}_t|\hat{\sigma}_t - \sigma_t| &= |\hat{\sigma} - \sigma|\\ \mathbb{E}_t|\hat{\Delta}_t - \Delta_t| &= |\hat{\Delta} - \Delta|,\end{aligned}$$

where μ_t is the temporal mean, σ_t is the standard deviation, and $\Delta_t = Q_{0.95}(|v_\tau - v_{\tau-1}|)$ represents the 95th percentile of absolute differences in consecutive VO_2 values.

Curriculum weights. At epoch e: $\lambda_{\text{base}} = \max(0.30,\ 1 - e/20)$, $\lambda_{\text{dynamic}} = 1 - \lambda_{\text{base}}$, $\lambda_{\text{aux}} = \min(0.30,\ 0.10 + 0.01e)$.

Model variants.

Variant	GRU hidden	GRU layers	Head MLP layers
Small	128	2	2
Large	256	4	4

Optimisation. AdamW (lr = 4×10^{-3}, weight-decay 10^{-5}) with cosine-annealing warm restarts ($\text{T}_0 = 10$, T_{mult}=2, $\eta_{\min} = 10^{-6}$). Batch 32, sequence length 60s (1Hz); gradient clipping $\|g\|_2 \leq 1 + 4e^{-e/10}$; early stopping after 50 epochs without validation-MAE improvement.

C Additional Results

C.1 Additional HR Figures

Figure 6 presents additional HR prediction using the standard prediction mode, while Figure 7 presents using the the generative mode.

Table 4: Comparison of Sequence-to-Sequence Instantaneous VO_2 Prediction for Models 256-2 and 128-4 using Leave-One-Out Cross-Validation. HR used is the predicted HR values using generation mode. Bold values indicate the better (lower) performance.

	KalmanVO$_2$ 256-2			KalmanVO$_2$ 128-4		
Runner	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
With ODE-based HR predictions (128-2)						
Runner-1	645.78	682.96	19.24	**519.27**	**565.93**	**15.74**
Runner-2	1326.20	1361.69	35.40	**463.79**	**518.90**	**12.94**
Runner-3	250.61	481.78	12.36	234.11	475.18	**12.04**
Runner-4	287.74	335.78	11.01	**128.56**	**182.10**	**6.03**
Runner-5	697.68	717.53	25.03	**549.20**	**571.91**	**19.88**
Runner-6	347.87	450.75	15.61	402.03	495.38	17.04
Runner-7	193.00	219.82	7.16	**122.36**	**160.31**	**4.70**
Runner-8	321.80	382.35	13.54	266.80	347.77	**11.98**
Runner-9	333.82	407.08	25.83	268.31	344.64	**21.32**
Runner-10	407.44	514.14	14.41	398.53	491.12	**14.23**
Aggregate	481.19	555.39	17.96	**335.30**	**415.32**	**13.59**
With Kalman-based HR predictions (128-2)						
Runner-1	747.67	781.70	21.97	**529.75**	**574.35**	**16.03**
Runner-2	1326.20	1361.69	35.40	1757.21	1804.52	46.63
Runner-3	262.15	479.42	12.59	290.56	495.09	13.59
Runner-4	439.11	500.90	15.78	**137.36**	**189.68**	**6.30**
Runner-5	697.75	717.59	25.03	**549.20**	**571.91**	**19.88**
Runner-6	383.54	473.57	16.73	463.29	550.49	18.92
Runner-7	192.96	219.79	7.15	**120.33**	**158.70**	**4.63**
Runner-8	321.80	382.35	13.54	288.13	345.58	**12.21**
Runner-9	396.71	449.44	29.59	274.75	354.99	**21.83**
Runner-10	468.62	559.96	16.36	432.30	529.91	**15.29**
Aggregate	523.65	592.64	19.41	484.29	557.52	17.53

C.2 Additional Oxygen Consumption results using Two Trained HR Models and figures

To quantify how HR–prediction quality propagates into downstream metabolic estimation, we trained two sequence-to-sequence VO_2 models that share the same architecture except for backbone size (**256-2** versus **128-4**; see Table 4. Both networks were evaluated in leave-one-runner-out cross-validation using HR signals that were *predicted* in generation mode by either our best ODE-based or Kalman-based HR model (both 128-2). The table below reports per-runner and aggregate MAE, RMSE, and MAPE. Boldface highlights the better of the two VO_2 variants under each HR source. The table shows that the results are worse when using the predicted HR compared to when using the true HR (see in the section), although the results are still well below the 20% MAPE threshold which is the considered standard.

D Ablation Study

We retrained three distinct architectures, TrendDirectVO$_2$Model, KalmanOnlyVO$_2$Model, and KalmanVO$_2$Model, under three loss configurations (base_only, base_statistical, and the full curriculum). This yielded nine experimental conditions, evaluated with consistent leave-one-runner-out splits.

D.1 Architectural Comparison

Table 5 summarizes the components that differentiate the predictors. TrendDirectVO$_2$Model employs a bidirectional GRU with learned momentum extrapolation and a blending mechanism while omitting explicit filtering. KalmanOnlyVO$_2$Model introduces a learned Kalman filter that models process and measurement variances, incorporates an observation network, and clamps innovations, but it does not include a direct VO_2 prediction head. KalmanVO$_2$Model unifies these approaches by combining the Kalman filter with a direct VO_2 head and a dynamics blending stage.

Table 5: Architectural features of VO_2 predictors.

Component	TrendDirect	KalmanOnly	Kalman (full)
Direct VO$_2$ head	✓	×	✓
Learned Kalman filter	×	✓	✓
Observation network	×	✓	✓
Dynamics blend	✓	✓	✓
Uncertainty propagation	×	✓	✓

D.2 Per runner Results

Table 6 reports MAPE for each runner across all architecture and loss combinations. Columns are grouped by architecture (Full hybrid, KalmanOnly, TrendDirect) and subcolumns correspond to the three loss functions. Bold values mark the best configuration for each runner, and the final row presents the average MAPE.

D.3 Evaluation Protocol

All experiments are evaluated in a strict sequence to sequence manner. The model receives the true HR samples rather than predicted ones. Only the first second of the first window is supplied as ground truth VO_2; for every subsequent window, the final VO_2 estimate of window i becomes the first input of window i+1, preventing error from resetting between windows.

D.4 Discussion

The full architecture with the complete curriculum achieves the lowest MAPE for nine of the ten runners and yields the best overall mean error. Its joint design captures both abrupt metabolic transitions and gradual trends. When pace increases sharply, oxygen uptake rises almost instantaneously; the direct VO_2 head accurately tracks these transients. KalmanOnlyVO$_2$Model, lacking this head, tends to over smooth and display latency. TrendDirectVO$_2$Model, although responsive, does not explicitly filter noise and therefore struggles to characterize sudden yet genuine spikes.

Our data were collected under two controlled protocols: each session was either an "all out" effort test or an incremental to

Figure 6: Ten random heart rate prediction using the *Standard* predictions mode on random sampled sessions with true HR (blue) and predicted HR (red). Each subfigure shows a single session's timeline, along with mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and correlation.

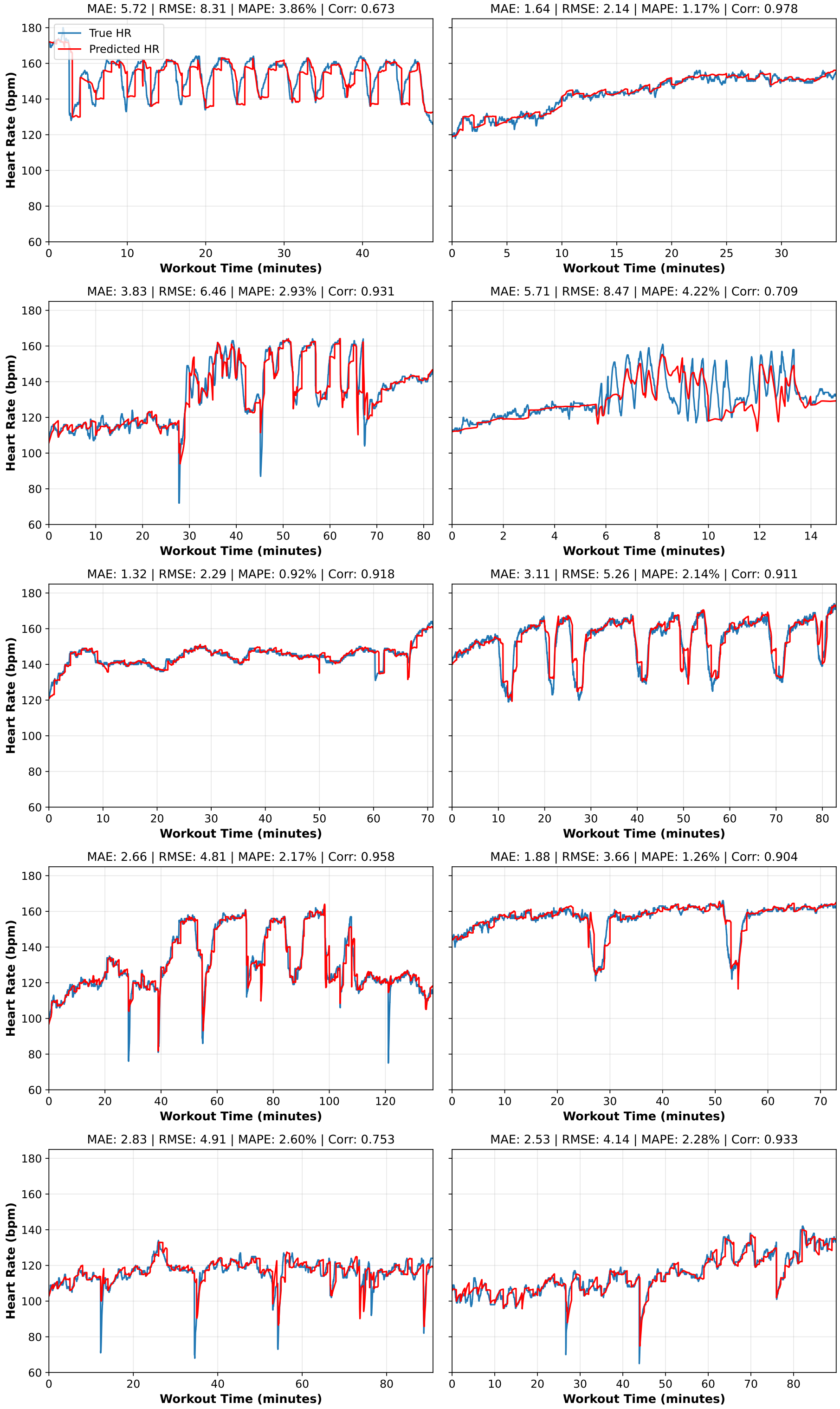


Figure 7: Ten random heart rate prediction using the *Generating* predictions mode on random sampled sessions with true HR (blue) and predicted HR (red). Each subfigure highlights a unique session with corresponding performance metrics (MAE, RMSE, MAPE, and correlation).

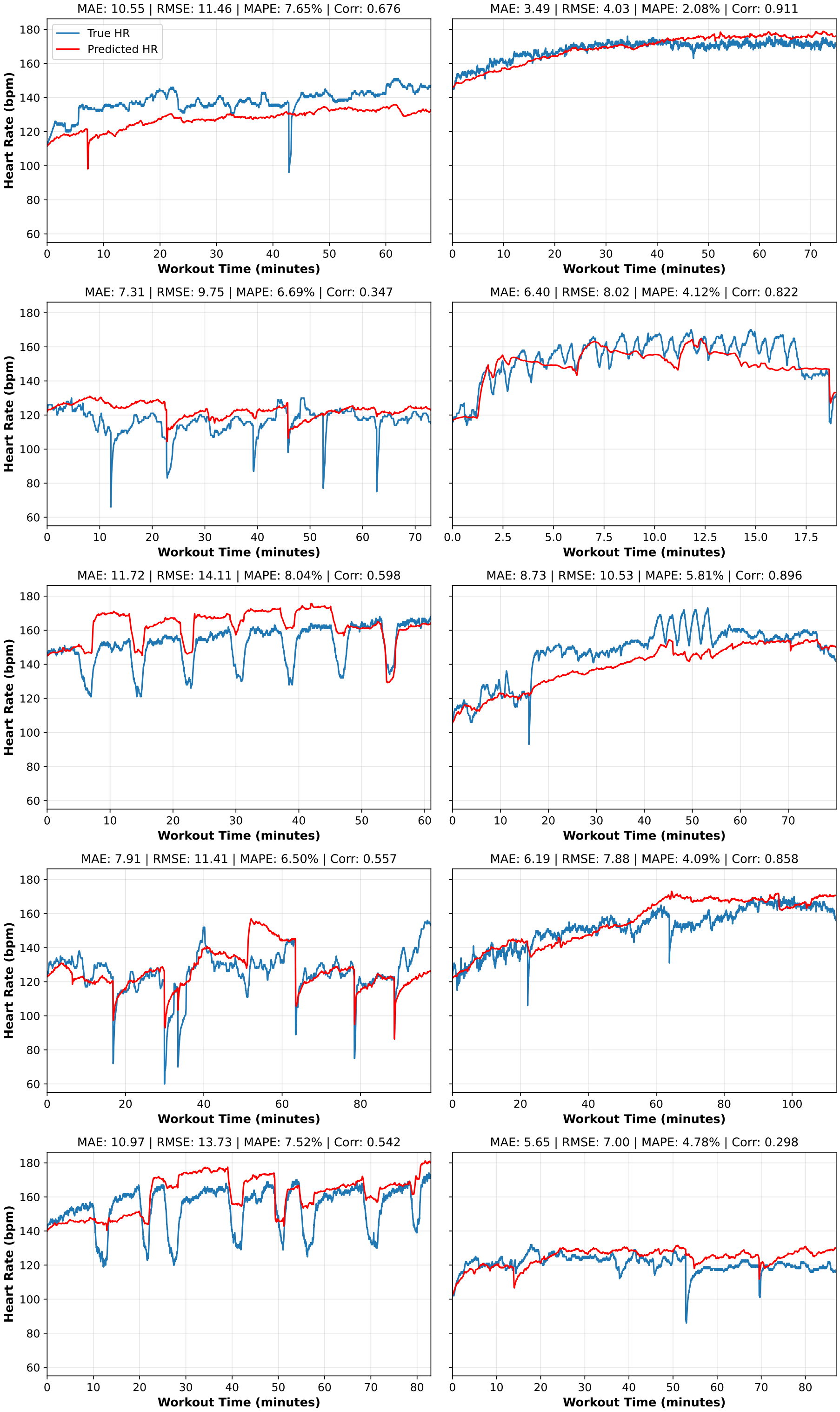


Table 6: Per runner MAPE (%) across architectures and loss configurations.

	Full (hybrid)			KalmanOnly			TrendDirect		
Runner	base_only	base_statistical	full	base_only	base_statistical	full	base_only	base_statistical	full
runner-1	4.81	5.57	11.57	4.79	4.73	4.72	**4.40**	15.05	5.44
runner-2	15.45	16.01	**11.87**	90.53	83.39	74.54	13.73	33.09	13.27
runner-3	9.39	11.02	**9.31**	80.66	66.37	65.41	14.42	10.79	9.37
runner-4	12.06	11.08	**10.31**	63.38	63.77	68.42	10.34	10.33	10.34
runner-5	21.50	17.98	**15.10**	70.79	70.94	94.55	21.40	20.62	20.17
runner-6	5.53	5.03	**4.98**	80.25	80.27	85.99	5.47	5.72	5.37
runner-7	**6.60**	30.12	23.33	22.44	22.01	20.57	7.96	27.99	10.47
runner-8	20.75	24.26	**20.25**	97.15	95.94	96.23	21.53	22.78	23.59
runner-9	**5.76**	27.01	7.14	84.13	87.47	83.49	5.88	26.36	26.79
runner-10	8.00	**4.03**	7.19	67.79	67.81	67.81	10.22	9.63	10.35
Mean	10.99	15.21	12.11	66.19	64.27	66.17	11.54	18.24	13.51

exhaustion ramp. Both protocols generate VO_2 traces that rise monotonically until failure, with few downward inflections. In such settings the bidirectional GRU in TrendDirectVO$_2$Model can look several steps forward and backward, learn a momentum based extrapolation, and implicitly smooth short horizons. The dynamics blend layer further dampens measurement noise by fusing the extrapolated value with the raw observation, effectively learning coefficients that behave like a simplified Kalman gain. Because most trajectories are dominated by steady ramps and plateaus rather than rapid oscillations, these two mechanisms capture the principal physiological variation, which explains why the trend only model approaches the full model in accuracy.

During steady state running, breath by breath variability introduces high frequency noise. The learned Kalman filter in the full model attenuates these fluctuations, yielding smoother estimates. Without a direct predictive branch, exclusive filtering suppresses peaks and allows error drift over erratic segments. The progressive curriculum further refines performance: early mean absolute error minimization aligns trajectories, intermediate dynamic consistency objectives enforce realistic multi step behavior, and final statistical penalties anchor distributional fidelity.

For runners exhibiting highly irregular VO_2 profiles, the full architecture with the complete curriculum reduces extreme overshoots by roughly one third relative to base_only training. For instance, runner−5 shows a drop in MAPE from 21.50 % (base_only) to 15.10 % (full), a 29.8 % relative improvement, while runner−2 falls from 15.45 % to 11.87 % (a 23.2 % gain). In contrast, athletes whose traces rise smoothly under the two test protocols see only marginal gains: runner−3 improves from 9.39 % to 9.31 % (below 1 % change) and runner−6 from 5.53 % to 4.98 % (about 10 %). Thus, the direct VO_2 head and uncertainty propagation act primarily as a safety net for edge cases rather than the main accuracy driver in routine conditions.

D.5 Conclusion

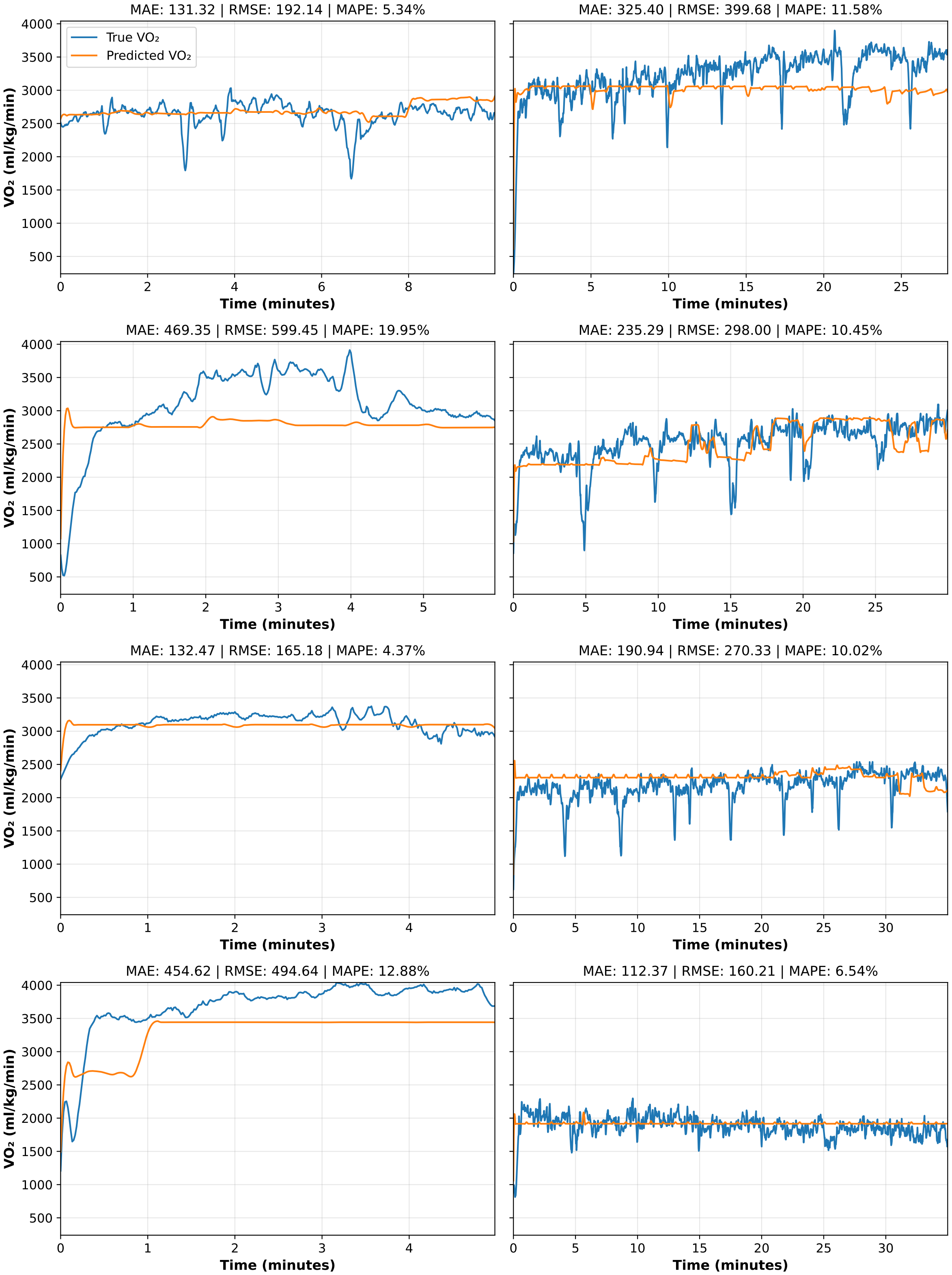
Accurate VO_2 estimation from wearable data benefits from uniting a direct prediction head that responds immediately to metabolic transitions with a learned Kalman filter that attenuates breath–by–breath noise. The staged curriculum balances pointwise accuracy, temporal coherence, and physiological plausibility. Empirically, the full architecture trained with the complete curriculum delivers the lowest mean error and mitigates extreme overshoots for runners with irregular traces, while the trend–only variant remains a competitive baseline under monotonic ramp protocols and in settings where model size or compute budget is limited. We therefore recommend deploying the full configuration when resources permit and reserving the trend Sonly model for lightweight or edge scenarios.

Looking ahead, we will extend evaluation beyond the present all-out and incremental-to-exhaustion sessions to include variable-intensity workouts such as intervals, fartlek, and tempo runs with recovery segments. These protocols generate VO_2 trajectories that rise and fall repeatedly, stressing both rapid responsiveness and robust smoothing. Under such non monotonic patterns we expect the combined architecture to deliver even larger gains over the trend only baseline, because the direct prediction head can follow sharp inflections while the learned Kalman filter suppresses transient noise between efforts.

E Additional Figures for Predicting Oxygen Consumption

Figure 8 presents additional VO_2 prediction visualizations.

Figure 8: Sequence-to-Sequence VO$_2$ Prediction Across Different Sessions using true HR values.


MAE: 131.32 | RMSE: 192.14 | MAPE: 5.34%
MAE: 325.40 | RMSE: 399.68 | MAPE: 11.58%
True VO₂
Predicted VO₂
VO₂ (ml/kg/min)
4000
3500
3000
2500
2000
1500
1000
500
0
2
4
6
8
5
10
15
20
25
Time (minutes)
MAE: 469.35 | RMSE: 599.45 | MAPE: 19.95%
MAE: 235.29 | RMSE: 298.00 | MAPE: 10.45%
1
3
MAE: 132.47 | RMSE: 165.18 | MAPE: 4.37%
MAE: 190.94 | RMSE: 270.33 | MAPE: 10.02%
30
MAE: 454.62 | RMSE: 494.64 | MAPE: 12.88%
MAE: 112.37 | RMSE: 160.21 | MAPE: 6.54%