FacePhys: State of the Heart Learning

Jiankai Tang* Tsinghua University
Kegang Wang* Tsinghua University
Yuntao Wang† Tsinghua University
Xin Liu University of Washington

Yuxuan Fan Tsinghua University
Jiatong Ji Tsinghua University
Yuanchun Shi Tsinghua University
Daniel McDuff† University of Washington

Abstract

Vital sign measurement using cameras presents opportunities for comfortable, ubiquitous health monitoring. Remote photoplethysmography (rPPG), a foundational technology, enables cardiac measurement through minute changes in light reflected from the skin. However, practical deployment is limited by the computational constraints of performing analysis on front-end devices and the accuracy degradation of transmitting data through compressive channels that reduce signal quality. We propose a memory efficient rPPG algorithm - *FacePhys* - built on a Linear Time-Invariant (LTI) state space model with temporal-spatial duality, which resolves the trilemma of model scalability, cross-dataset generalization, and real-time operation. Leveraging a transferable heart state, FacePhys captures subtle periodic variations across video frames through streaming per-frame waveform inference, enabling training on extended video sequences and supporting low-latency operation. FacePhys establishes a new state-of-the-art, with a substantial 49% reduction in heart rate estimation error. Our solution enables streaming waveform inference with a memory footprint of 3.6 MB and per-frame latency of 9.46 ms – surpassing existing methods by 83% to 99%. A real-time demo is available at facephys.com.

1 Introduction

Remote photoplethysmography (rPPG) is a non-contact technique for measuring cardiac activity by analyzing light reflected from skin (McDuff, 2023). The subtle pulsatile variations are typically not visible to the naked eye, but they can be captured by cameras and analyzed to recover heart rate (Wu et al., 2012), heart rate variability and other vital signs that can be derived from the waveform (e.g., blood pressure (Curran et al., 2023)). The attraction of non-invasive measurement of cardiac activity has led to rPPG technology being integrated into a number of applications, including health monitors (Lu et al., 2021; Liu et al., 2024), emotion recognition systems (Wang et al., 2024b; Song et al., 2020), and fatigue detection models (Du et al., 2022).

Deep learning is the predominant and superior approach used in rPPG models (McDuff, 2023), which can be effective at disentangling the cardiac pulse from head motions, ambient lighting changes, and other confounders (Tang et al., 2023; Nowara et al., 2020). However, there are several reasons that performing inference *on-device* is very desirable, and this means high computational efficiency *and* accuracy are both important.

First, the PPG is degraded more than other information by video compression (McDuff et al., 2017) meaning that it is optimal to process the video as close to the source as possible (i.e., on-device) rather than send it over a bandwidth limited network. **Second**, application of rPPG in resource limited settings, such as by mobile health workers in remote settings, means it may not be possible to rely on a cloud service. **Third**, cardiac measurement is a highly privacy sensitive application, with videos often

*Equal contribution. †Corresponding authors.

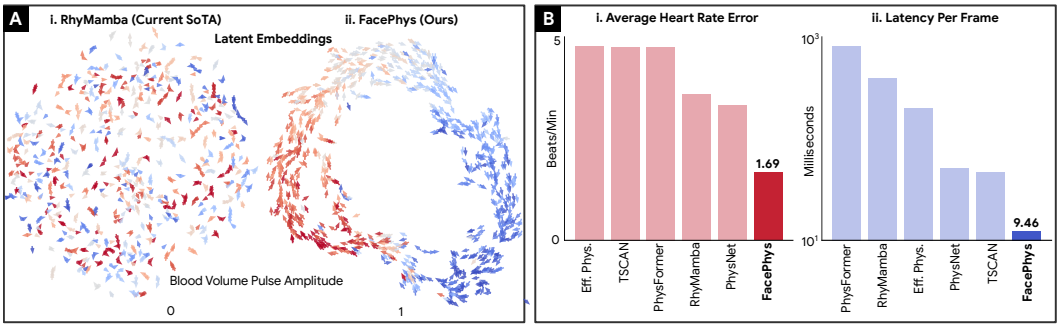


Figure 1: **The FacePhys State Space Model**, provides an effective representation of the cyclical nature of heart beats, as shown by the latent embeddings (A), combining high accuracy (B-i) and high efficiency (B-ii). We achieve gains of 49% in heart rate estimation and 83% in per-frame latency compared to the current state-of-the-art.

containing a patient's face and sensitive physiological signals. Further adding to why streaming and uploading to a server is not ideal. **Fourth**, the ability to run at a high frame rates enables opportunistic sensing (e.g., obtaining measurements each time you look at your phone) and helps capture waveform dynamics that could be used to detect atrial fibrillation (Liu et al., 2022) or hypertension (Yoshioka and Bounyong, 2020) where high-frame rates (at least 100Hz) are a requirement to yield precise measurements. However, rather than leading to a decrease, the use of deep learning has meant that the computational resources required to run rPPG algorithms has *increased*, making practical applications and real-time inference challenging.

Some studies attempt to address efficiency issues by reducing the resolution of frames (Botina-Monsalve et al., 2022) or simplifying the model complexity (Liu et al., 2021); however, these design choices invariably impact accuracy. Creating smaller, more memory efficient models (He et al., 2019; Lin et al., 2019) often requires slicing videos into shorter segments (e.g., 180 frames, approximately 6 seconds) (Liu et al., 2023; Kuang et al., 2023), an approach that overlooks the importance of longer-term temporal dependencies, leading to the loss of important information during training and inference, and making algorithms more susceptible to noise.

This paper proposes an accurate and memory-efficient rPPG model (**FacePhys**) built on a Linear Time-Invariant (LTI) state space model. FacePhys captures subtle spatial and temporal color changes corresponding to blood volume pulse, enabling accurate prediction of model states and the cardiac signal across arbitrarily long sequences. Leveraging temporal-spatial state space attention duality (TSD), FacePhys supports streaming per-frame waveform inference through recurrent state updates, significantly reducing computational load and achieving real-time operation. Heart rate is computed from the predicted waveform using standard spectral analysis. By learning a strong internal representation, FacePhys demonstrates superior performance across all datasets compared to state-of-the-art (SOTA) methods (see Fig. 1). The three core contributions of this paper are as follows:

1. We built FacePhys, an LTI state space model with TSD that supports streaming *per-frame waveform* inference while preserving long-range periodic information.
2. We developed a discrete-time formulation that cuts the model memory to **3.6 MB** (up to **98.4%** reduction) and achieves sub-**10 ms** per-frame latency, verified on mobile browsers.
3. We evaluated FacePhys on large-scale datasets and demonstrated that it outperforms SOTA methods, including a **42.3%** improvement on the challenging MMPD dataset and up to **49%** overall improvement.

2 Related Work

Efficient rPPG Methods. Remote physiological sensing has evolved from handcrafted signal processing techniques such as POS (Wang et al., 2017) to deep learning methods (Gideon and Stent, 2021; Li and Yin, 2023). Early CNN-based rPPG models extract features from facial videos using convolutions and pooling. Seq-rPPG (Wang et al., 2024b) employs 1D-CNN on temporal features, achieving stable performance with minimal 8x8 facial regions. DeepPhys (Chen and McDuff, 2018) and TS-CAN (Liu et al., 2020a) use 2D-CNNs with temporal shift modules for

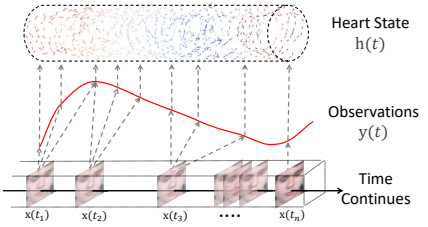


Figure 2: **LTI heart state.** The time-continuous heart state is discretized as a state space model for efficiency.

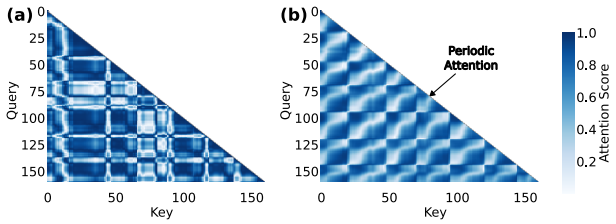


Figure 3: **Periodic attention.** Introducing trainable complex numbers into the diagonalized state transition matrix generates oscillatory terms for periodic attention.

improved heart rate estimation. PhysNet (Yu et al., 2019) and iBVPNet (Joshi and Cho, 2024) leverage 3D-CNNs for spatio-temporal feature extraction. PHASE-Net (Zhao et al., 2026) derives a physics-informed temporal model and uses a causal gated temporal convolutional network for pulse recovery. Despite improvements, CNNs struggle with long-range dependencies, limiting effectiveness in complex scenarios (Tang et al., 2025b). Transformer-based methods address this via self-attention: PhysFormer (Yu et al., 2022b) focuses on temporal features through long-range dependency modeling, while RhythmFormer (Zou et al., 2024) introduces hierarchical periodicity. However, Transformers' quadratic complexity limits practical deployment. Recent Mamba-based approaches (Luo et al., 2024; Wu et al., 2025; Zou et al., 2025; Yan et al., 2024) improve efficiency but remain challenging for embedded devices, motivating more efficient architectures.

Table 1: Computational Efficiency on CPUs

Method	Params.↓	Mem.↓	Latency↓	FLOPs↓
EFFPhys	2,189	609	371	39.0
TSCAN	533	228	55.2	8.07
PhysNet	770	230	62.1	6.16
PhysFormer	7,395	576	2330	95.7
RhyMamba	4,936	347	889	27.4
FacePhys	719	3.6	9.46	0.15

Params. (K) = Number of parameters. Mem (MB) = Memory usage. Latency (ms) = Inference time on CPUs. FLOPs (G) = Floating point operations.

State Space Models (SSMs). Early SSMs such as HiPPO (Gu et al., 2020) and S4 (Gu et al., 2022a) achieve linear-time complexity via continuous-time systems discretized for sequential data, compared to Transformers' quadratic cost. Variants like S5 (Smith et al., 2023), H3 (Fu et al., 2023), Mamba (Gu and Dao, 2023), and Mamba-2 (Dao and Gu, 2024) enhance capacity and extend to vision. Mamba introduces selective state updates, while Mamba-2 establishes the Structured State Space Duality (SSD) framework unifying SSMs and Transformers (Dao and Gu, 2024). In vision, VideoMamba achieves strong performance on video understanding (Li et al., 2024), with Mamba variants succeeding across diverse tasks (Xu et al., 2024). We propose FacePhys as an rPPG-specific SSM leveraging SSD for streaming waveform inference with physiologically-motivated periodic attention.

3 Method

3.1 Basic Model

FacePhys extracts the heart state from continuous input $x(t)$, as shown in Fig. 2. The remote PPG signal $y(t)$ can be regarded as an observation of the heart state $h(t)$. The entire model can be structured as a controlled Linear Time-Invariant (LTI) system, as in Equation (1).

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t) \\ y(t) &= \mathbf{C}h(t) + \mathbf{D}x(t) \end{aligned} \tag{1}$$

where trainable $\mathbf{A}$ is the state transition matrix, $\mathbf{B}$ is the input matrix, $\mathbf{C}$ is the output matrix, $\mathbf{D}$ is the feedforward matrix, $x(t)$ is the control term (face information), $h(t)$ is the heart state, and $y(t)$ is the observation term (rPPG measurement).

We treat the underlying blood volume pulse as a latent dynamical system rather than a pure temporal dependency, and thus the state transition matrix encodes physiological periodicity instead of generic

recurrence. As reported by PhysNet (Yu et al., 2019), the RNN variant underperforms its CNN counterpart; we therefore adopt ODEs as a continuous-time state-space formulation better aligned with physiological dynamics than discrete-time RNNs.

3.2 State Space Duality

State space duality (Dao and Gu, 2024) is an established method for introducing efficient attention. Solving the continuous-time system in Eq. (1) requires discretization, as real videos consist of discrete frames. Following Mamba-2 (Dao and Gu, 2024), we use zero-order hold (ZOH) discretization to obtain a numerically stable and efficient SSM with strong periodic long-range temporal modeling while retaining the continuous-time inductive bias.

Using the ZOH discretization method, any Linear Time-Invariant (LTI) system can be discretized as follows: $\bar{A} = e^{A\Delta t}$ and $\bar{B} = \left(\int_0^{\Delta t} e^{A\tau} d\tau\right) B$. This property ensures the equivalence of state transitions between the continuous and discrete systems, forming the core theorem of SSD.

The recursive relationship of the hidden state h_t is transformed into matrix multiplication via global convolution expansion: $h_t = \sum_{k=1}^{t} \bar{A}^{t-k} \bar{B} x_k$. By constraining matrix $\bar{A}$ to be a diagonal matrix, matrix multiplication and inversion can be avoided. The global hidden state H, composed of hidden states from all time steps, is expressed as (2). Here, L is a lower triangular matrix, whose elements are defined as $L_{i,j} = \bar{A}^{i-j}$ when $i \geq j$, and $\odot$ denotes the Hadamard product.

$$H = (L \odot (C\bar{B}^{\top})) \cdot X \tag{2}$$

By making matrices C and $\bar{B}$ dependent on the input, this formulation is equivalent to linear attention, expressed as $(L \odot (QK^T)) \cdot V$. Thus, SSD exhibits long-range modeling capabilities similar to those of transformers.

Due to the step-by-step inference characteristics of SSMs, linear complexity can be achieved during inference as shown in Table 2, which enables efficient solving of the heart state ODE.

Table 2: **Memory Efficiency.** Comparison Between TSD and Traditional Structures.

Structure	Training Mem.	Inference Mem.
3D CNN	$O(Tk^3D)$	$O(k^3D)$
Transformer	$O(T^2D)$	$O(TD)$
TSD (Ours)	$O(TD)$	$O(D+L)$

L = SSD state dim. k = 3D kernel size. D = Feature dim. T = Time duration.

3.3 General Framework

Our framework is designed as a discretized LTI solver, which is a state space model with a structured dual form.

Due to the three-dimensional nature of video, we employ temporal-spatial decoupled modeling rather than a simple patch embedding. During training, it inputs video frames and outputs the blood volume pulse (BVP) waveform $\mathbf{y} \in \mathbb{R}^{B \times T}$. During inference, the model supports streaming per-frame waveform prediction through recurrent state updates, where each frame produces one waveform sample. Heart rate is then computed from the predicted waveform using standard spectral analysis (FFT/PSD) over a temporal window.

The model processes video sequence tensors $\mathbf{X} \in \mathbb{R}^{B \times T \times H \times W \times C}$ as input and extracts BVP signals through temporal-spatial state space blocks. As illustrated in Figure 4, the computational workflow comprises two critical stages:

1. Temporal Normalization (TN): Independently performs detrending and standardization across the temporal dimension for each spatial coordinate.
2. Temporal-Spatial Modeling: Achieves periodic long-range temporal modeling through Temporal-Spatial SSM-Attention Duality (TSD).

Unlike traditional spatio-temporal joint modeling, TSD decouples spatial and temporal dimensions, utilizing 2D convolutions for spatial modeling while employing SSM duality for temporal modeling. This separation enables memory-efficient inference, achieved efficient attention with linear computational complexity. For a detailed comparison of memory efficiency with conventional architectures, refer to Table 2.

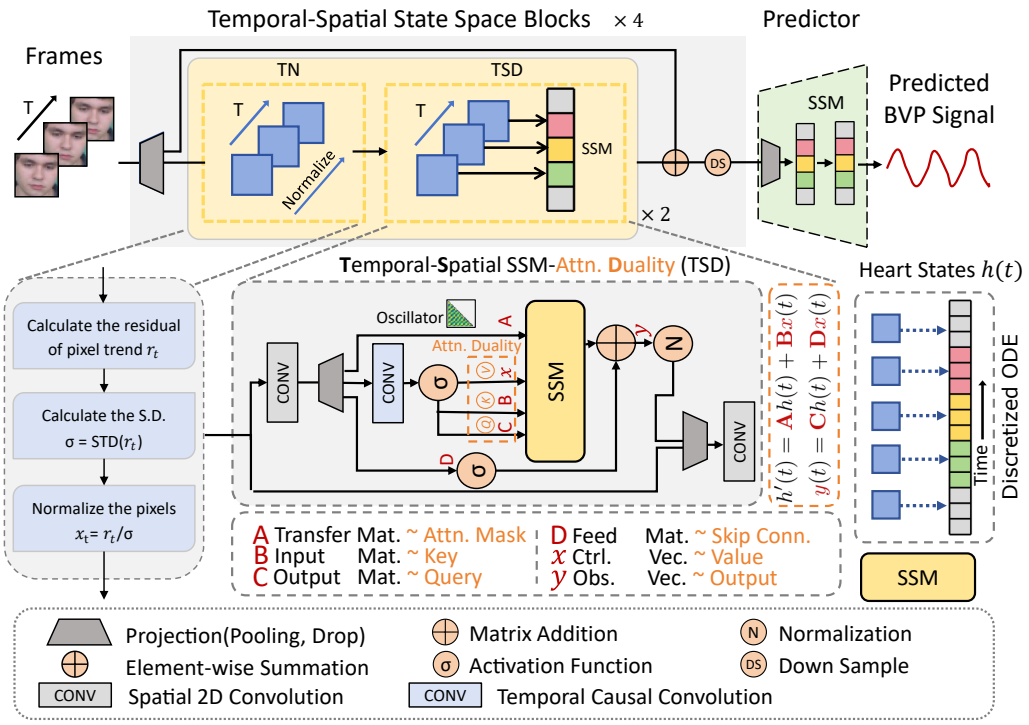


Figure 4: **FacePhys Framework.** Our framework utilizes the SSM dual as the core component, which serves both as an efficient discretization solver for the heart state LTI system and as a linear attention processor. It employs Temporal Normalization (TN) to stabilize the extraction of temporal features and introduces a complex state transition matrix A to enable periodic attention.

3.4 Temporal Normalization Module

The TN module (Wang et al., 2024a) ensures the stability of temporal characteristics in long sequences. It begins by calculating the linear trend term using the least squares method, as formulated in Equation (3). During training, the trend estimation can be computed in parallel.

$$(\beta_0, \beta_1) = \arg\min \beta \sum_{t=1}^{T} (x_t - (\beta_0 t + \beta_1))^2 \tag{3}$$

After obtaining the trend coefficients (β_0, β_1), the residual of each pixel relative to the trend is computed and subsequently standardized as the final output.

$$\begin{cases} \text{Residual} & r_t = x_t - (\beta_0 t + \beta_1) \\ \text{S.D.} & \sigma = \sqrt{\frac{1}{T}\sum_{t=1}^{T} r_t^2} \\ \text{Output} & \tilde{x}_t = r_t / \sigma \end{cases}$$

To achieve constant time and space complexity during inference, the linear trend needs to be replaced with a Recursive Moving Average (RMA) trend.

$$\begin{cases} \text{RMA Trend} & \mu_t = \alpha \mu_{t-1} + (1-\alpha) x_t \\ \text{Residual} & r_t = x_t - \mu_t \\ \text{S.D.} & \sigma_t = \sqrt{\alpha \sigma_{t-1}^2 + (1-\alpha) r_t^2} \\ \text{Output} & \tilde{x}_t = r_t / \sigma_t \end{cases}$$

By updating the trend using RMA, RMA maintains fixed-size running statistics without buffering the full input history. Moreover, RMA exhibits a strong similarity to the linear trend, making it an effective alternative for real-time processing.

3.5 Complex Diagonal State Transition Matrix A

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t) \tag{4}$$

The oscillatory term introduces periodic representation, see Figure 3. The solution to the controlled LTI system (4) consists of a homogeneous solution and a particular solution, where the particular solution depends on the control term, while the homogeneous solution describes the dissipation and oscillation components during state transition, which are determined by the eigenvalues of the trainable matrix A. Since the control term $x(t)$ is governed by the input, we focus on the behavior of the homogeneous term. The cardiac pattern exhibits physiological periodicity, so matrix A should include oscillatory terms. However, since S4D (Gu et al., 2022b), many SSMs fix A to a real diagonal form for efficiency; this removes complex eigenvalues and suppresses oscillations (Eq. (5)). To match physiology without sacrificing speed, we keep the trainable diagonal structure but allow complex diagonal entries, which restores oscillatory behavior.

$$h(t) = \sum_{j=1}^{k} c_j e^{\lambda_j t} v_j + \sum_{l=1}^{m} e^{\alpha_l t} \left[A_l \cos(\beta_l t) + B_l \sin(\beta_l t) \right] \tag{5}$$

Where: c_j are real constants (corresponding to real eigenvalues),A_l and B_l are real constant vectors (corresponding to pairs of complex eigenvalues), α_l and β_l are the real and imaginary parts of the complex eigenvalues. Clearly, a real diagonal matrix A has no complex eigenvalues and cannot produce periodic behavior. However, if we remove the diagonal constraint on A, the computational complexity increases from linear to $O(n^3)$, which cannot meet efficiency requirements. Therefore, introducing complex diagonal elements is considered to improve this. As shown in Eq. (2), the matrix L is equivalent to a causal mask, which describes the causality and decay behavior of state transitions. Its elements are determined by $L_{i,j} = \bar{A}^{i-j}$ when $i \geq j$, and the complex eigenvalues of A lead to periodic $L_{i,j}$, thus it is dual to a periodic attention mask.

4 Experiments

4.1 Datasets

We use five benchmark datasets to evaluate our method: UBFC-rPPG (Bobbia et al., 2019), PURE (Stricker et al., 2014), MMPD (Tang et al., 2023), RLAP-rPPG (Wang et al., 2024b), and VitalVideo (Toye, 2023). UBFC-rPPG includes 43 RGB videos under controlled indoor lighting. PURE contains 59 videos with various head movements for motion robustness testing. MMPD comprises 660 videos under diverse lighting and activities. RLAP features 58 subjects with high synchronization. VitalVideo, the largest dataset, covers 893 subjects with varied skin tones and indoor lighting. These datasets were selected for accessibility, reliability, diversity, and synchronization quality, making them suitable for remote physiological signal extraction evaluation. All baselines are strictly reproduced according to the original papers and use the same training data and processing methods. FacePhys training details are provided in the supplementary material.

4.2 Comparison with State-of-the-Art Methods

Intra-dataset Experiments. We first compare our method with five SOTA methods (Efficient-Phys (Liu et al., 2021), TSCAN (Liu et al., 2020b), PhysNet (Yu et al., 2019), PhysFormer (Yu et al., 2022a), RhythmMamba (Zou et al., 2025)) on three representative datasets (MMPD (Tang et al., 2023), VitalVideo (Toye, 2023), and UBFC (Bobbia et al., 2019)) using 5-fold subject-independent intra-dataset settings. The results are shown in Table 3. Our method outperforms all other methods on almost all datasets, demonstrating its effectiveness in remote PPG and HR estimation. Specifically, FacePhys achieves the lowest MAE of 5.58 on MMPD and the highest Pearson Correlation of 1.00 on UBFC dataset, highlighting its superior accuracy and reliability.

Cross-dataset Experiments. To assess the generalization ability of our method, we evaluate its performance in cross-dataset settings. The results in Tables 4 and 5 indicate that FacePhys achieves

Table 3: **Intra-Dataset Results.** FacePhys performed best across all datasets, when performing five-fold cross-validation.

Model	MMPD			VitalVideo			UBFC		
	MAE	RMSE	R	MAE	RMSE	R	MAE	RMSE	R
EFFPhys	14.8	20.3	0.31	2.35	7.17	0.83	3.41	6.55	0.89
TSCAN	14.2	19.4	0.39	2.93	7.54	0.83	2.36	4.08	0.98
PhysNet	8.13	12.4	0.56	0.69	**2.66**	**0.98**	1.76	2.92	0.98
PhysFormer	9.47	14.0	0.47	0.66	2.76	**0.98**	5.42	8.99	0.85
RhyMamba	8.21	12.0	0.61	0.83	3.69	0.96	0.53	0.73	**1.00**
FacePhys	**5.58**	**10.2**	**0.72**	**0.64**	**2.66**	**0.98**	**0.46**	**0.67**	**1.00**
`Gains (%)`	**+31.4**	**+15.0**	**+18.0**	**+3.0**	**+0.0**	**+0.0**	**+13.2**	**+8.2**	**+0.0**

MAE ↓ = Mean Absolute Error in HR estimation (Beats/Min). RMSE ↓ = Root Mean Squared Error in HR estimation (Beats/Min). R ↑ = Pearson Correlation in HR estimation.

Table 4: **Cross-dataset results.** FacePhys performs best in all comparisons (trained on **RLAP**).

Model	MMPD			VitalVideo			PURE		
	MAE	RMSE	R	MAE	RMSE	R	MAE	RMSE	R
EFFPhys	9.18	13.7	0.56	6.75	19.3	0.51	2.48	5.48	0.98
TSCAN	10.2	15.2	0.47	4.20	8.86	0.78	3.90	7.05	0.96
PhysNet	11.2	16.4	0.42	0.89	3.56	0.96	0.63	2.34	1.00
PhysFormer	12.8	18.9	0.28	4.97	10.1	0.70	0.91	3.32	0.99
RhyMamba	9.62	14.9	0.47	2.85	8.91	0.80	1.63	4.55	0.98
FacePhys	**5.30**	**10.0**	**0.76**	**0.77**	**3.07**	**0.97**	**0.24**	**0.68**	**1.00**
`Gains`	**+42.3**	**+27.0**	**+35.7**	**+13.5**	**+13.8**	**+1.0**	**+61.9**	**+70.9**	**+2.0**

MAE ↓ = Mean Absolute Error in HR estimation (Beats/Min). RMSE ↓ = Root Mean Squared Error in HR estimation (Beats/Min). R ↑ = Pearson Correlation in HR estimation.

Table 5: **Cross-dataset results.** FacePhys performs best in all comparisons (trained on **PURE**).

Model	MMPD			VitalVideo			UBFC		
	MAE	RMSE	R	MAE	RMSE	R	MAE	RMSE	R
EFFPhys	17.6	23.0	0.26	3.83	8.42	0.79	2.96	5.11	0.97
TSCAN	18.1	23.8	0.22	3.24	8.03	0.82	2.71	5.70	0.95
PhysNet	15.1	20.5	0.29	1.32	5.13	0.93	1.62	3.49	0.98
PhysFormer	13.0	18.1	0.44	1.08	4.07	0.95	1.15	2.06	0.99
RhyMamba	10.9	15.2	0.42	1.74	5.87	0.91	0.63	1.07	**1.00**
FacePhys	**8.45**	**13.1**	**0.61**	**0.90**	**3.37**	**0.97**	**0.48**	**0.72**	**1.00**
`Gains`	**+22.5**	**+13.8**	**+38.6**	**+16.7**	**+17.2**	**+2.1**	**+23.8**	**+32.7**	**+0.0**

MAE ↓ = Mean Absolute Error in HR estimation (Beats/Min). RMSE ↓ = Root Mean Squared Error in HR estimation (Beats/Min). R ↑ = Pearson Correlation in HR estimation.

the best performance across *all* the cross-dataset settings, demonstrating strong generalization. Specifically, FacePhys achieves the lowest MAE of 5.30 and the highest R of 0.76 on the challenging MMPD dataset when trained on the RLAP dataset. FacePhys outperforms all methods on the largest VitalVideo dataset (MAE: 0.77, R: 0.97) when trained on RLAP, confirming its robustness and reliability. Compared to cross-dataset results, FacePhys maintains consistent performance across different datasets, whereas other methods show significant performance drops and less consistent performance. This demonstrates our method's effectiveness in handling diverse real-world scenarios and generalizing well across datasets.

Computational Efficiency. We further assess the computational efficiency of our method by comparing its inference time against other SOTA approaches. As shown in Table 1, FacePhys achieves the best trade-off between accuracy and efficiency, underscoring its practicality for real-world applications. FacePhys achieves the lowest latency of 9.46 ms, which is even lower than the frame capture time of 33.3 ms at 30 fps, making it suitable for real-time PPG estimation on resource-constrained devices.

4.3 Ablation Studies

Table 6: **Training and Test Chunk Length Ablation.** Longer video chunks lead to lower HR estimation error.

Train\Test	160	320	640	1280	1800
160	6.13	6.05	6.11	6.15	6.11
320	6.17	6.19	6.17	6.09	5.78
640	6.24	6.05	6.02	5.86	5.79
1280	5.92	5.91	5.80	5.66	5.75
1600	5.93	5.73	5.51	5.42	**5.30**

Experiments are trained on RLAP and tested on MMPD. MAE is calculated in HR estimation (Beats/Min). Each video is divided into chunks of different lengths.

Ablation Study on Components. We conduct an ablation study to validate the effectiveness of each component in FacePhys. We compare the performance after removing different components, including TN, SSD modules and oscillator matrix A. The results in Table 7 show that our method performs best when all components are included. The TN module maintains the stability of temporal features, while the SSD module provides attention to assign weights to different frames. After the trainable state transition matrix A is replaced with complex values, oscillatory terms emerge, which enhances the model's ability to represent periodicity.

Table 7: **Architecture Ablation.** Removing the Temporal Normalization, State Space or Oscillator Matrix A will harm performance.

Model	MMPD			VitalVideo			PURE		
	MAE	RMSE	R	MAE	RMSE	R	MAE	RMSE	R
w/o TN	11.6	17.6	0.43	0.88	3.45	0.97	0.80	3.97	0.99
w/o SSM Duality	10.8	16.7	0.51	1.10	3.80	0.96	0.34	0.87	1.00
w/o Oscillator	7.04	11.7	0.71	0.80	3.08	0.97	0.26	0.72	1.00
FacePhys	**5.30**	**10.0**	**0.76**	**0.77**	**3.07**	**0.97**	**0.24**	**0.68**	**1.00**
`Gains`	**+24.5**	**+14.5**	**+7.0**	**+3.7**	**+0.3**	**+0.0**	**+7.7**	**+5.6**	**+0.0**

MAE ↓ = Mean Absolute Error in HR estimation (Beats/Min). RMSE ↓ = Root Mean Squared Error in HR estimation (Beats/Min). R ↑ = Pearson Correlation in HR estimation.

Ablation Study on Training and Test Chunk Length. We also examine the impact of different training and test chunk lengths on model performance. The results in Table 6 show that longer chunks lead to better performance by capturing more temporal information. This finding aligns with previous studies (Zou et al., 2025) that longer test lengths can improve rPPG performance. With our memory-efficient design, we can train on entire videos, further proving that longer training lengths also enhance performance. Our method achieves the best results with a training chunk length of 1600 and a test chunk length of 1800, demonstrating its effectiveness in handling long-term temporal dependencies.

4.4 Temporal Analysis

Effect of HR estimation window length. We examine how the duration of the HR analysis window affects estimation accuracy using saved BVP predictions from RLAP-trained FacePhys on UBFC. Unlike the training and test chunk lengths in Table 6, this experiment varies the temporal window used for spectral HR estimation. As shown in Figure 5, HR MAE decreases from 7.67 BPM at 1 s to 2.78 BPM at 3 s and 1.83 BPM at 5 s. Longer windows further reduce the error to 1.21 BPM at 15

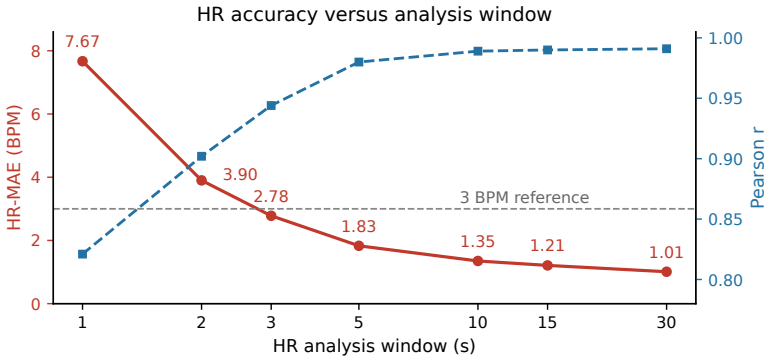


Figure 5: **Effect of HR analysis window length.** Longer windows improve HR estimation accuracy on UBFC, with diminishing gains.

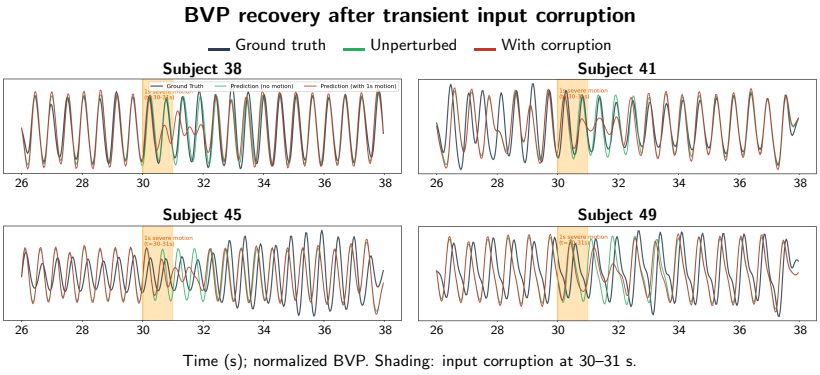


Figure 6: **Recovery after transient input corruption.** Across 12 UBFC subjects, recovery takes 0.92 s on average (maximum: 1.4 s), measured by correlation ≥ 0.9 with the unperturbed output.

s and 1.01 BPM at 30 s, with diminishing improvements. These results characterize the trade-off between observation duration and HR accuracy. Although FacePhys produces BVP samples frame by frame, HR estimation requires accumulating a temporal window. The reported 9.46 ms per-frame inference time therefore describes waveform computation rather than the time required to obtain a stable HR estimate.

Recovery after transient input corruption. To assess whether a brief disturbance causes persistent errors during streaming inference, we evaluate RLAP-trained FacePhys on 12 UBFC subjects. We replace the input frames between 30 and 31 s with random-noise images and then resume the original video. Recovery time is measured from the end of the perturbation until the predicted BVP regains a correlation of at least 0.9 with the output of an unperturbed run. The mean recovery time is 0.92 s, with a maximum of 1.4 s. Figure 6 illustrates the output trajectories for four subjects. The perturbed predictions return toward their unperturbed trajectories after clean input resumes, supporting recovery from transient corruption in this controlled setting. This experiment evaluates output recovery under synthetic corruption; it does not directly measure hidden-state convergence or establish robustness to arbitrary real-world motion.

5 Discussion

5.1 An Effective and Generalizable Model

FacePhys achieves state-of-the-art performance in both intra-dataset and cross-dataset testing, excelling on challenging MMPD and large-scale VitalVideo datasets compared to nearly saturated PURE and UBFC. In cross-dataset validation, FacePhys maintains consistent performance while other methods degrade significantly. When trained on RLAP, MAE on VitalVideo remains below 1 BPM (0.77), with MMPD achieving 5.30 BPM error. When trained on PURE, MAE on both VitalVideo (0.90) and UBFC (0.48) also remains below 1 BPM, demonstrating strong performance.

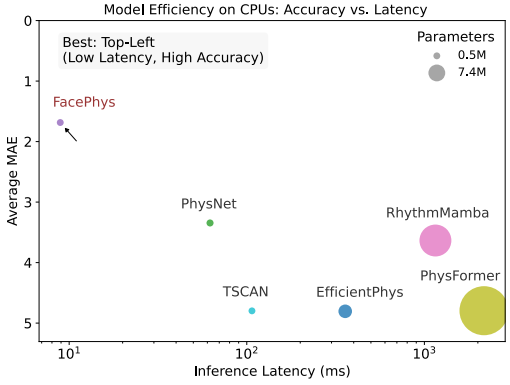


Figure 7: In terms of model accuracy and latency, FacePhys's efficiency performs best.

This robust generalization stems from FacePhys's TN and TSD design, which embeds physiological priors and decoupled spatial and temporal modeling, distinguishing it from prior Mamba works as a physiology-driven SSM. Studies (Yu et al., 2019) found longer videos did not benefit CNNs due to limited receptive fields, while Table 2 shows Transformers incur impractical quadratic costs with increased input length. FacePhys efficiently captures long-term dependen-

cies with minimal overhead, enabling training on extended segments and enhancing cross-dataset generalization.

5.2 Efficient and Real-time FacePhys

Real-time performance is crucial for medical monitoring and video streaming applications. Traditional methods like Green (Verkruysse et al., 2008) process single frames but suffer from motion artifacts, while deep learning methods improve accuracy but require buffering long segments, limiting real-time use. Models like RTrPPG (Botina-Monsalve et al., 2022) claim real-time inference yet require pre-recorded input, and TS-CAN (Liu et al., 2020a) requires 30-second recording before output. With 719K parameters, FacePhys uses 3.6 MB memory and achieves 9.46 ms per-frame latency, enabling streaming waveform inference on mobile devices.

FacePhys's efficiency stems from its streaming inference paradigm: unlike methods constructing temporal dependencies within fixed segments, FacePhys maintains a transferable heart state updated frame-by-frame through recurrent transitions, outputting one waveform sample per frame while accumulating forward information for stable predictions. Heart rate is computed from the waveform via spectral analysis (FFT/PSD) over a temporal window, introducing unavoidable lag inherent to frequency-based HR measurement. The lightweight TN and TSD design enable this streaming architecture for time-sensitive and resource-limited applications. The temporal analyses in Section 4.4 further show that streaming waveform inference can recover from brief synthetic input corruption, while HR accuracy remains dependent on the duration of the spectral analysis window.

5.3 Limitations and Future Work

Although FacePhys has made significant progress in performance and real-time capabilities, certain limitations remain. First, to ensure single frame inference, our model cannot access future information; this introduces lag. Second, although we validated our model on large scale diverse datasets, we have not yet tested it in hospital clinical scenarios, where performance for users with cardiovascular diseases remains to be evaluated.

Future work can extend FacePhys to predict additional cardiovascular signals from facial videos, including blood oxygen (Tang et al., 2025a) and blood pressure (Curran et al., 2023). Enhancing rPPG accuracy through multimodal fusion with thermal imaging or IMU sensors is another potential direction (Ma et al., 2025). Optimizing computational efficiency will enable real-time processing at higher frame rates on lightweight hardware, expanding its practical applications in remote healthcare monitoring.

5.4 Broader Impacts

FacePhys enables accessible cardiovascular monitoring through low-cost deployment on consumer devices, democratizing healthcare for underserved populations. On-device processing minimizes data transmission, enhancing privacy protection compared to cloud-based solutions. However, facial video capture raises privacy concerns, requiring explicit user consent and transparent data policies. The technology could be misused for unauthorized surveillance, necessitating ethical frameworks and regulatory compliance. As rPPG cannot replace clinical-grade devices, systems must clearly communicate uncertainty and recommend physician consultation for diagnosis. Performance may vary across skin tones and demographics; we advocate rigorous subgroup validation and continuous bias monitoring to ensure equitable healthcare access across diverse populations.

6 Conclusion

In this paper, we propose FacePhys, a memory-efficient rPPG algorithm based on temporal-spatial state space duality, designed for arbitrary length video training and inference. Our approach effectively captures subtle temporal changes and spatial representations of facial vascular information, achieving significant improvements over state-of-the-art methods on large-scale datasets such as MMPD and VitalVideo. FacePhys demonstrates superior accuracy and memory efficiency, with improvements of up to 60.3% in accuracy and 98.4% in memory efficiency. To the best of our knowledge, FacePhys is the first rPPG model that supports streaming waveform inference through recurrent state updates, reducing computational overhead. Our method provides new directions for future research on rPPG, effectively balancing accuracy and efficiency for practical deployment.

References

Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019.

Deivid Botina-Monsalve, Yannick Benezeth, and Johel Miteran. Rtrppg: An ultra light 3dcnn for real-time remote photoplethysmography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2146–2154, 2022.

Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018.

Theodore Curran, Xin Liu, Daniel McDuff, Shwetak Patel, and Eugene Yang. Camera-based remote photoplethysmography to measure heart rate and blood pressure in ambulatory patients with cardiovascular disease: Preliminary analysis. *Journal of the American College of Cardiology*, 81 (8_Supplement):2301–2301, 2023.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. *ArXiv*, 2024.

Guanglong Du, Linlin Zhang, Kang Su, Xueqian Wang, Shaohua Teng, and Peter X. Liu. A Multimodal Fusion Fatigue Driving Detection Method Based on Heart Rate and PERCLOS. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):21810–21820, 2022.

Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models, 2023.

John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3975–3984, 2021.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *ArXiv*, 2023.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Re. Hippo: Recurrent memory with optimal polynomial projections, 2020.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces, 2022a.

Albert Gu, Ankit Gupta, Karan Goel, and Christopher Ré. On the parameterization and initialization of diagonal state space models, 2022b.

Junfeng He, Khoi Pham, Nachiappan Valliappan, Pingmei Xu, Chase Roberts, Dmitry Lagun, and Vidhya Navalpakkam. On-device few-shot personalization for real-time gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

Jitesh Joshi and Youngjun Cho. Ibvp dataset: Rgb-thermal rppg dataset with high resolution signal quality labels. *Electronics*, 13(7):1334, 2024.

Hailan Kuang, Can Ao, Xiaolin Ma, and Xinhua Liu. Shuffle-rppgnet: Efficient network with global context for remote heart rate variability measurement. *IEEE Sensors Journal*, 2023.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. VideoMamba: State Space Model for Efficient Video Understanding. arXiv, 2024.

Zhihua Li and Lijun Yin. Contactless pulse estimation leveraging pseudo labels and self-supervision. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20531–20540, 2023.

Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

Ke Liu, Jiankai Tang, Zhang Jiang, Yuntao Wang, Xiaojing Liu, Dong Li, and Yuanchun Shi. Summit vitals: Multi-camera and multi-signal biosensing at high altitudes. In *The 21st IEEE International Conference on Ubiquitous Intelligence and Computing (UIC 2024)*, 2024.

Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020a.

Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. In *Advances in Neural Information Processing Systems*, pages 19400–19411. Curran Associates, Inc., 2020b.

Xin Liu, Brian L. Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based vitals measurement, 2021.

Xuenan Liu, Xuezhi Yang, Dingliang Wang, Alexander Wong, Likun Ma, and Longwei Li. VidAF: A Motion-Robust Model for Atrial Fibrillation Screening From Facial Videos. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1672–1683, 2022.

Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Roni Sengupta, Shwetak Patel, Yuntao Wang, and Daniel McDuff. rppg-toolbox: Deep remote ppg toolbox. In *Advances in Neural Information Processing Systems*, pages 68485–68510. Curran Associates, Inc., 2023.

Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12404–12413, 2021.

Chaoqi Luo, Yiping Xie, and Zitong Yu. Physmamba: Efficient remote physiological measurement with slowfast temporal difference mamba. In *Chinese Conference on Biometric Recognition*, pages 248–259. Springer, 2024.

Xulin Ma, Jiankai Tang, Zhang Jiang, Songqin Cheng, Yuanchun Shi, Dong Li, Xin Liu, Daniel McDuff, Xiaojing Liu, and Yuntao Wang. Non-contact health monitoring during daily personal care routines. In *IEEE BSN 2025*, 2025.

Daniel McDuff. Camera measurement of physiological vital signs. *ACM Computing Surveys*, 55(9): 1–40, 2023.

Daniel McDuff, Ethan B. Blackford, and Justin R. Estepp. The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In *IEEE FG*, 2017.

Ewa Nowara, Daniel McDuff, and Ashok Veeraraghavan. The Benefit of Distraction: Denoising Remote Vitals Measurements using Inverse Attention. *arXiv:2010.07770 [cs, eess]*, 2020. arXiv: 2010.07770.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling, 2023.

Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Yu Liu, and Xun Chen. PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography. *arXiv:2006.02699 [eess]*, 2020. arXiv: 2006.02699.

Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 1056–1062, 2014.

Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. Mmpd: Multi-domain mobile video physiology dataset, 2023.

Jiankai Tang, Xin Liu, Daniel McDuff, Zhang Jiang, Hongming Hu, Luxi Zhou, Nodoka Nagao, Haruta Suzuki, Yuki Nagahama, Wei Li, et al. Camera measurement of blood oxygen saturation. *arXiv preprint arXiv:2503.01699*, 2025a.

Jiankai Tang, Tao Zhang, Jia Li, Yiru Zhang, Mingyu Zhang, Kegang Wang, Yuming Hao, Bolin Wang, Haiyang Li, Xingyao Wang, et al. M3pd dataset: Dual-view photoplethysmography (ppg) using front-and-rear cameras of smartphones in lab and clinical settings. *arXiv preprint arXiv:2511.02349*, 2025b.

Pieter-Jan Toye. Vital videos: A dataset of face videos with ppg and blood pressure ground truths. 2023.

Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

Kegang Wang, Jiankai Tang, Yantao Wei, Mingxuan Liu, Xin Liu, and Yuntao Wang. A plug-and-play temporal normalization module for robust remote photoplethysmography, 2024a.

Kegang Wang, Yantao Wei, Jiankai Tang, Yuntao Wang, Mingwen Tong, Jie Gao, Yujian Ma, and Zhongjin Zhao. Camera-based hrv prediction for remote learning environments, 2024b.

Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017.

Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM transactions on graphics (TOG)*, 31(4):1–8, 2012.

Zheng Wu, Yiping Xie, Bo Zhao, Jiguang He, Fei Luo, Ning Deng, and Zitong Yu. Cardiacmamba: A multimodal rgb-rf fusion framework with state space models for remote physiological measurement. *arXiv preprint arXiv:2502.13624*, 2025.

Rui Xu, Shu Yang, Yihui Wang, Yu Cai, Bo Du, and Hao Chen. Visual Mamba: A Survey and New Outlooks, 2024.

Zhixin Yan, Yan Zhong, Hongbin Xu, Wenjun Zhang, Shangru Yi, Lin Shu, and Wenxiong Kang. Physmamba: State space duality model for remote physiological measurement. *arXiv preprint arXiv:2408.01077*, 2024.

Mototaka Yoshioka and Souksakhone Bounyong. Analysis of pulse transit time derived from imaging photoplethysmography and microwave sensor-based ballistocardiography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *Proc. BMVC*, 2019.

Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *CVPR*, 2022a.

Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4186–4196, 2022b.

Bo Zhao, Dan Guo, Junzhe Cao, Yong Xu, Bochao Zou, Tao Tan, Yue Sun, and Zitong Yu. Phase-net: Physics-grounded harmonic attention system for efficient remote photoplethysmography measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21198–21207, 2026.

Bochao Zou, Zizheng Guo, Jiansheng Chen, and Huimin Ma. Rhythmformer: Extracting rppg signals based on hierarchical temporal periodic transformer. *ArXiv*, abs/2402.12788, 2024.

Bochao Zou, Zizheng Guo, Xiaocheng Hu, and Huimin Ma. Rhythmmamba: Fast, lightweight, and accurate remote physiological measurement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10):11077–11085, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe FacePhys as a memory-efficient rPPG algorithm based on LTI state space models with SSD, achieving 49% error reduction and real-time streaming inference (3.6 MB memory, 9.46 ms latency), which are validated in the experiments section.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5.3 (Limitations and Future Work) discusses limitations including inability to access future information for single frame inference, lack of clinical validation in hospital scenarios, and untested performance for users with cardiovascular diseases.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper is primarily empirical, presenting a novel architecture for rPPG estimation. While it describes the mathematical formulation of the state space model (Eq. 2-6), it does not claim formal theoretical proofs or theorems.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the architecture in detail (Section 3), datasets (Section 4.1), and states that training details are in supplementary material. The abstract mentions training code is provided in supplementary materials.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The abstract states "The training code and real-time inference demo for mobile browsers can be found in the supplementary materials." All datasets used are publicly available benchmark datasets.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 describes datasets and evaluation protocol. The paper states that training details including data splits, hyperparameters, and experimental configurations are provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The paper reports mean performance metrics (MAE, RMSE, Pearson R) across 5-fold cross-validation but does not include error bars or standard deviations. Statistical significance testing would be computationally expensive for the large-scale experiments conducted.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The supplementary material provides detailed compute resource information including hardware specifications, software environment, and training configuration. Table 1 in the paper reports inference latency and memory usage.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available benchmark datasets for rPPG estimation and discusses ethical considerations including privacy, consent, and fairness in Section 5.4 (Broader Impacts).

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 5.4 (Broader Impacts) discusses positive impacts (democratizing healthcare, privacy-preserving on-device processing) and negative impacts (privacy concerns from facial video capture, potential surveillance misuse, performance variations across demographics, limitations vs clinical-grade devices).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The rPPG model for heart rate estimation from facial videos does not pose high misuse risks comparable to generative models. Section 5.4 discusses ethical considerations and recommends user consent and transparent data policies.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets (UBFC-rPPG, PURE, MMPD, RLAP-rPPG, VitalVideo) are properly cited in Section 4.1 and were obtained through author permissions on request. Baseline methods are properly cited throughout the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces the FacePhys model with detailed architecture documentation in Section 3. Training code and real-time inference demo are provided in supplementary materials as stated in the abstract.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper evaluates exclusively on publicly available benchmark datasets and does not involve crowdsourcing or new human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper only uses publicly available benchmark datasets and does not involve new human subjects research requiring IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology of FacePhys is based on state space models and does not involve LLMs as a component. Any LLM usage would be limited to writing assistance, which does not require declaration.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.