PHASE-Net: Physics-Grounded Harmonic Attention System for Efficient Remote Photoplethysmography Measurement

Supplementary Material

A. Introduction to the Datasets

UBFC-rPPG [1] contains 42 RGB facial videos from 42 distinct subjects. Each video is captured at 640×480 pixel resolution and 30 frames per second (fps). Recordings take place under varied lighting conditions, including natural sunlight and indoor artificial illumination. Ground-truth physiological signals are recorded via a CMS50E pulse oximeter at 60 Hz, ensuring precise temporal alignment for evaluation.

PURE [24] comprises 60 high-quality RGB videos collected from 10 subjects performing six different head movement scenarios (static, talking, translation movements, etc.). Videos are recorded at 30 fps under consistent indoor lighting and controlled background settings, minimizing external interference. Synchronized physiological measurements are obtained using a CMS50E oximeter sampling at 60 Hz. PURE is particularly valuable for evaluating rPPG performance during facial movements.

BUAA [32] is designed to assess algorithmic robustness across varying illumination intensities. The dataset features video sequences recorded under a range of controlled lighting conditions, from low-light (below 10 lux) to normal brightness. In our experiments, we only utilize videos captured under illumination levels $\geq$10 lux, as extremely dim lighting introduces significant image degradation requiring specialized enhancement techniques beyond this study's scope.

MMPD [26] comprises 660 videos, each lasting one minute, collected from 33 subjects with diverse skin tones and gender distributions. Each video is recorded at 30 fps with a resolution of 320×240 pixels, under four distinct lighting conditions (bright, warm, dim, and colored lighting). Subjects perform various daily activities, introducing intra-subject variability and further increasing dataset complexity.

B. Implementation Details

Our PHASE-Net is implemented using PyTorch. The input to the network is a sequence of 128 frames, resized to 128×128. We trained the model for 15 epochs using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 4. The loss function hyperparameter was set to $\lambda = 0.1$. All experiments were conducted on a single NVIDIA H100 GPU.

C. Detailed Derivation of the Physics-Informed Temporal Model

This appendix provides the detailed mathematical derivations for the physics-informed temporal model, as summarized in Section 3.1.

C.1. Derivation of the Damped Wave Equation (PDE)

Our goal is to derive a single equation for the pressure pulsation p' from the 1D linearized equations for momentum and continuity:

$$\rho \frac{\partial u'}{\partial t} + k u' = -\frac{\partial p'}{\partial x} \tag{18}$$

$$\frac{\partial Q'}{\partial x} = -C \frac{\partial p'}{\partial t} \tag{19}$$

where $Q' = Au'$ is the flow rate, and A is the cross-sectional area of the vessel. The derivation proceeds in the following steps:

1. We take the partial derivative of the momentum equation (Eq. 18) with respect to the spatial variable x:

   $$\frac{\partial}{\partial x}\left(\rho \frac{\partial u'}{\partial t} + k u'\right) = \frac{\partial}{\partial x}\left(-\frac{\partial p'}{\partial x}\right)$$

   Assuming fluid properties ρ, k are locally uniform and swapping the order of differentiation, we get:

   $$\rho \frac{\partial}{\partial t}\left(\frac{\partial u'}{\partial x}\right) + k\left(\frac{\partial u'}{\partial x}\right) = -\frac{\partial^2 p'}{\partial x^2} \tag{20}$$

2. We relate the velocity gradient $\frac{\partial u'}{\partial x}$ to the flow rate gradient $\frac{\partial Q'}{\partial x}$. Since $Q' = Au'$, under the small pulsation assumption, the area A can be approximated by its mean value $\bar{A}$, so $Q' \approx \bar{A} u'$. Taking the spatial derivative yields:

   $$\frac{\partial u'}{\partial x} \approx \frac{1}{\bar{A}} \frac{\partial Q'}{\partial x} \tag{21}$$

3. We substitute Eq. 21 into Eq. 20 to replace the velocity gradient with the flow rate gradient:

   $$\rho \frac{\partial}{\partial t}\left(\frac{1}{\bar{A}} \frac{\partial Q'}{\partial x}\right) + \frac{k}{\bar{A}}\left(\frac{\partial Q'}{\partial x}\right) = -\frac{\partial^2 p'}{\partial x^2}$$

4. Finally, we use the continuity equation (Eq. 19) to replace the flow rate gradient term $\frac{\partial Q'}{\partial x}$ with the pressure term $-C\frac{\partial p'}{\partial t}$:

   $$\frac{\rho}{\bar{A}} \frac{\partial}{\partial t}\left(-C \frac{\partial p'}{\partial t}\right) + \frac{k}{\bar{A}}\left(-C \frac{\partial p'}{\partial t}\right) = -\frac{\partial^2 p'}{\partial x^2}$$

Rearranging the terms, we obtain:

$$\frac{\rho C}{\bar{A}}\frac{\partial^2 p'}{\partial t^2} + \frac{kC}{\bar{A}}\frac{\partial p'}{\partial t} = \frac{\partial^2 p'}{\partial x^2}$$

5. By defining new physical constants for wave speed squared ($c^2 := \frac{\bar{A}}{\rho C}$) and a damping-related coefficient, we arrive at the final Damped Wave Equation presented in the main text:

$$\frac{\partial^2 p'}{\partial t^2} + \alpha\frac{\partial p'}{\partial t} = c^2\frac{\partial^2 p'}{\partial x^2} \tag{22}$$

C.2. Discretization and State-Space Formulation

We start with the second-order ODE for the damped harmonic oscillator:

$$\frac{d^2 z(t)}{dt^2} + \alpha\frac{dz(t)}{dt} + \omega^2 z(t) = u(t) \tag{23}$$

First, we convert this into a system of two first-order ODEs by defining the state vector $\mathbf{x}(t) = [z(t), v(t)]^T$, where $v(t) = \frac{dz(t)}{dt}$ is the velocity.

$$\frac{dz(t)}{dt} = v(t)$$
$$\frac{dv(t)}{dt} = -\alpha v(t) - \omega^2 z(t) + u(t)$$

We discretize this system using a semi-implicit Euler method with a time step Δt. Let $z_t \approx z(t\Delta t)$ and $a_t \approx u(t\Delta t)$. The update rules are:

$$v_t = v_{t-1} + \Delta t \cdot (-\alpha v_t - \omega^2 z_{t-1} + a_t) \tag{24}$$
$$z_t = z_{t-1} + \Delta t \cdot v_t \tag{25}$$

We first solve for v_t from Eq. 24:

$$(1+\alpha\Delta t)v_t = v_{t-1} - \omega^2\Delta t z_{t-1} + \Delta t a_t$$
$$v_t = \frac{1}{1+\alpha\Delta t}v_{t-1} - \frac{\omega^2\Delta t}{1+\alpha\Delta t}z_{t-1} + \frac{\Delta t}{1+\alpha\Delta t}a_t$$

Substituting this into Eq. 25 gives the update for z_t:

$$z_t = z_{t-1} + \Delta t\left(\frac{1}{1+\alpha\Delta t}v_{t-1} - \frac{\omega^2\Delta t}{1+\alpha\Delta t}z_{t-1} + \frac{\Delta t}{1+\alpha\Delta t}a_t\right)$$
$$z_t = \left(1 - \frac{\omega^2\Delta t^2}{1+\alpha\Delta t}\right)z_{t-1} + \frac{\Delta t}{1+\alpha\Delta t}v_{t-1} + \frac{\Delta t^2}{1+\alpha\Delta t}a_t$$

We can now write these two update rules in the standard LTI State-Space Model form $\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}a_t$, where $\mathbf{x}_t = [z_t, v_t]^T$:

$$\mathbf{x}_t = \underbrace{\begin{bmatrix} 1 - \frac{\omega^2\Delta t^2}{1+\alpha\Delta t} & \frac{\Delta t}{1+\alpha\Delta t} \\ -\frac{\omega^2\Delta t}{1+\alpha\Delta t} & \frac{1}{1+\alpha\Delta t} \end{bmatrix}}_{\mathbf{A}}\mathbf{x}_{t-1} + \underbrace{\begin{bmatrix} \frac{\Delta t^2}{1+\alpha\Delta t} \\ \frac{\Delta t}{1+\alpha\Delta t} \end{bmatrix}}_{\mathbf{B}}a_t \tag{26}$$

The output equation is simply $z_t = \mathbf{C}\mathbf{x}_t$, with $\mathbf{C} = [1 \quad 0]$.

C.3. Proofs of Propositions

Proposition 5 (Equivalence to Causal Convolution)**.** *The solution z_t of the LTI system $\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}a_t, z_t = \mathbf{C}\mathbf{x}_t$ can be expressed as a causal convolution of all past inputs.*

Proof. By unrolling the state-space recurrence relation, we get:

$$\begin{aligned}\mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}a_t \\ &= \mathbf{A}(\mathbf{A}\mathbf{x}_{t-2} + \mathbf{B}a_{t-1}) + \mathbf{B}a_t \\ &= \mathbf{A}^2\mathbf{x}_{t-2} + \mathbf{A}\mathbf{B}a_{t-1} + \mathbf{B}a_t \\ &= \dots \\ &= \mathbf{A}^t\mathbf{x}_0 + \sum_{m=0}^{t-1}\mathbf{A}^m\mathbf{B}a_{t-m}\end{aligned}$$

Assuming zero initial conditions ($\mathbf{x}_0 = \mathbf{0}$), the state is solely determined by the history of inputs:

$$\mathbf{x}_t = \sum_{m=0}^{t-1}\mathbf{A}^m\mathbf{B}a_{t-m}$$

Applying the output equation $z_t = \mathbf{C}\mathbf{x}_t$:

$$z_t = \mathbf{C}\sum_{m=0}^{t-1}\mathbf{A}^m\mathbf{B}a_{t-m} = \sum_{m=0}^{t-1}(\mathbf{C}\mathbf{A}^m\mathbf{B})a_{t-m}$$

We can extend the sum to infinity by defining the kernel $g[m] = \mathbf{C}\mathbf{A}^m\mathbf{B}$ for $m \geq 0$ and assuming a causal system where $a_k = 0$ for $k < 0$. This gives the convolution form:

$$z_t = \sum_{m=0}^{\infty} g[m]a_{t-m}$$

For a damped system, the spectral radius $\rho(\mathbf{A}) < 1$, ensuring the IIR filter is stable. □

Proposition 6 (FIR Approximation)**.** *The IIR convolution can be approximated with arbitrary precision ε by a Finite Impulse Response (FIR) filter of sufficient length R.*

Proof. The error introduced by truncating the infinite sum (the IIR filter kernel $g[m]$) at length $R-1$ is the tail of the sum:

$$e_t = \left|\sum_{m=0}^{\infty} g[m]a_{t-m} - \sum_{m=0}^{R-1} g[m]a_{t-m}\right| = \left|\sum_{m=R}^{\infty} g[m]a_{t-m}\right|$$

Let the input be bounded, $\|a_t\|_\infty \leq M_{in}$, and the matrix norms be bounded such that $\|\mathbf{A}^m\| \leq K\rho^m$ for some constants $K > 0$ and $0 < \rho < 1$ (guaranteed for a stable

system). We can bound the error:

$$\begin{aligned}\|e_t\|_\infty &\le \sum_{m=R}^{\infty} \|\mathbf{C}\|\|\mathbf{A}^m\|\|\mathbf{B}\|\|a_{t-m}\|_\infty \\ &\le \sum_{m=R}^{\infty} \|\mathbf{C}\|(K\rho^m)\|\mathbf{B}\|M_{in} \\ &= KM_{in}\|\mathbf{C}\|\|\mathbf{B}\| \sum_{m=R}^{\infty} \rho^m\end{aligned}$$

The last term is a geometric series, which sums to $\frac{\rho^R}{1-\rho}$. Therefore:

$$\|e_t\|_\infty \le KM_{in}\|\mathbf{C}\|\|\mathbf{B}\|\frac{\rho^R}{1-\rho}$$

To ensure the error is less than a desired precision ε, we require:

$$KM_{in}\|\mathbf{C}\|\|\mathbf{B}\|\frac{\rho^R}{1-\rho} \le \varepsilon$$

Solving for R gives the required receptive field length (filter size):

$$R \ge \frac{\log\left(\frac{KM_{in}\|\mathbf{C}\|\|\mathbf{B}\|}{\varepsilon(1-\rho)}\right)}{\log(1/\rho)}$$

This shows that a finite kernel length R is sufficient to approximate the true physical dynamics to any desired precision. □

D. Generalization Theory of PHASE-Net

Problem Setup. Consider the stable linear time–invariant (LTI) system derived from the physics model:

$$\begin{aligned}\mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}a_t, \\ z_t &= \mathbf{C}\mathbf{x}_t = \sum_{m=0}^{\infty} g[m]\, a_{t-m}, \\ g[m] &= \mathbf{C}\mathbf{A}^m\mathbf{B}.\end{aligned}$$

In the network implementation we use a finite–length causal convolution. Let the temporal window length be R, define the input vector

$$\phi_t = (a_t, a_{t-1}, \ldots, a_{t-R+1}) \in \mathbb{R}^R,$$

and the truncated FIR coefficient vector

$$w = (g[0], g[1], \ldots, g[R-1]).$$

The predictor can be written as

$$f(\phi_t) = \langle w, \phi_t \rangle.$$

Physical Facts. Fact 1 (Stability). Causality and spectral normalization guarantee $\rho(\mathbf{A}) < 1$. Hence there exist constants $K > 0$ and $0 < \rho < 1$ such that

$$\|\mathbf{A}^m\| \le K\rho^m, \quad \forall m \ge 0.$$

Fact 2 (Magnitude and Norm Bounds). The input amplitude is bounded by M_{in}. Weight regularization ensures $\|\mathbf{B}\| \le B_0$ and $\|\mathbf{C}\| \le C_0$. Therefore the ℓ_1 norm of the convolution kernel satisfies

$$\|w\|_1 = \sum_{m=0}^{R-1} |g[m]| \le \sum_{m=0}^{\infty} C_0KB_0\rho^m = \frac{U}{1-\rho}, \qquad U \triangleq C_0KB_0.$$

Fact 3 (FIR Truncation Error). Because $|g[m]| \le U\rho^m$,

$$\sum_{m=R}^{\infty} |g[m]| \le \frac{U\rho^R}{1-\rho}.$$

Since $\|a_t\|_\infty \le M_{\text{in}}$, the difference between the infinite IIR output and the length–R FIR output satisfies

$$|z_t - z_t^{(R)}| \le \frac{U}{1-\rho}M_{\text{in}}\rho^R \triangleq \Gamma\rho^R.$$

This term can be made arbitrarily small by increasing R.

Rademacher Complexity. Consider samples $\{\phi_i\}_{i=1}^n$ with $\|\phi_i\|_\infty \le M_{\text{in}}$. The empirical Rademacher complexity is

$$\widehat{\mathfrak{R}}_n = \mathbb{E}_\sigma\Big[\sup_{\|w\|_1 \le L} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\langle w, \phi_i\rangle\Big],$$

where σ_i are independent Rademacher variables and $L = U/(1-\rho)$.

Step 1 (Dual Norm Representation). By ℓ_1–ℓ_∞ duality,

$$\widehat{\mathfrak{R}}_n = \frac{L}{n}\,\mathbb{E}_\sigma\Big\|\sum_{i=1}^{n}\sigma_i\phi_i\Big\|_\infty.$$

Step 2 (Bounding the Maximal Coordinate). For any coordinate $j \le R$, the random variable $\sum_{i=1}^n \sigma_i\phi_{i,j}$ has magnitude at most nM_{in}. Khintchine–Kahane inequality together with a union bound yields

$$\mathbb{E}_\sigma \max_{1\le j\le R}\Big|\sum_{i=1}^{n}\sigma_i\phi_{i,j}\Big| \le M_{\text{in}}\sqrt{2n\log(2R)}.$$

Step 3 (Complexity Bound). Substituting the above into the dual form gives

$$\widehat{\mathfrak{R}}_n \le LM_{\text{in}}\sqrt{\frac{2\log(2R)}{n}}.$$

Taking expectation shows that the true Rademacher complexity satisfies

$$\mathfrak{R}_n \le \frac{U}{1-\rho}M_{\text{in}}\sqrt{\frac{2\log(2R)}{n}}.$$

Source-Domain Generalization. Let the loss ℓ be L_ℓ–Lipschitz and bounded in $[0,1]$. By the standard Rademacher generalization inequality, with probability at least $1-\delta$ over the random draw of the training set,

$$\mathcal{E}_{\mathrm{src}}(f) \le \widehat{\mathcal{E}}_n(f) + 2L_\ell \mathfrak{R}_n + 3\sqrt{\frac{\log(2/\delta)}{2n}} + O(\rho^R).$$

Plugging in the bound on $\mathfrak{R}_n$ gives

$$\mathcal{E}_{\mathrm{src}}(f) \le \widehat{\mathcal{E}}_n(f) + O\left(\sqrt{\tfrac{\log R}{n}}\right) + O(\rho^R).$$

Target-Domain Risk. Let $\mathbb{P}_{\mathrm{src}}$ and $\mathbb{P}_{\mathrm{tgt}}$ denote the source and target distributions, and W_1 their 1–Wasserstein distance. Since f is L_f–Lipschitz with

$$L_f \le \|w\|_1 \le \frac{U}{1-\rho},$$

the discrepancy between source and target satisfies

$$\mathrm{Disc} \le L_\ell L_f \mathrm{W}_1(\mathbb{P}_{\mathrm{src}}, \mathbb{P}_{\mathrm{tgt}}) \le L_\ell \frac{U}{1-\rho} \mathrm{W}_1(\mathbb{P}_{\mathrm{src}}, \mathbb{P}_{\mathrm{tgt}}).$$

By the triangle inequality,

$$\mathcal{E}_{\mathrm{tgt}}(f) \le \mathcal{E}_{\mathrm{src}}(f) + \mathrm{Disc}.$$

Combining with the source bound yields

$$\mathcal{E}_{\mathrm{tgt}}(f) \le \widehat{\mathcal{E}}_n(f) + O\left(\sqrt{\tfrac{\log R}{n}}\right) + O(\rho^R) + L_\ell \frac{U}{1-\rho} \mathrm{W}_1(\mathbb{P}_{\mathrm{src}}, \mathbb{P}_{\mathrm{tgt}}).$$

Choice of R. To make the truncation error $O(\rho^R)$ smaller than the statistical term, choose

$$R \gtrsim \frac{2\log n}{\log(1/\rho)} = \Theta(\log n).$$

With this choice, ρ^R is negligible and the bound simplifies to

$$\mathcal{E}_{\mathrm{tgt}}(f) \le \widehat{\mathcal{E}}_n(f) + O\left(\sqrt{\tfrac{\log\log n}{n}}\right) + L_\ell \frac{U}{1-\rho} \mathrm{W}_1(\mathbb{P}_{\mathrm{src}}, \mathbb{P}_{\mathrm{tgt}}).$$

Comparison with Unconstrained Models. For an unconstrained temporal model with hypothesis class $\mathcal{F}_{\mathrm{base}}$, one typically has

$$\mathfrak{R}_n(\mathcal{F}_{\mathrm{base}}) = O\left(\sqrt{\tfrac{C}{n}}\right),$$

where the capacity constant C depends on depth, width, or spectral norm and is usually much larger than $\log\log n$. Thus the physics–informed class enjoys a strictly smaller statistical term $O(\sqrt{\log\log n/n})$ under the same sample size n.

E. Detailed Description of ZAS

The Zero-FLOPs Axial Swapper (ZAS) is a lightweight spatial mixing operator designed to enrich long-range dependencies without adding computational burden. By selectively permuting a small subset of feature channels through block-wise transposition, ZAS introduces cross-region interactions that enhance the receptive field while keeping the temporal dimension untouched. Because the operation is purely an index reordering, it adds no learnable parameters and incurs zero FLOPs.

Algorithm 1: Zero-FLOPs Axial Swapper (ZAS)

Feature tensor $X \in \mathbb{R}^{B\times C\times T\times H\times W}$
Output tensor $\tilde{X} \in \mathbb{R}^{B\times C\times T\times H\times W}$

Step 1. Channel partition.
Split X into two disjoint parts:

$$X = \big[X_{\mathrm{id}},\, X_{\mathrm{swap}}\big],$$

where X_{id} contains the first $C-k$ channels and X_{swap} contains the last $k = \lfloor pC \rfloor$ channels to be permuted.

Step 2. Block partition.
Given a block size b, crop the core region $H_2 = \lfloor H/b \rfloor \cdot b$, $W_2 = \lfloor W/b \rfloor \cdot b$, and reshape each spatial slice of X_{swap}

$$\mathcal{P} : \mathbb{R}^{H_2\times W_2} \to \mathbb{R}^{\frac{H_2}{b}\times\frac{W_2}{b}\times b\times b}$$

into a grid of non-overlapping $b\times b$ blocks.

Step 3. Block-wise transpose.
For each $b\times b$ block Z, apply the inner transpose

$$\mathcal{T}(Z)_{u,v} = Z_{v,u}.$$

This operation is performed independently for every block and for all batches, channels, and time frames.

Step 4. Reconstruction.
Recover the spatial layout by the inverse partition

$$\mathrm{ZAS}(X_{\mathrm{swap}}) = \mathcal{P}^{-1}\big(\mathcal{T}(\mathcal{P}(X_{\mathrm{swap}}))\big).$$

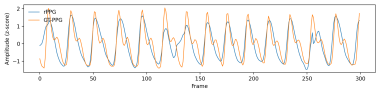
Concatenate with the unchanged channels to obtain the output:

$$\tilde{X} = \big[X_{\mathrm{id}},\, \mathrm{ZAS}(X_{\mathrm{swap}})\big].$$

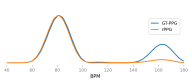
Remark.
ZAS performs only index reordering and introduces *zero learnable parameters* and *zero FLOPs*; its Jacobian is a permutation matrix, ensuring gradient safety and perfect energy preservation.

F. Visualization of the Predicted and Ground-truth BVP

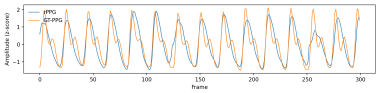
We randomly select representative clip samples from the UBFC-rPPG [1] and PURE [24] datasets and visualize both the predicted rPPG waveforms and their corresponding power spectral density (PSD) curves in Fig. 5 and Fig. 6. These qualitative results provide an intuitive view of model behavior: the predicted signals not only closely follow the ground-truth BVP in amplitude and phase but also exhibit highly consistent dominant frequency peaks in the PSD domain, indicating accurate heart-rate estimation. Across both controlled (PURE) and more unconstrained (UBFC) scenarios, PHASE-Net preserves the fine-grained temporal structure of the pulse waveform and maintains sharp, well-aligned spectral peaks, further validating its ability to recover clean physiological rhythms despite variations in illumination, motion, and sensor noise.



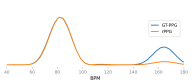
(a) rPPG signal (segment 1).



(b) PSD (segment 1).

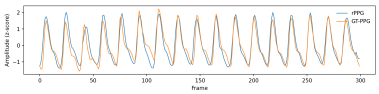


(c) rPPG signal (segment 2).

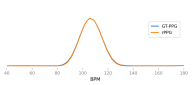


(d) PSD (segment 2).

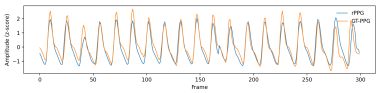
Figure 5. Visual comparison of rPPG signals predicted by PHASE-Net and their corresponding power spectral densities (PSDs), along with ground-truth references, on the PURE dataset [24].



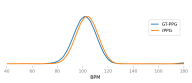
(a) rPPG signal (subject 22).



(b) PSD (subject 22).



(c) rPPG signal (subject 9).



(d) PSD (subject 9).

Figure 6. Visual comparison of rPPG signals predicted by PHASE-Net and their corresponding PSDs, with ground-truth references, on the UBFC-rPPG dataset [1].