Continuous Heart Rate Variability Estimation From PPG via State-Space Modeling

Berken Utku Demirel and Christian Holz

***Abstract*— Objective: Heart rate variability (HRV) reflects autonomic regulation and is used in cardiovascular monitoring. Photoplethysmography (PPG) is commonly used for heart rate (HR) tracking in daily life, but deriving reliable HRV from PPG is highly difficult because of motion artifacts and drift effects from variability in pulse arrival time (PAT).**

Methods: We propose a multimodal framework that combines encoders for PPG, inertial measurements, and temperature signals with a learnable state-space model for inter-beat inference. The state-space dynamics adapt to non-linear changes and PAT-related shifts. A trust gate uses predicted uncertainty to down-weight corrupted intervals.

Results: Using a single model configuration across three public datasets (DaLiA, WildPPG, BIDMC), our method consistently improves inter-beat interval accuracy and HRV indices compared to prior work. For SDNN, we reduce error by up to 80% relative to traditional peak detection, while improving agreement with ECG-derived references.

Conclusion: Uncertainty-aware multimodal observations with an adaptive state-space model (SSM) yield robust HRV estimation under real-world artifacts.

Significance: Our method enables robust HRV monitoring in realistic settings from common wearable sensors and provides strong baselines and results to support research and future applications.

Code: `github.com/eth-siplab/multimodal-hrv`

***Index Terms*— Heart rate variability, Photoplethysmography, Signal processing, Wearable devices**

I. Introduction

Heart rate variability (HRV) is a widely used noninvasive marker of cardiac autonomic activity and has been associated with multiple indicators of cardiovascular health. Variations in beat-to-beat intervals reflect the balance between sympathetic and parasympathetic nervous system inputs [1], making HRV a sensitive measure of autonomic regulation [2]. Lower or abnormal HRV has been linked to elevated risk of cardiovascular morbidity and mortality, as well as other systemic health outcomes [3]. For instance, alterations in HRV have been associated with diabetes [4], mental health disorders such as depression and anxiety [5], and even sleep quality [6].

PPG has become the dominant modality, enabling continuous monitoring of blood volume changes in everyday settings [7]–[10]. The widespread adoption of smartwatches means that a large portion of the population collects PPG during daily activities [11]. This accessibility has positioned PPG as the foundation for real-time heart rate (HR) estimation and monitoring outside clinical environments [8]. Therefore, extensive research has focused on improving the accuracy of HR tracking from PPG for everyday use [12]–[15].

However, estimating HRV in the real-world is considerably more challenging than HR. HRV analysis requires the detection of beat timings; even small inaccuracies in identifying beat intervals lead to large errors in variability indices (e.g., RMSSD), making them sensitive to motion artifacts [16]. While deep learning approaches have been developed to enhance HR estimation [17], [18], most rely on feed-forward architectures that process windows in isolation. Because these models do not enforce temporal continuity across windows, small errors accumulate over time, producing drift [15]. This failure of temporal consistency renders fine-grained HRV analysis unreliable in high-motion conditions.

Motion artifacts and signal distortions are the primary challenge for HRV estimation from wearable PPG, as they obscure beat-to-beat dynamics and undermine methods that depend on clean waveforms [19]–[21].

Unlike in controlled settings, wearable data collected in the wild is contaminated by motion, posture changes, and environmental noise, making accurate peak detection difficult [14]. This challenge highlights the need for assessing the quality of a given interval and adaptively estimating HRV.

In this work, we propose a unified framework that couples neural networks with a learnable state-space model. Unlike static filters, it updates its dynamics and observation uncertainty continuously using contextual signals (e.g., IMU, temperature) to handle motion and physiological transitions. We evaluate our method on three public datasets, including the largest collected in naturalistic conditions, WildPPG and DaLiA. Comparing against traditional techniques, our method substantially reduces error, demonstrating that robust HRV monitoring is feasible even in high-motion daily life scenarios.

We summarize our main contributions as follows:

- a deep learning–SSM hybrid that models beat intervals with second-order dynamics to capture inter-beat interval oscillations and PAT-driven variability,
- an adaptive trust gate that leverages learned uncertainty to suppress corrupted intervals for improving validity, and
- extensive experiments in which our method outperformed traditional approaches by more than 80%, providing robust HRV estimates under everyday conditions.

The authors are with the Department of Computer Science, ETH Zürich, Stampfenbachstrasse 48, 8092 Zurich, Switzerland (e-mail: berken.demirel@inf.ethz.ch, christian.holz@inf.ethz.ch).

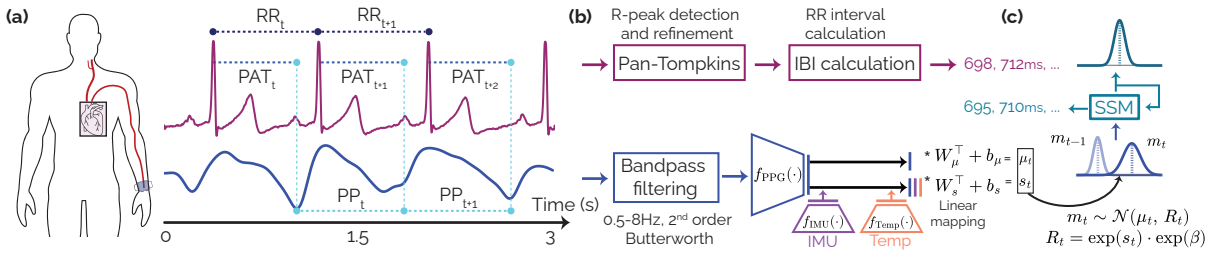


Fig. 1: Our proposed method directly estimates HRV based on signals from wearable devices in daily-life conditions. **(a)** PPG and ECG signals, their corresponding inter-beat intervals (IBIs), and their relationship via pulse arrival time (PAT). **(b)** Pre-processing and neural networks, where PPG and auxiliary modalities are used to estimate the IBI distribution, parameterized by mean (μ) and log-variance (s_t) under a negative log-likelihood loss. **(c)** State-space modeling via a parametrizable Kalman filter, which captures IBI transitions while incorporating environmental variables and the learned observation uncertainty.

II. Related work

A. HRV Estimation

Estimating HRV from blood volume changes has been explored using both contact-based and remote PPG [19], [22]–[24]. Early studies primarily relied on contact PPG placed on the fingertip to extract inter-beat intervals, demonstrating the feasibility of HRV analysis without ECG [19]. In a related direction, Ballistocardiogram (BCG) signals have also been employed to detect beat timings and measure HRV [20].

More recently, researchers have investigated remote PPG, which extracts blood volume pulse signals from facial videos [22], [23]. These approaches enable contactless HRV monitoring with applications in telehealth. However, both contact and remote PPG methods assume high-quality waveforms obtained under stationary conditions. For example, authors in [23] achieved accurate HRV estimation from rPPG but required subjects to remain still under consistent lighting.

Despite these advances, robust HRV estimation from PPG in real-world conditions remains largely underexplored. Most existing approaches rely on clean morphological features of the PPG waveform [25], [26], which are highly sensitive to motion artifacts and environmental noise [21]. Others require ECG to detect peaks [25], [27], a modality rarely available in wearable devices. Consequently, methods developed for laboratory environments often fail to generalize to data collected in the wild, where motion and variability alter the waveform [28]. Recent efforts have sought to address these challenges by using signal quality indices [29] or deep learning–based denoising techniques [30], [31], yet these approaches remain limited under dynamic conditions. Along similar lines, prior work either combines deep denoising with post-hoc PRV tracking [21] or directly regresses HRV indices from processed PPG [32], [33]; however, both remain limited in enforcing beat-level temporal consistency under nonstationary motion. Moreover, rigorous evaluations and open-source implementations for HRV estimation are still lacking, limiting comparability and leaving open the need for methods to provide reliable HRV indices.

These limitations motivate the development of our method, which is designed to operate under noisy, unconstrained conditions while incorporating complementary sensing modalities such as inertial measurements to provide motion context and improve HRV estimation in real-world environments.

B. State-Space Approaches

State-space models such as the Kalman filter have been used for modeling physiological dynamics and handling noisy signals [34]–[36]. These approaches provide a principled way to represent latent physiological states and update them in the presence of noise or uncertainty, including motion artifacts. For example, authors in [36] parameterize Kalman filter updates using respiratory signals. However, most SSM implementations are relatively straightforward, focusing on generic fusion without explicitly considering the temporal structure or nonlinear relationship between different modalities. In particular, they do not fully exploit how HRV can be modeled as a dynamic process. As a result, their applicability to HRV estimation from wearable PPG in real-world conditions remains limited. In this work, we propose a tailored deep learning–based state-space framework for HRV estimation, incorporating complementary sensor modalities to achieve robustness in real-world settings.

III. Method

We estimate HRV by combining a per-segment neural network with a lightweight, learnable SSM that filters inter-beat intervals (IBIs) over time. PPG-derived intervals differ from ECG-based RR intervals due to variations in pulse arrival time (PAT) [37], as well as distortions introduced by motion and environmental factors. Although PAT is related to IBIs, explicit estimation is infeasible without ECG signals. Our approach consists of two stages: (i) learning a per-window IBI distribution from PPG and auxiliary modalities, and (ii) filtering the resulting IBIs with a causal, parameterized Kalman filter. Crucially, to address the non-linear and oscillatory nature of HRV, the filter's dynamics and hemodynamic offsets are not fixed but adapted at every step based on contextual features. The following sections describe each component in detail.

A. Pre-processing

PPG signals (green channel, 128 Hz) were segmented into 8-second windows with 6-second overlap for IBI estimation. We found shorter windows (4 s) too sensitive to noise, while longer windows reduced the model's ability to track rapid changes. Each segment was bandpass-filtered using a zero-phase second-order Butterworth filter (0.5–8 Hz) to suppress

baseline drift and high-frequency noise. ECG signals were processed using the Pan–Tompkins algorithm [38] to detect R-peaks, from which reference RR intervals were derived.

Inertial measurement unit (IMU) signals (auxiliary modality) were segmented in the same way as PPG and resampled to the same frequency (128 Hz), without additional filtering or normalization. Temperature signals (auxiliary modality, 1 Hz) were also segmented, but since variance within short windows was small, each segment was represented by its average value.

B. Estimating IBIs

We used modality-specific neural encoders to extract features from each signal type. For PPG, we employed a three-block UNet [39], denoted $f_{\text{PPG}}(\cdot)$. IMU signals were processed with a two-layer convolutional network ($f_{\text{IMU}}(\cdot)$; kernel size 9, 64 channels), and temperature signals with a two-layer multilayer perceptron ($f_{\text{Temp}}(\cdot)$; dimensions 1–32–64). We also experimented with convolutional networks and LSTM backbones, but the UNet consistently yielded the best performance, likely due to its multi-scale feature extraction capability [40].

The models were trained using the negative log-likelihood (NLL) on IBI prediction. Specifically, the predicted mean μ was obtained from a linear projection of PPG features, while the log-variance s was obtained from a linear projection of the concatenated features from all modalities (Figure 1). This design enables the model to rely on the PPG waveform for IBI estimates while leveraging auxiliary signals for predictive uncertainty, e.g., higher uncertainty in high-motion segments.

C. State-Space for Refining IBIs

PP intervals are generally measured as foot-to-foot distances in the PPG [41], while pulse arrival time corresponds to the delay between ECG R-peaks and PPG foot points (Figure 1). These timing quantities are related by Equation 1.

$$PP_t - RR_t = (PAT_{t+1} - PAT_t), \tag{1}$$

which shows that discrepancies between PPG and ECG-derived inter-beat intervals arise from changes in PAT. Importantly, this implies that even with a perfectly clean PPG waveform and accurate detection of foot points, HRV estimation remains challenging unless the variability in PAT is also accounted for. This observation motivates the use of a latent representation that captures both the underlying cardiac intervals and peripheral delays for accurate IBI estimations.

a) Latent state: To capture the quasi-periodic nature of HRV without over-smoothing, we model the latent cardiac rhythm as a second-order system. The state vector is defined as $\mathbf{x}_t = [RR_t, RR_{t-1}, RR_{t-2}]^\top$, representing the current and previous inter-beat intervals. Unlike simple first-order models, this formulation allows the filter to represent oscillatory phenomena such as respiratory sinus arrhythmia. The 3D state also lets us write the second-order dynamics as a first-order Markov process, where at each step, the filter updates RR_t and shifts the lagged components forward.

b) Dynamics: The latent RR trajectory evolves as a second-order autoregressive process. While standard Kalman filters assume fixed linear dynamics, physiological transitions are complex and state-dependent. Therefore, we employ a *Linear Time-Varying (LTV)* formulation where the transition parameters are predicted by the neural network at every step:

$$RR_t = \alpha_{t-1}\, RR_{t-1} + \alpha_{t-2}\, RR_{t-2} + w_t, \qquad w_t \sim \mathcal{N}(0,\, q_t). \tag{2}$$

Here, α_t is the transition coefficient and $q_t \geq 0$ specifies the instantaneous process noise. Both parameters are predicted from auxiliary features (IMU, temperature). This allows the model to switch regimes. For example, increasing process noise q_t during activity onset to track HR accelerations, or adjusting α to lock into stable sinus rhythm during rest.

c) Observations: The observation model links the latent state RR_t to the inter-beat interval m_t which is estimated from PPG. Crucially, we do not treat PAT changes as random noise. Instead, we model the measurement as the sum of the true cardiac interval (RR_t), a learned hemodynamic offset (capturing PAT), and measurement noise (v_t), as in Equation 3.

$$m_t = RR_t + \delta_{\text{PAT},t} + v_t, \qquad v_t \sim \mathcal{N}(0, R_t), \tag{3}$$

where $\delta_{\text{PAT},t}$ is a learned *hemodynamic offset* predicted from auxiliary features, and R_t is the observation uncertainty:

$$R_t = \exp(s_t) \cdot \exp(\beta), \tag{4}$$

where s_t is the log-variance from the encoders and β is a learnable scale. In this formulation, $\delta_{\text{PAT},t}$ accounts for systematic shifts in PAT caused by posture changes or physical exertion (measurable via IMU), while v_t captures random measurement errors. This disentanglement aims to allow the model to recover the underlying rhythm RR_t even when peripheral vascular dynamics distort the mechanical timing.

d) Learning the parameters: Traditional state-space models rely on hand-tuned parameters which are difficult to set optimally for diverse real-world conditions. Instead, we learn the parameters in a data-driven manner. A lightweight NN head $g_\theta(\cdot)$ predicts the transition parameters (α_t, q_t) and the hemodynamic offset $\delta_{\text{PAT},t}$ from auxiliary features. By learning these parameters across varied activities and subjects, the model adapts its dynamics to different motion and noise contexts for robustness in real-world environments.

D. Training

Our method is trained in two stages. In the first stage, the models $f_{PPG}(\cdot)$, $f_{IMU}(\cdot)$, and $f_{Temp}(\cdot)$ are optimized with a negative log-likelihood loss $\mathcal{L}_{\text{NLL}}$ using per-window IBI labels to estimate the mean μ_t and log-variance s_t. In the second stage, these models are frozen and used to generate (μ_t, s_t) over full subject sequences. End-to-end training can admit degenerate solutions where the observation network and the SSM co-adapt to maximize likelihood by inflating gains or collapsing uncertainty, rather than improving beat timing. In experiments, we observed this as divergence of (α, β) and overconfident s_t, even with gradient clipping and parameter constraints. Thus, we adopt a two-stage training scheme.

The SSM is optimized with a composite loss with three objectives: (1) accuracy, measured by the ℓ_1 error between predicted ($\hat{x}$) and reference (x) RR intervals; (2) innovation consistency, which ensures that the filter's predicted uncertainty S_t is consistent with the actual squared prediction error e_t^2. We penalize deviations of their ratio from one, $\mathbb{E}[|\frac{e_t^2}{S_t} - 1|]$, so that the model neither underestimates nor overestimates its confidence. And, (3) physiological plausibility which encourages the predicted RRs ($\hat{x}$) to remain within 300–1900 ms. Formally, the loss is defined as in Equation 5.

$$\mathcal{L}_{\text{SSM}} = \|\hat{x}_t - x_t\|_1 + \lambda_{\text{innov}}\, \mathbb{E}\left[\left|\frac{e_t^2}{S_t} - 1\right|\right] + \lambda_{\text{bounds}}\, \mathcal{P}(\hat{x}_t), \quad (5)$$

where $\mathcal{P}(\cdot)$ applies a soft penalty (quadratic outside the valid range) to RR intervals that fall outside physiologically plausible limits. The weights $\lambda_{\text{innov}} = 0.05$ and $\lambda_{\text{bounds}} = 0.5$ were chosen through a grid search on the training set and then fixed across all datasets, without dataset-specific retuning to have a realistic evaluation.

To stabilize optimization, a short warm-up period blends the learned parameters toward conservative defaults while still updating all prediction heads of the state space model. Both training and evaluation are performed causally, ensuring that predictions depend only on past and present observations.

For Stage 1, the models were trained with a learning rate of 5e-4, batch size 128 for 100 epochs. In Stage 2, SSM was optimized with a learning rate of 1e-3, batch size 256 for 15 epochs. We used the Adam optimizer in both stages. Early stopping on the validation was applied in Stage 1 with a patience of 15, whereas in Stage 2 no early stopping was used since the smaller model showed no signs of overfitting.

E. Quality gating for corrupted intervals

Wearable PPG is prone to motion artifacts [17]. Thus, we use a trust gate based on three complementary failure modes. First, *statistical inconsistency* is monitored via the normalized innovation squared, $z_t = |e_t|/\sqrt{S_t}$. Large deviations of z_t^2 indicate motion or ectopic beats. Since HRV analysis requires Normal-to-Normal intervals [42], gating these outliers is necessary to preserve validity. Second, *aleatoric uncertainty* ($\tilde{r}_t$) is derived from the network's predicted variance R_t to capture intrinsic low-SNR. Third, *physiological plausibility* ($\tilde{d}_t$) penalizes absolute beat-to-beat jumps $|\Delta RR_t|$. Beat-to-beat RR changes are physiologically limited, as autonomic control mechanisms modulate over multiple beats rather than within a single cycle [16], [42]. Thus, abrupt jumps in consecutive RR intervals are indicative of artifacts rather than variability. We combine these metrics into a scalar anomaly score.

$$\text{score}_t = w_1 z_t + w_2 \tilde{r}_t + w_3 \tilde{d}_t, \quad (6)$$

where weights are set to $\mathbf{w} = [1.0, 0.5, 0.5]$ to prioritize statistical consistency (z_t). We set these weights based on validation set performance and showed minimal sensitivity within a ± 0.2 range. Rather than enforcing a fixed exclusion, we employ an adaptive rejection strategy. We calibrate a global threshold τ (0.73) on the validation set to maximize the coverage ($\sim$ 0.82) accuracy tradeoff. Within each cross-validation fold, we reserve a random 10% subset of the training data as a validation set and use it to tune $\mathbf{w}$ and τ; no test subjects are used for calibration. For test, the resulting retained coverage was $\sim$0.80 across datasets. Because the score distribution shifts with motion intensity, the same threshold functions differently across conditions to retain good intervals while masking intense motion segments. Since standard HRV analysis is defined on normal-to-normal intervals, we treat gated-out windows as abnormal/corrupted and exclude them from HRV computation. The resulting technique ensures only valid segments contribute to the HRV estimate.

IV. Experiments

We evaluate both the performance and the generalization of methods under diverse conditions, ranging from resting periods to extreme motion. To ensure fairness, we keep all parameters (architectures, learning rates, and loss weights) of our method fixed across datasets and subjects to assess generalization without tuning to specific datasets or individuals.

A. Datasets

We use three public datasets: DaLiA [14], WildPPG [17], and BIDMC [43]. DaLiA and WildPPG are important, as they are the largest datasets collected in real-life and thus provide a strong benchmark for HRV under motion and environmental variability. We also include BIDMC, since its clean clinical recordings allow us to validate methods without motion.

DaLiA: DaLiA consists of approximately 2 hours of recordings from 15 participants (total $\approx$30 hours) during naturalistic activities such as walking, running, and cycling [14].

WildPPG: WildPPG contains 13.5 hours of recordings from 16 participants (total $\approx$216 hours) during outdoor activities, capturing PPG and ECG under real-world conditions [17]. The dataset includes substantial noise due to motion, temperature fluctuations, and environmental changes, making it especially valuable for assessing algorithm robustness in real-world.

BIDMC: BIDMC dataset consists of 8-minute resting-state recordings from hospital patients [43]. PPG waveforms in this dataset are extremely clean compared to real-life recordings, making it well-suited as a proof of concept to demonstrate if methods perform reliably with minimal motion artifacts.

B. Evaluation

To ensure a fair assessment of our method, we adopt cross-validation (CV) tailored to the size and characteristics of each dataset. For DaLiA and WildPPG, we perform 5-fold CV at the subject level to balance robustness and computational cost since these datasets are larger in scale. For BIDMC, we apply 10-fold CV, similar to the previous work [40]. In all cases, CV is conducted at the subject level, ensuring that test subjects are excluded from the training. For learning models, we trained with three different random seeds and report the mean performance. If the variance across seeds was negligible relative to the mean, we omit it from the tables.

TABLE I: MAE (ms) between ECG and PPG-derived HRV indices over 5 and 10-min windows.

Metric	Method	DaLiA 5-min	DaLiA 10-min	WildPPG 5-min	WildPPG 10-min	BIDMC 5-min
SDNN	F2F	29.35±17.3	22.32±16.2	75.29±35.5	71.45±36.0	9.27±15.8
	Elgendi	44.66±20.8	36.91±19.4	91.64±37.3	111.0±39.3	11.26±24.2
	DCL	23.21±10.4	19.20±11.6	32.25±7.25	31.74±6.90	15.01±8.7
	Ours	**18.12±4.57**	**14.23±5.37**	**14.96±7.12**	**14.52±9.76**	**4.08±3.8**
	Gain (%)	21.9	25.9	53.6	54.2	56.0
RMSSD	F2F	39.08±18.3	38.80±17.5	91.35±40.3	91.38±41.0	11.50±17.8
	Elgendi	64.64±22.8	66.20±22.8	111.0±39.3	111.2±39.7	12.26±23.9
	DCL	31.12±10.8	33.56±10.5	36.15±5.46	36.45±5.31	19.79±13.6
	Ours	**28.50±6.12**	**28.90±8.57**	**13.81±10.5**	**14.61±10.8**	**6.90±8.8**
	Gain (%)	8.4	13.9	61.8	59.9	40.0

C. Metrics

We evaluate performance by comparing IBIs derived from our method against those obtained from reference ECG. For beat-level accuracy, we report the mean absolute error (MAE) and Pearson correlation coefficient (r).

Time-domain HRV indices: We compute the standard deviation of Normal-to-Normal intervals (SDNN) and the root mean square of successive differences (RMSSD). These metrics are calculated over non-overlapping 5 and 10-minute windows. We report HRV indices only for the 5-minute window for BIDMC as it has 8-minute recordings.

Frequency-domain HRV indices: We compute low-frequency (LF, 0.04–0.15 Hz) and high-frequency (HF, 0.15–0.40 Hz) power from the IBIs using Welch's method [44]. We report LF and HF in normalized units (i.e., LF/(LF+HF); unitless). Agreement with ECG-derived indices is obtained using Bland–Altman bias and limits of agreement (LOA) [37].

D. Baselines

We compared our method against three baselines. First, a traditional signal processing pipeline [26], [28], where PPG peaks are detected, foot points identified by differentiation [37], and IBIs calculated as foot-to-foot (F2F) distances. In addition to the signal-processing baseline, we include (i) the Elgendi rule-based peak detector [45] (via NeuroKit2 [46]) and (ii) a deep learning baseline using a CNN-LSTM (DCL) model commonly used for PPG [15] (without state-space refinement).

We applied standard IBI cleaning because traditional PPG-based estimates for HRV are extremely noise sensitive. First, we keep RR intervals only in 300–2000 ms and remove outliers that deviate by more than $\pm 50\%$ from the window median. Second, we quality-gate each PPG window by SNR after a 0.5–8 Hz band-pass, treating 0.7–3.0 Hz as signal and the rest (0.15–0.5 Hz, 3–8 Hz) as noise [14], [40]. Per subject, we retain the top 80% highest-SNR windows (equivalently, SNR $\geq \tau$ dB) and discard the rest before IBI estimation and aggregation. These steps reduce spurious detections and corrupted intervals; without them, performance drops significantly.

V. Results

We evaluated our method across datasets using IBI estimates as well as HRV indices in both the time and frequency domain.

For HRV indices (Tables I and III), our method reduces estimation error by more than 60% on average (up to 84.88%). The improvements can be explained by the gains observed at the IBI level (Table II), where our method achieves up to 50% performance improvements. These gains are especially pronounced in the real-world datasets, highlighting the robustness of our approach under challenging conditions.

We also present Bland–Altman and correlation plots for comparison (Figures 2 to 5). The traditional F2F shows low level of agreement, whereas our approach yields narrower limits and reduced systematic error. These results demonstrate that our method improves accuracy and provides HRV indices aligned with the ECG, while extending usable monitoring time. We observe that error rates for SDNN and RMSSD can be very high for traditional signal processing methods. This is largely due to the fact that PPG in real-world conditions is susceptible to noise. Even after applying filtering and window-rejection techniques, many corrupted segments remain, as the distortions are not always easily distinguishable. Because HRV indices are derived from IBIs, any error compounds over the window, leading to large errors in SDNN, RMSSD, and frequency-domain measures. In other words, while filtering may remove the most corrupted intervals, the remaining misestimations accumulate, amplifying errors. This highlights both the sensitivity of HRV to small inaccuracies and the need for robust modeling approaches in real-world settings.

However, our method adopts a gating mechanism to quantify the reliability of each interval using indicators (innovation statistics, observation variance, and beat-to-beat stability). Rather than discarding segments based on fixed heuristics, the

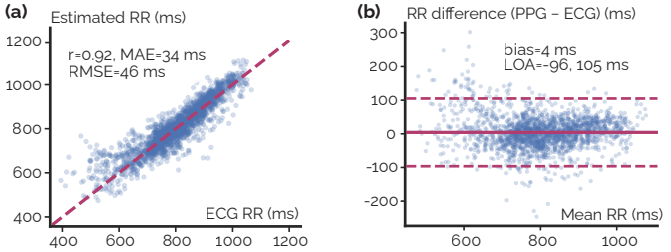


Fig. 2: **(a)** Correlation and **(b)** Bland-Altman plot obtained with our method for WildPPG Subject 1. Approximately 80% of intervals were preserved after gating, resulting in a correlation of r ≈ .92 and a Bland–Altman bias of 4 ms compared to ECG IBIs.

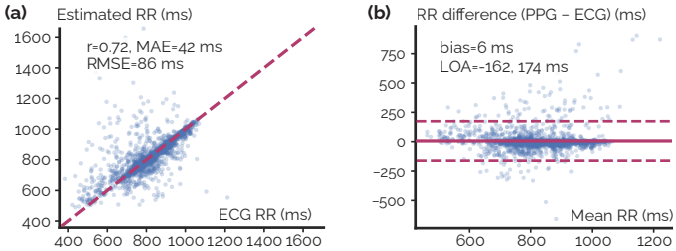


Fig. 3: **(a)** Correlation and **(b)** Bland-Altman plot obtained with traditional processing methods for WildPPG Subject 1. Approximately 80% of intervals were preserved after eliminating noisy segments, resulting in a correlation of ≈0.72 with ECG.

TABLE II: Comparison of PPG based IBI estimation performance in ms between methods across datasets in three metrics.

Method	DaLiA			WildPPG			BIDMC		
	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r
F2F	89.0±50.2	123.6±59.0	0.50±0.12	118.4±40.2	179.5±46.5	0.25±0.21	16.23±31.8	27.1±36.2	0.50±0.35
Elgendi	74.0±25.7	120.4±34.3	0.63±0.20	147.7±55.8	224.4±61.2	0.15±0.17	12.41±24.8	27.5±43.8	0.59±0.40
DCL	67.0±10.3	90.1±15.4	0.70±0.10	105.7±20.8	144.43±23.9	0.53±0.10	43.55±10.5	51.4±22.5	0.34±0.18
Ours	**60.0±7.3**	**83.8±8.4**	**0.81±0.11**	**89.4±17.1**	**113.5±25.6**	**0.60±0.11**	**8.65±3.12**	**13.2±10.4**	**0.91±0.10**
Gain (%)	10.4	7.1	15.7	15.4	21.4	13.2	30.3	51.3	54.2

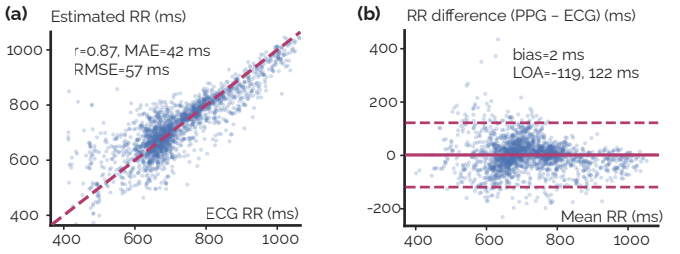


Fig. 4: **(a)** Correlation and **(b)** Bland-Altman plot obtained with our method for DaLiA Subject 1. 82% of intervals were preserved after gating, resulting in strong agreement with ECG (correlation r ≈ .87, Bland-Altman bias of 2 ms).

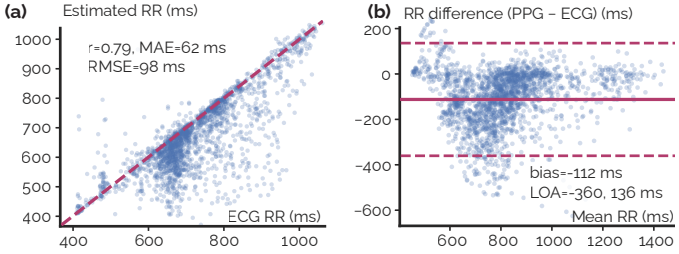


Fig. 5: **(a)** Correlation and **(b)** Bland-Altman plot obtained with traditional processing methods for DaLiA Subject 1. With 82% of intervals, the correlation reaches r ≈ .79 with ECG-derived ground-truth IBIs.

gating adapts to the data distribution and selectively masks corrupted intervals before HRV computation to improve both IBI accuracy and the robustness of derived metrics.

Overall, our method shows consistent gains over prior work. By combining deep learning with a tailored SSM, we achieve lower IBI errors, stronger agreement with ECG across HRV indices, and improved robustness in realistic conditions.

A. Ablation Studies

We performed ablation studies to evaluate the contribution of each component, with results shown in Tables IV and V. First, removing the SSM (w/o SSM) and relying solely on NN predictions increased the IBI error by ≈20% and reduced the alignment with ECG-derived HRV indices (SDNN, RMSSD) ≈50%. We also replaced the SSM with a 2-layer causal Transformer (w/ Trans.), but it did not close this gap, suggesting that structured uncertainty-aware filtering (rather than generic sequence modeling) is key to robustness.

Second, excluding auxiliary modalities (w/o AM) such as inertial and temperature signals caused a drop in performance, especially in DaLiA and WildPPG. In WildPPG, temperature information was particularly important, as recordings were collected in extreme environments such as mountains [17]. Under these conditions, removing auxiliary modalities increased error by 25–30% and substantially reduced agreement. Finally, replacing the learned transition and noise parameters (α_t, q_t, R_t, and $\delta_{\text{PAT},t}$) with constants led to a clear drop in performance, with errors rising by up to 100%. In this ablation, we set $\alpha_{t-1} = 1$, $\alpha_{t-2} = 0$ (random-walk), $q_t = 0.01$ (low process noise), $R_t = 10$ (fixed observation variance), and $\delta_{\text{PAT},t} = 150$ ms (mean hemodynamic offset). The degradation confirms that learning state-space parameters is essential, and manually chosen constants cannot capture the variability. Overall, ablation studies show that each component contributes to the overall performance.

Computational Efficiency: Our backbone (3-block UNet) runs in real time on mobile-class hardware [40]. Our SSM adds ≈ 5k MACs (for F=10, hidden 64), which is about ≥10× cheaper than a single UNet pass over the same segment. Overall, our model has ∼305k parameters. In addition, our preprocessing is lightweight. Because the SSM is causal and low-dimensional, it adds negligible latency. Our pipeline thus supports real-time, on-device deployment on wearables.

TABLE III: Bland–Altman analysis of frequency-domain HRV indices (bias [95% LOA]).

Metric	Method	DaLiA 5-min	DaLiA 10-min	WildPPG 5-min	WildPPG 10-min	BIDMC 5-min
LF	F2F	-14.8 [-29.8, 3.6]	-14.6 [-25.3, -2.1]	-13.4 [-41.2, 29.7]	-11.6 [-37.6, 27.7]	-13.6 [-29.8, 3.6]
	Elgendi	-19.4 [-36.3, 2.3]	-19.5 [-32.4, -4.0]	-13.4 [-41.2, 29.2]	-11.8 [-37.3, 26.1]	10.6 [0.02, 23.4]
	DCL	-17.4 [-32.1, 0.6]	-17.3 [-28.6, -4.1]	-7.77 [-37.3, +37.9]	-5.62 [-33.8, 37.0]	-24.0 [-35.9, -8.78]
	Ours	**-2.5 [-9.6, 5.2]**	**-2.3 [-7.5, 3.2]**	**8.0 [-14.9, 39.8]**	**9.7 [-13.2, 41.4]**	**8.6 [-5.12, 25.4]**
	Gain	83.1%	84.2%	3.0%	72.6%	18.9%
HF	F2F	17.5 [-3.3, 42.9]	17.2 [2.3, 34.4]	16.8 [-19.0, 70.9]	14.8 [-16.8, 61.3]	29.6 [17.0, 44.3]
	Elgendi	24.3 [-1.9, 58.0]	24.5 [4.6, 48.7]	16.9 [-18.9, 70.6]	15.1 [-16.1, 60.1]	-4.24 [-11.8, 4.63]
	DCL	21.37 [-0.4, 48.1]	21.2 [4.5, 40.8]	10.3 [-23.5, 61.6]	8.18 [-22.3, 53.3]	61.2 [36.7, 92.0]
	Ours	**2.6 [-4.9, 10.6]**	**2.4 [-3.0, 8.1]**	**-6.8 [-25.0, 18.2]**	**-8.0 [-25.5, 15.8]**	**1.10 [-8.27, 13.2]**
	Gain	85.1%	86.0%	34.0%	2.2%	74.1%

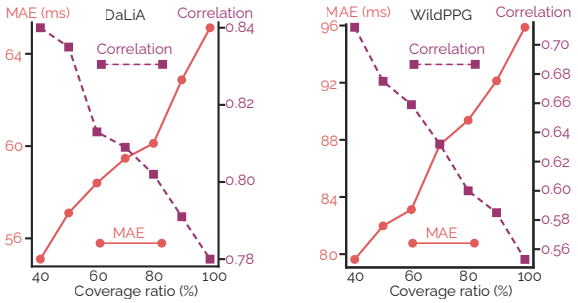
TABLE IV: Ablation experiments of our method in three datasets compared with estimated IBIs from ECG.

Method	DaLiA MAE	DaLiA RMSE	DaLiA r	WildPPG MAE	WildPPG RMSE	WildPPG r	BIDMC MAE	BIDMC RMSE	BIDMC r
Ours	**60.06**	**83.85**	**0.81**	**89.37**	**113.5**	**0.60**	**8.65**	**13.21**	**0.91**
w/o SSM	73.12	98.71	0.69	115.3	167.9	0.45	12.83	17.90	0.74
w/ Trans.	110.8	137.5	0.60	178.5	200.1	0.37	20.17	28.68	0.55
w/o AM	78.33	100.1	0.61	132.1	178.3	0.39	8.65	13.21	0.91
w/o LP	89.61	120.4	0.57	143.7	190.2	0.33	18.29	23.41	0.54
Gain	17.86%	15.05%	17.39%	22.49%	32.40%	33.33%	32.58%	26.20%	22.97%

TABLE V: Ablation experiments of our method.

Metric	Method	DaLiA 5-min	DaLiA 10-min	WildPPG 5-min	WildPPG 10-min	BIDMC 5-min	BIDMC 10-min
SDNN	Ours	**18.12**	**14.23**	**14.96**	**14.52**	**4.08**	—
	w/ Trans.	54.73	68.96	43.04	46.80	10.26	—
	w/o SSM	35.61	33.47	42.57	40.13	6.80	—
	w/o AM	48.14	43.56	50.43	49.01	4.08*	—
	w/o LP	52.67	50.12	61.93	60.88	9.43	—
RMSSD	Ours	**28.50**	**28.90**	**13.81**	**14.61**	**6.90**	—
	w/ Trans.	30.42	31.08	29.00	30.36	12.20	—
	w/o SSM	43.87	44.33	22.74	24.05	9.73	—
	w/o AM	55.21	56.01	30.12	30.89	6.90*	—
	w/o LP	60.09	61.73	40.58	41.71	12.54	—

* Single modality dataset



(a) DaLiA—15 subjects (b) WildPPG—16 subjects

Fig. 6: Relationship between coverage (included segments after quality gating) and IBI estimation accuracy, reported as MAE and correlation with intervals obtained from ECG.

VI. Discussion

Experiments show that our method improves HRV estimation over traditional techniques and learning baselines that ignore the temporal structure. The observed performance gains highlight the effectiveness of combining NNs with a learnable SSM, which enables handling of noise and physiological variability. Ablation studies confirm the importance of modeling transitions between IBIs and incorporating auxiliary modalities for improved robustness in real-world conditions.

a) Comparison with Previous Approaches: Conventional HRV estimation from blood volume pulse signals relies on handcrafted signal processing pipelines, combining filtering with heuristic peak detection. Although such methods perform well in laboratory settings, their accuracy deteriorates sharply under motion artifacts and variable wearing conditions that dominate real-world recordings. In contrast, our results demonstrate that combining state-space with learned representations addresses these challenges, resulting in more reliable estimation and improved HRV indices across diverse activities, and improving over pure deep learning approaches that do not explicitly account for temporal dynamics.

b) Clinical Relevance: Robust HRV estimation in free-living conditions is important [2], [47], as it enables continuous monitoring of autonomic function without the constraints of ECG [1]. Such capability has potential applications in cardiovascular disease stratification, stress and mental health monitoring outside the clinic. By demonstrating improved robustness under motion and environmental variability, our method advances the feasibility of population-scale HRV monitoring using widely available modalities in wearables.

c) Limitations and Future Work: Despite these strengths, several limitations should be acknowledged for our work. First, a distinct challenge in wearable sensing is the physiological difference between electrical HRV and mechanical Pulse Rate Variability (PRV). While our model targets ECG-derived labels, the discrepancy between HRV and PRV in subjects is typically low < 20 ms [47], [48], which is an order of magnitude smaller than the error introduced by motion artifacts ($>$100 ms). Moreover, inferring *approximate* HRV from 5 s averaged HR has been explored [49], but it discards beat-level dynamics and limits artifact-aware modeling. Thus, correcting PRV towards HRV labels acts as a beneficial calibration rather than an artificial hallucination. PRV-HRV divergence can increase with vasomotor/thermoregulatory changes [50], posture shifts, or cardiovascular pathology [51]; dedicated evaluation in such regimes remains important.

Second, our treatment of PAT is simplified, focusing on its changes as an additional latent component. More sophisticated modeling of dynamics, and their physiological coupling with both RR intervals and vascular properties, could further improve accuracy in unconstrained environments.

Third, our estimates are defined at a 2 s update rate (overlapping windows), rather than as a beat-indexed sequence obtained from explicit peaks. This improves robustness under motion but introduces averaging within each update interval. The impact is small: on WildPPG, ECG-derived IBIs aggregated to 2 s bins show only $\sim$8 ms variability, and HRV computed from averaged versus beat-level ECG differ by only $\sim$3 ms, which is negligible relative to motion-induced errors.

Finally, although our approach benefits from large-scale supervised training using WildPPG and DaLiA, we have not explored self-supervised learning. Given the abundance of unlabeled data, future work could employ representation learning to pretrain models on large datasets for generalization.

Overall, our findings establish a strong foundation for robust HRV estimation from wearables in real-world settings while demonstrating improvements over traditional methods.

VII. Conclusion

We have introduced a multimodal, learning-based state-space framework for continuous HRV estimation from wearable PPG with robust estimations under everyday motion and activities as well as under environmental noise. Across two public datasets collected in free-living conditions, the proposed model and uncertainty-aware quality gating consistently outperformed traditional peak-detection pipelines and feed-forward baselines, reducing SDNN error by up to 80%. These results show that reliable HRV monitoring from consumer wearables is feasible in uncontrolled settings. By providing an end-to-end pipeline with reproducible baselines and evaluations, this work supports more trustworthy and comparable HRV research for real-world health monitoring.

REFERENCES

[1] E. B. Schroeder and et al., "Hypertension, blood pressure, and heart rate variability," *Hypertension*, 2003.

[2] M. Malik and et al., "Crosstalk proposal: Heart rate variability is a valid measure of cardiac autonomic responsiveness," *The Journal of Physiology*, 2019.

[3] H. Lee and et al., "Real-time machine learning model to predict in-hospital cardiac arrest using heart rate variability in ICU," *npj Digital Medicine*, 2023.

[4] F. Marzbanrad and et al., "Methodological comparisons of heart rate variability analysis in patients with type 2 diabetes and angiotensin converting enzyme polymorphism," *IEEE Journal of Biomedical and Health Informatics*, 2016.

[5] Z. Wang and et al., "Heart rate variability in mental disorders: an umbrella review of meta-analyses," *Translational Psychiatry*, 2025.

[6] C. da Estrela and et al., "Heart rate variability, sleep quality, and depression in the context of chronic stress," *Annals of Behavioral Medicine*, 2020.

[7] Y. Sun and N. Thakor, "Photoplethysmography revisited: From contact to noncontact, from point to imaging," *IEEE Transactions on Biomedical Engineering*, 2016.

[8] J. Park and et al., "Photoplethysmogram analysis and applications: An integrative review," *Frontiers in Physiology*, 2022.

[9] C. Holz and E. J. Wang, "Glabella: Continuously sensing blood pressure behavior using an unobtrusive wearable device," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2017.

[10] B. Braun and et al., "egoppg: Heart rate estimation from eye-tracking cameras in egocentric systems to benefit downstream vision tasks," *arXiv preprint arXiv:2502.20879*, 2025.

[11] R.-J. Shei and et al., "Wearable activity trackers–advanced technology or advanced marketing?" *European Journal of Applied Physiology*, 2022.

[12] V. Bieri and et al., "Beliefppg: Uncertainty-aware heart rate estimation from ppg signals via belief propagation," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023.

[13] Z. Zhang, Z. Pi, and B. Liu, "Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise," *IEEE Transactions on Biomedical Engineering*, 2015.

[14] A. Reiss and et al., "Deep ppg: Large-scale heart rate estimation with convolutional neural networks," *Sensors*, 2019.

[15] B. U. Demirel and C. Holz, "Temporal cardiovascular dynamics for improved ppg-based heart rate estimation," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[16] D. C. Sheridan and et al., "Heart rate variability analysis: How much artifact can we remove?" *Psychiatry Investigation*, 2020.

[17] M. Meier and et al., "WildPPG: A real-world PPG dataset of long continuous recordings," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[18] B. Bent and et al, "Investigating sources of inaccuracy in wearable optical heart rate sensors," *npj Digital Medicine*, 2020.

[19] A. Sahroni and et al., "Hrv assessment using finger-tip photoplethysmography (pulserate) as compared to ecg on healthy subjects during different postures and fixed breathing pattern," *Procedia Computer Science*, 2019.

[20] J. Paalasmaa and et al., "Adaptive heartbeat modeling for beat-to-beat heart rate measurement in ballistocardiograms," *IEEE Journal of Biomedical and Health Informatics*, 2015.

[21] T. Wittenberg and et al., "Evaluation of hrv estimation algorithms from ppg data using neural networks," *Current Directions in Biomedical Engineering*, 2020.

[22] A. Gudi and et al., "Efficient real-time camera based estimation of heart rate and its variability," in *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019.

[23] L. Liu and et al., "Camera-based seismocardiogram for heart rate variability monitoring," *IEEE JBHI*, 2024.

[24] L. Iozzia, L. Cerina, and L. Mainardi, "Relationships between heart-rate variability and pulse-rate variability obtained from video-ppg signal using zca," *Physiological Measurement*, 2016.

[25] A. Aygun and et al., "Robust interbeat interval and heart rate variability estimation method from various morphological features using wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[26] A. Gupta and et al., "Heart rate and hrv estimation using ppg based on superlet transform and lstm network," *IEEE Transactions on Instrumentation and Measurement*, pp. 1–1, 2025.

[27] C. León and et al., "Early detection of late onset sepsis in premature infants using visibility graph analysis of heart rate variability," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[28] B.-J. Singstad and et al.., "Estimation of heart rate variability from finger photoplethysmography during rest, mild exercise and mild mental stress," *Journal of Electrical Bioimpedance*, 2021.

[29] K. Tyapochkin and et al., "Smartphone ppg: signal processing, quality assessment, and impact on hrv parameters," in *IEEE EMBC*, 2019.

[30] R. Ahmed and et al., "A deep learning and fast wavelet transform-based hybrid approach for denoising of ppg signals," *IEEE Sensors Letters*, 2023.

[31] M. Meier and et al., "Tri-spectral ppg: Robust reflective photoplethysmography by fusing multiple wavelengths for cardiac monitoring," in *IEEE International Conference on Body Sensor Networks (BSN)*, 2024.

[32] Y. Zhang and et al., "Efficient and direct inference of heart rate variability using both signal processing and machine learning," in *Proceedings of the 8th ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, 2024.

[33] J. Xu and et al., "Real-time intelligent on-device monitoring of heart rate variability with ppg sensors," *Journal of Systems Architecture*, 2024.

[34] Lee and Ju-Won, "Design of Kalman Filter to Estimate Heart Rate Variability from PPG Signal for Mobile Healthcare," *journal of information and communication convergence engineering*, 2010.

[35] S. Ismail and et al., "Heart rate tracking in photoplethysmography signals affected by motion artifacts: a review," *EURASIP Journal on Advances in Signal Processing*, 2021.

[36] L. S. Goldoozian and et al., "Time-varying assessment of heart rate variability parameters using respiratory information," *Computers in Biology and Medicine*, 2017.

[37] B. E. Ajtay and et al., "The oscillating pulse arrival time as a physiological explanation regarding the difference between ecg- and photoplethysmogram-derived heart rate variability parameters," *Biomedical Signal Processing and Control*, 2023.

[38] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Transactions on Biomedical Engineering*, 1985.

[39] O. Ronneberger and et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

[40] B. U. Demirel and C. Holz, "An unsupervised approach for periodic source detection in time series," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24. JMLR.org, 2024.

[41] M. Nitzan and et al., "The difference in pulse transit time to the toe and finger measured by photoplethysmography," *Physiological Measurement*, 2001.

[42] T. F. of the European Society of Cardiology the North American Society of Pacing Electrophysiology, "Heart rate variability," *Circulation*, 1996.

[43] M. A. F. Pimentel and et al., "Toward a robust estimation of respiratory rate from pulse oximeters," *IEEE Transactions on Biomedical Engineering*, 2017.

[44] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, 1967.

[45] M. Elgendi and et al., "Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions," *PLOS ONE*, 2013.

[46] D. Makowski and et al., "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, 2021.

[47] E. Mejía-Mejía and et al., "Differential effects of the blood pressure state on pulse rate variability and heart rate variability in critically ill patients," *npj Digital Medicine*, 2021.

[48] E. Gil and et al., "Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions," *Physiological Measurement*, 2010.

[49] X. Li and et al., "Calculation of approximate heart rate variability indicators based on low-resolution heart rate data provided by widely used commercially available wearable devices," *Biomedical Signal Processing and Control*, 2026.

[50] E. Mejía-Mejía and et al., "Heart Rate Variability (HRV) and Pulse Rate Variability (PRV) for the Assessment of Autonomic Responses," *Frontiers in Physiology*, 2020.

[51] A. B. Kantrowitz and et al., "Pulse rate variability is not the same as heart rate variability: findings from a large, diverse clinical population study," *Frontiers in Physiology*, 2025.