A Frequency-Selective Multimodal Remote Physiological Signal Sensing System Based on Mamba Architecture*

Xiuze Jia[2], Xuenan Liu[1,*], Siyi Wang[1], Lizhong Zhang[2] and Shuai Ding[1]

[1]School of Computer Science and Information Engineering,Hefei University of Technology, Hefei, 230601, Anhui, China
[2]School of Software,Hefei University of Technology, Hefei, 230601, Anhui, China

Abstract

Remote heart rate estimation from video still faces three main challenges in real-world scenarios: (1) the absence of adaptive, frequency-selective modeling allows low-frequency physiological rhythms to be overwhelmed by noise; (2) single-modality inputs suffer from instability under varying illumination, occlusions, or device changes; and (3) conventional temporal encoders are computationally expensive and lack long-sequence generalization. To address these limitations, we propose FSMamba, a frequency-selective multimodal perception system built upon the Mamba state-space framework. FSMamba employs a dual-branch feature extractor for RGB and NIR streams and a Joint Cross Attention (JCA) module to enable bidirectional, multi-head cross-modal interaction. In its encoder, we combine the standard MambaBlock with a parallel Frequency-Selective Filter (FSFilter) that uses a learnable time step—derived from trainable heart-rate bounds—and an SSMKernel-based causal recurrence to implicitly generate a band-pass convolution kernel. A channel-wise gating further refines the heart-rate-focused features. The decoder fuses raw temporal and frequency-enhanced representations via joint classification and class-wise regression to predict the final heart rate. Experiments on VIPL-HR demonstrate that FSMamba achieves competitive RMSE performance across the majority of diverse conditions, and ablation studies confirm the effectiveness of each module.

Keywords

remote heart rate estimation, Joint Cross Attention, multimodal fusion, Frequency-Selective Filter, Mamba state-space modeling, SSMKernel

1. Introduction

Remote photoplethysmography (rPPG) is a non-contact technique for estimating physiological indicators such as heart rate and respiration from facial video. It has gained significant attention in recent years for applications in health monitoring and emotion recognition [1, 2, 3]. Classical methods like CHROM [4] and POS [5] rely on color space transformation and fixed bandpass filters, which perform reliably under ideal conditions but are sensitive to lighting changes, head movement, and occlusions [6].

To enhance robustness, recent research has turned to deep learning models such as DeepPhys, PhysNet, MTTS-CAN, and PhysFormer [7, 8, 9, 10], which leverage CNNs or Transformer-based architectures [11] for end-to-end modeling and achieve promising results. However, these models still suffer from sensitivity to input quality variations and limited generalization to unseen devices or challenging environments [12].

Multimodal fusion has emerged as a practical approach to address these limitations, especially through the integration of near-infrared (NIR) signals, which provide more stable skin reflectance under low-light or occluded conditions. While some studies have attempted to combine RGB and NIR modalities [13], most fusion strategies rely on early fusion or simple concatenation, lacking mechanisms to model the dynamic and complementary relationships between modalities.

The 4th RePSS – Multimodal Fusion Learning for Remote Physiological Signal Sensing, to be held at IJCAI 2025, August 28th, 2025, Guangzhou, China.

*This work was supported by the National Natural Science Foundation of China (No. 62401194) and the Fundamental Research Funds for the Central Universities (JZ2024HGTA0192)

*Corresponding author.

†These authors contributed equally.

xiuxejia@gmail.com (X. Jia); xuenanliu@mail.hfut.edu.cn (X. Liu); 2022212064@mail.hfut.edu.cn (S. Wang); 2022213497@mail.hfut.edu.cn (L. Zhang); 2022212134@mail.hfut.edu.cn (S. Ding)

From a spectral perspective, rPPG signals are primarily concentrated in the frequency band of 0.7 – 2.5 Hz. Traditional bandpass filters are fixed and not adaptive to individual or context variations. Learnable filter approaches such as SincNet [14] have shown great potential in related domains like speech processing and physiological signal estimation.

Transformer-based models[11] offer strong global modeling capabilities but suffer from quadratic complexity, making them less suitable for long sequences and edge deployment. In contrast, state-space models (SSMs), including S4 and Mamba [15, 16], provide a more efficient and interpretable way to model long-term dependencies in temporal signals.

On the training side, recent works have explored multi-objective loss formulations—combining regression, interval classification, and distributional alignment—to mitigate label noise and account for sample uncertainty, thereby improving model robustness and generalization [17].

In summary, the current development of rPPG systems is oriented toward three key directions: multi-modal fusion, frequency-aware modeling, and efficient sequential modeling. A major challenge remains in balancing model complexity, accuracy, and deployability, particularly for real-world applications.

2. Modifications

2.1. System Introduction

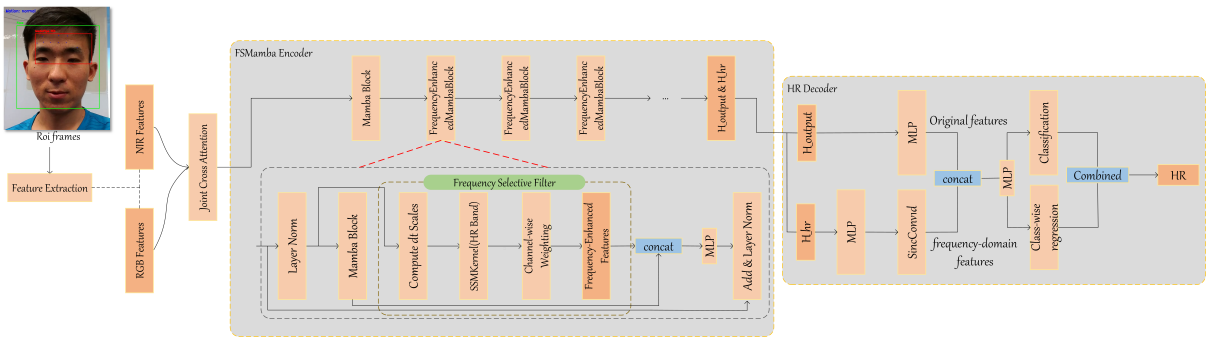

Figure 1: Overview of the proposed FSMamba framework. The system consists of four key modules: (1) a dual-branch *Feature Extractor* that generates spatiotemporal features from RGB and NIR video inputs; (2) a *Joint Cross Attention (JCA)* module that models cross-modal interactions and outputs fused features; (3) the *FSMamba Encoder*, which integrates state-space modeling and frequency-selective filtering to extract both global and heart-rate-specific representations; and (4) an *HR Decoder* that combines temporal and spectral features through MLP and SincConv, and predicts the heart rate via joint classification and class-wise regression.

Traditional rPPG systems face significant challenges, such as sensitivity to lighting variations, motion artifacts, and limited modeling of long-term temporal dependencies. To address these issues, we propose a modular end-to-end frequency-selective multimodal framework with four key modules: (1) a dual-stream feature extractor that encodes spatiotemporal dynamics from RGB and NIR inputs; (2) a Joint Cross Attention (JCA) module for cross-modal interaction and feature alignment; (3) an FSMamba encoder that combines state-space modeling with frequency-aware filtering to capture global and heart-rate-focused representations; and (4) a spectrum-aware decoder that fuses temporal and frequency features for robust heart rate prediction through joint classification and regression.

2.2. Feature Extraction Module

This module employs a dual-branch design based on inter-frame differencing and lightweight convolutional encoding to extract spatiotemporal features from RGB and NIR sequences.

(1) Temporal Difference Construction (STMap) Following PhysFormer [10], we compute inter-frame differences to construct spatiotemporal maps (STMap), which highlight subtle pulse-induced

variations between frames:

$$\mathbf{X}_t = \mathbf{I}_{t+1} - \mathbf{I}_t, \quad t = 1, \ldots, T-1 \tag{1}$$

where $\mathbf{I}_t$ denotes the t-th frame of the input video sequence, and $\mathbf{X}_t$ is the resulting difference map that emphasizes temporal color fluctuations caused by blood flow.

(2) Spatial Encoder Each temporal difference map $\mathbf{X}_t$ is processed through a 5-layer CNN, where each layer consists of a 3×3 convolution, ReLU activation, and Batch Normalization:

$$\mathbf{F}_i = \text{BN}(\text{ReLU}(\text{Conv}_{3\times3}(\mathbf{F}_{i-1}))) \tag{2}$$

where $\mathbf{F}_i$ denotes the intermediate feature map at the i-th layer. The channel dimension doubles progressively across layers. After temporal stacking, the final output $\mathbf{F}_{\text{final}} \in \mathbb{R}^{T\times d}$ is obtained by applying global average pooling (GAP) across spatial dimensions, where T is the number of frames and d is the feature dimension.

(3) Dual-Modality Processing Both RGB and NIR video streams undergo independent STMap generation and spatial encoding. The process for each modality is defined as:

$$\mathbf{X}_{\text{rgb}} = \text{SpatialEncoder}(\text{STMap}(\mathbf{I}_{\text{rgb}})), \quad \mathbf{X}_{\text{nir}} = \text{SpatialEncoder}(\text{STMap}(\mathbf{I}_{\text{nir}})) \tag{3}$$

where $\mathbf{I}_{\text{rgb}}$ and $\mathbf{I}_{\text{nir}}$ are the original RGB and NIR frame sequences, respectively. The outputs $\mathbf{X}_{\text{rgb}}, \mathbf{X}_{\text{nir}} \in \mathbb{R}^{T\times d}$ represent temporally encoded features for each modality, with T as the sequence length and d the feature dimension.

This dual-stream structure ensures that both modalities preserve their complementary spectral information while enabling robust downstream multimodal fusion.

2.3. Multimodal Fusion Mechanism: Bidirectional Cross-Modal Attention

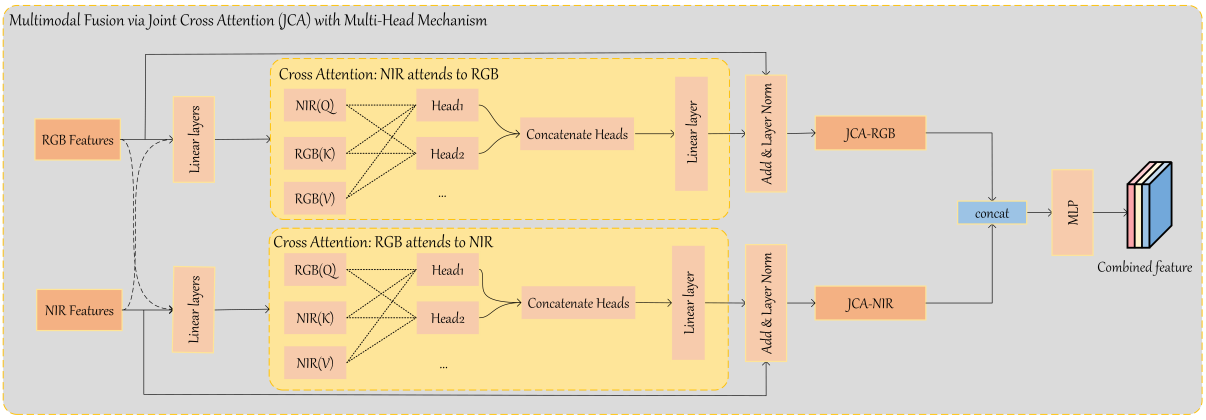

Figure 2: Overview of the Joint Cross Attention (JCA) module. RGB and NIR features are fused via bidirectional multi-head cross-attention, where each modality attends to the other. Each branch includes linear projections, multi-head attention, and residual connections with LayerNorm. The outputs (JCA-RGB and JCA-NIR) are concatenated and refined via an MLP to produce the final fused representation.

To model the interaction between RGB and NIR modalities, we introduce the **Joint Cross Attention (JCA)** module based on the Transformer architecture [11, 13]. Given the sequence features:

$$\mathbf{X}_{\text{rgb}}, \mathbf{X}_{\text{nir}} \in \mathbb{R}^{B\times T\times d} \tag{4}$$

For each attention head h, we compute the attention flow as follows. The attention mechanism is based on the scaled dot-product attention, where we compute the attention for each head using different queries (Q), keys (K), and values (V).

$$\mathbf{F}^h_{\text{rgb}\to\text{nir}} = \text{Attention}(\mathbf{Q}^h_{\text{rgb}}, \mathbf{K}^h_{\text{nir}}, \mathbf{V}^h_{\text{nir}}), \quad \mathbf{F}^h_{\text{nir}\to\text{rgb}} = \text{Attention}(\mathbf{Q}^h_{\text{nir}}, \mathbf{K}^h_{\text{rgb}}, \mathbf{V}^h_{\text{rgb}}) \tag{5}$$

where $\mathbf{Q}^h_{\text{rgb}}$, $\mathbf{K}^h_{\text{nir}}$, and $\mathbf{V}^h_{\text{nir}}$ are the query, key, and value matrices for the h-th attention head, respectively, derived from the RGB and NIR features, and similarly, $\mathbf{Q}^h_{\text{nir}}$, $\mathbf{K}^h_{\text{rgb}}$, and $\mathbf{V}^h_{\text{rgb}}$ are the corresponding matrices for the reverse attention.

The attention function is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{6}$$

where d is the dimensionality of the query and key vectors.

After calculating each attention head for both directions, we concatenate the outputs of all heads. Let H be the number of heads, and denote the concatenation of all heads as:

$$\mathbf{F}_{\text{rgb}\to\text{nir}} = \text{Concat}\left(\mathbf{F}^1_{\text{rgb}\to\text{nir}}, \mathbf{F}^2_{\text{rgb}\to\text{nir}}, \cdots, \mathbf{F}^H_{\text{rgb}\to\text{nir}}\right) \tag{7}$$

$$\mathbf{F}_{\text{nir}\to\text{rgb}} = \text{Concat}\left(\mathbf{F}^1_{\text{nir}\to\text{rgb}}, \mathbf{F}^2_{\text{nir}\to\text{rgb}}, \cdots, \mathbf{F}^H_{\text{nir}\to\text{rgb}}\right) \tag{8}$$

Then, we update each stream via residual connection:

$$\mathbf{X}'_{\text{rgb}} = \mathbf{X}_{\text{rgb}} + \mathbf{F}_{\text{nir}\to\text{rgb}}, \quad \mathbf{X}'_{\text{nir}} = \mathbf{X}_{\text{nir}} + \mathbf{F}_{\text{rgb}\to\text{nir}} \tag{9}$$

Here, $\mathbf{X}'_{\text{rgb}}$ and $\mathbf{X}'_{\text{nir}}$ represent the attended features after cross-modal integration.

Finally, we concatenate and project the updated features from both modalities:

$$\mathbf{X}_{\text{fused}} = \text{MLP}\left(\mathbf{X}'_{\text{rgb}} \parallel \mathbf{X}'_{\text{nir}}\right) \tag{10}$$

2.4. FSMamba Encoder: Frequency-Selective State-Space Modeling

To overcome the limitations of conventional encoders in frequency selectivity and long-range modeling, we propose the FSMamba encoder, which combines the Mamba state-space framework [16] with a frequency-guided path that focuses on the heart rate band ($0.7 \sim 2.5$ Hz).

(1) Overall Architecture Given an input sequence $\mathbf{X}_0 \in \mathbb{R}^{B\times T\times D}$, FSMamba employs a layered encoder structure where the first layer is a standard **MambaBlock**, followed by $L-1$ layers of **Frequency-Enhanced MambaBlocks**

Each enhanced layer processes input as:

$$\mathbf{Y}^{\text{ssm}}_\ell = \text{MambaBlock}_\ell(\text{LN}(\mathbf{X}_\ell)), \quad \mathbf{Y}^{\text{hr}}_\ell = \text{FSFilter}_\ell(\text{LN}(\mathbf{X}_\ell)) \tag{11}$$

where $\mathbf{X}_\ell$ is the input to layer ℓ, $\mathbf{Y}^{\text{ssm}}_\ell$ and $\mathbf{Y}^{\text{hr}}_\ell$ are the outputs of the MambaBlock and FSFilter, respectively.

$$\mathbf{X}_{\ell+1} = \text{LN}\left(\mathbf{X}_\ell + \text{MLP}\left(\mathbf{Y}^{\text{ssm}}_\ell \parallel \mathbf{Y}^{\text{hr}}_\ell\right)\right) \tag{12}$$

where $\mathbf{X}_{\ell+1}$ is the residual-updated output of the current layer, with feature fusion performed via an MLP.

$$\mathbf{H}_{\text{output}} = \text{LN}(\mathbf{X}_L), \quad \mathbf{H}_{\text{hr}} = \frac{1}{L-1}\sum_{\ell=1}^{L-1}\mathbf{Y}^{\text{hr}}_\ell \tag{13}$$

where $\mathbf{X}_L$ is the final output of the encoder, which passes through L layers of MambaBlock and Frequency-Enhanced MambaBlock, and $\mathbf{H}_{\text{output}} = \text{LN}(\mathbf{X}_L)$ is the final encoder representation after layer normalization. Additionally, $\mathbf{H}_{\text{hr}}$ aggregates heart-rate-enhanced features from all FSFilter layers by averaging them across layers.

(2) State-Space Path in MambaBlock: Local Convolution and Global Temporal Modeling The state-space path follows the standard **MambaBlock** [16], which models both short-term and long-range temporal dependencies through a combination of local convolution and global state recurrence. Given input $\mathbf{X}_\ell \in \mathbb{R}^{T\times D}$, the block first applies a depthwise 1D convolution to capture local motion patterns across time. The result is then passed through a dynamic gating unit and projected into a state-space model (SSM) for global modeling.

The SSM core operates via a linear recurrence:

$$\mathbf{s}_{t+1} = \mathbf{A}\mathbf{s}_t + \mathbf{B}\mathbf{x}_t, \quad \mathbf{y}_t = \mathbf{C}\mathbf{s}_t + \mathbf{D}\mathbf{x}_t \tag{14}$$

where $\mathbf{x}_t$ is the input at time t, $\mathbf{s}_t$ is the latent state, and $\mathbf{y}_t$ is the output. The matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are learnable and define a causal filter with global temporal receptive field.

The output is fused with the input via a residual connection and normalized. This design enables MambaBlock to efficiently learn hierarchical temporal features by combining local convolution, global recurrence, and data-driven gating in a lightweight and scalable architecture.

(3) Frequency-Enhanced MambaBlock: Learning Frequency-Focused Representations To enhance sensitivity to periodic physiological dynamics such as heart rate oscillations, we augment the MambaBlock with a parallel **Frequency-Selective Filter (FSFilter)**. This module is implemented via a modified state-space kernel, denoted as $\mathrm{SSM}_{\mathrm{hr\_band}}$, which introduces frequency-domain awareness by dynamically modulating its temporal resolution.

We first define a learnable time step:

$$\Delta t = \frac{2}{f^{\mathrm{hr}}_{\min} + f^{\mathrm{hr}}_{\max}}, \tag{15}$$

where $f^{\mathrm{hr}}_{\min}, f^{\mathrm{hr}}_{\max} > 0$ are trainable scalars initialized to 0.7 and 2.5 Hz respectively. During training, Δt adapts to shift the effective center frequency of the filter.

The FSFilter applies the following causal state-space recurrence for each time step k:

$$s_{k+1} = A_{\Delta t}\, s_k + B_{\Delta t}\, x_k, \quad y_k = C\, s_k, \tag{16}$$

where

- $x_k \in \mathbb{R}^D$ is the input at time k,
- $s_k \in \mathbb{R}^H$ is the latent state,
- $A_{\Delta t}, B_{\Delta t} \in \mathbb{R}^{H\times H}$ are transition matrices modulated by Δt,
- $C \in \mathbb{R}^{D\times H}$ is the output projection.

By unrolling the recurrence, this is equivalent to a causal convolution

$$y_k \;=\; \sum_{n=0}^{k} h_n\, x_{k-n}, \quad h_n = C(A_{\Delta t})^n\, B_{\Delta t}, \tag{17}$$

where $\{h_n\}$ is the implicit filter kernel whose effective bandwidth is controlled by Δt.

Finally, we apply a channel-wise gating to the filter output:

$$\mathbf{Y}^{\mathrm{hr}}_\ell = \mathrm{SSM}_{\mathrm{hr\_band}}(\mathbf{X}_\ell; \Delta t) \;\odot\; \mathbf{w}^{\mathrm{hr}}, \quad \mathbf{w}^{\mathrm{hr}} \in \mathbb{R}^D \tag{18}$$

where $\mathbf{w}^{\mathrm{hr}}$ is a trainable vector and $\odot$ denotes element-wise multiplication. This gating further emphasizes or suppresses specific channels according to their relevance to the heart-rate frequency band.

Through (i) learnable frequency bounds via Δt, (ii) state-space – derived kernel h_n, and (iii) channel-wise gating $\mathbf{w}^{\mathrm{hr}}$, the FSFilter functions as a data-driven band-pass filter centered on the heart rate range, providing explicit frequency-domain selectivity in the FSMamba encoder.

2.5. Heart Rate Decoder Module

To address both interval discrimination and frequency sensitivity, the decoder integrates temporal and frequency-aware features with joint classification and regression objectives [14, 17].

(1) Feature Extraction and Spectral Enhancement: The features from the FSMamba encoder are processed in two branches:

- **Raw Feature Processing:** The temporal features are passed through an MLP to extract raw features:

$$\mathbf{F}_{\text{raw}} = \text{MLP}(\mathbf{H}_{\text{output}}) \tag{19}$$

where $\mathbf{H}_{\text{output}}$ is the global representation from the FSMamba encoder.

- **Frequency Feature Enhancement:** The heart-rate focused features $\mathbf{H}_{\text{hr}}$ are processed through a Sinc convolutional layer followed by an MLP to enhance frequency-specific features:

$$\mathbf{F}_{\text{freq}} = \text{SincConv1d}(\text{MLP}(\mathbf{H}_{\text{hr}})) \tag{20}$$

where $\mathbf{H}_{\text{hr}}$ is the heart-rate focused representation from the FSMamba encoder. The SincConv1d layer is a learnable filter designed to enhance the relevant frequency components for heart rate estimation.

(2) Feature Fusion and Regression: After extracting raw and frequency-enhanced features, they are fused as:

$$\mathbf{F}_{\text{fused}} = \text{MLP}\left(\mathbf{F}_{\text{raw}} \parallel \mathbf{F}_{\text{freq}}\right) \tag{21}$$

Here, the raw temporal features $\mathbf{F}_{\text{raw}}$ and the frequency-enhanced features $\mathbf{F}_{\text{freq}}$ are concatenated and fused using an MLP.

Finally, the heart rate prediction $\hat{y}_{\text{final}}$ is computed using a **softmax** function for classification and a regression head to output the final estimation:

$$\hat{y}_{\text{final}} = \sum_{i=1}^{C} \text{softmax}(\mathbf{F}_{\text{fused}})_i \cdot \text{Reg}_i(\mathbf{F}_{\text{fused}}) \tag{22}$$

where C represents the number of interval classes, and Reg_i denotes the regression output for class i.

This design enhances both frequency focus and interval-level adaptation, yielding robust heart rate predictions even under challenging conditions.

2.6. Loss Function Design

We adopt a multi-objective loss function to jointly optimize prediction accuracy, classification sensitivity, and distributional stability [17]:

$$\mathscr{L}_{\text{total}} = \lambda_{\text{reg}} \cdot \mathscr{L}_{\text{reg}} + \lambda_{\text{cls}} \cdot \mathscr{L}_{\text{cls}} + \lambda_{\text{dist}} \cdot \mathscr{L}_{\text{dist}}, \tag{23}$$

where $\lambda_{\text{reg}} = 1.0$, $\lambda_{\text{cls}} = 1.0$, and $\lambda_{\text{dist}} = 0.5$.

Specifically, the three components are defined as follows:

$$\mathscr{L}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} \text{SmoothL1}(y_i - \hat{y}_{\text{final},i}) \tag{24}$$

$$\mathscr{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^{N} \log(p_{i,c_i}) \tag{25}$$

$$\mathscr{L}_{\text{dist}} = (\mu_{\hat{y}} - \mu_y)^2 + (\sigma_{\hat{y}} - \sigma_y)^2 \tag{26}$$

Here, $\hat{y}_{\text{final}}$ denotes the predicted heart rate, y is the ground truth, p_{i,c_i} is the probability assigned to the true interval class, and μ, σ are the empirical mean and standard deviation. This formulation improves both accuracy and robustness under uncertainty [17].

3. Experimental Design and Result Analysis

We evaluate the proposed FSMamba system through comprehensive comparative and ablation experiments on the VIPL-HR dataset [12], with detailed analysis to validate the effectiveness and contribution of each module.

3.1. Datasets

VIPL-HR contains 2,378 RGB and 752 NIR facial video sequences from 107 subjects under varied conditions (rest, motion, illumination), with synchronized PPG, heart rate, and SpO_2 annotations [12].

Oulu Bio Face Database (OBF) comprises videos from 100 healthy volunteers and 6 AF patients (two 5-minute sessions per subject in RGB + NIR), along with simultaneous ECG, PPG, and respiration signals, for heart rate, respiratory rate, and AF detection benchmarking [18].

3.2. Experimental Settings and Evaluation

We adopt the **Root Mean Square Error (RMSE)** as the primary evaluation metric to assess model robustness under outlier predictions, defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{27}$$

All experiments are conducted on VIPL-HR using the subject-independent protocol [12], with 684/68/198 samples for training, validation, and testing. The model is trained for 15 epochs with Adam optimizer (lr = 1×10^{-4}, batch size = 2) on 180-frame segments.

The architecture includes four modules: a feature extractor, a cross-modal fusion block (4-head attention, 64 dim/head), a 6-layer FSMamba encoder ($d_{\text{model}} = 256$, $d_{\text{state}} = 128$, 30 Hz), and a heart rate decoder (16 SincConv layers, kernel size = 33, band = 0.7 – 2.5 Hz, regression head dim = 128) [16, 14]. Inputs are resized to 128×128, and all features are 256-dimensional. This setup ensures a good trade-off between temporal modeling and frequency selectivity for accurate rPPG estimation.

3.3. Results

As shown in Table 1, the proposed FSMamba system achieves consistent performance across diverse VIPL-HR scenarios. In typical settings (e.g., sitting, talking, bright lighting), RMSE stays within 11 – 13 BPM. Even under low light, long distance, or mobile capture (v4, v6, v8/v9), it remains below 15 BPM, demonstrating robustness to noise and input variation. The highest RMSE (24.17 BPM) appears in the post-exercise recovery scenario (v7), suggesting future improvement is needed for modeling rapid physiological changes. The solid performance in mobile cases further indicates strong potential for real-world deployment.

Ablation results (Table 2) demonstrate the importance of each loss component and system module. Removing $\mathscr{L}_{\text{reg}}$, cross-modal attention, or the NIR modality notably degrades RMSE, confirming their complementary roles. Despite limited NIR data, its inclusion enhances low-light robustness. Overall, FSMamba' s design effectively balances accuracy, generalization, and real-world applicability through frequency-aware modeling and multimodal fusion.

As shown in Table 3, our proposed system (Team: **xiuxejia**, Hefei University of Technology) ranked **6th** on the official RE-PSS leaderboard, evaluated on the VIPL-HR and OBF test sets, with an RMSE of 16.25 BPM.

Due to local resource constraints, we submitted a lightweight version without pre-trained weights or extensive hyperparameter tuning. However, the system still performed well, demonstrating its robustness and potential for deployment.

Table 1
Performance evaluation across different scenarios on the VIPL-HR dataset (unit: BPM)

Scenario	#Samples	RMSE↓	MAE↓	Mean Error	Std. Dev.↓	Max Error↓
v1	22	13.120	10.568	7.042	11.070	28.073
v2	22	12.873	9.891	4.408	12.095	27.821
v3	22	11.943	9.565	2.527	11.672	21.699
v4	22	12.605	10.394	1.551	12.510	28.531
v5	22	10.977	8.807	0.080	10.977	24.656
v6	22	14.578	12.425	1.595	14.490	28.028
v7	22	24.173	18.135	−17.118	17.068	59.033
v8	22	9.743	8.324	−1.592	9.612	24.094
v9	22	10.448	8.763	−3.589	9.812	27.198

Table 2
Ablation Study on VIPL-HR: Module and Loss Impact (RMSE, BPM)

Model Configuration	Loss Terms	RMSE↓ (BPM)
RGB+NIR+JCA	$\mathscr{L}_{\text{reg}} + \mathscr{L}_{\text{cls}} + \mathscr{L}_{\text{dist}}$	**13.98**
Without $\mathscr{L}_{\text{reg}}$	$\mathscr{L}_{\text{cls}} + \mathscr{L}_{\text{dist}}$	15.45
Without $\mathscr{L}_{\text{cls}}$	$\mathscr{L}_{\text{reg}} + \mathscr{L}_{\text{dist}}$	14.15
Without $\mathscr{L}_{\text{dist}}$	$\mathscr{L}_{\text{reg}} + \mathscr{L}_{\text{cls}}$	14.22
Without JCA	$\mathscr{L}_{\text{reg}} + \mathscr{L}_{\text{cls}} + \mathscr{L}_{\text{dist}}$	14.81
RGB-only Modality	$\mathscr{L}_{\text{reg}} + \mathscr{L}_{\text{cls}} + \mathscr{L}_{\text{dist}}$	15.32

Table 3
Top 6 challenge team results ranked by RMSE (unit: BPM) on the official VIPL-HR and OBF evaluation set

Ranking	Team Name	Captain Affiliation	Score (RMSE↓)
1	HFUT-VUT	Hefei University of Technology	11.89505
2	IST	Nanjing University	12.31846
3	xjgroupscu	Sichuan University	12.70790
4	NJU_TEAM	Nanjing University	14.51449
5	Sgt. Pepper's	Hefei University of Technology	14.69105
6	**xiuxejia**	**Hefei University of Technology**	**16.25080**

4. Conclusion

We propose FSMamba, a frequency-selective multimodal framework for heart rate estimation based on the Mamba architecture. It combines RGB − NIR fusion, frequency-aware encoding, and multi-branch loss to enhance robustness. Experiments on VIPL-HR show strong generalization, with future work focusing on adaptability and deployment efficiency.

Declaration on Generative AI

During the preparation of this work, the author(s) used **ChatGPT (OpenAI)** for *translation between Chinese and English of author-written text* and *grammar/spelling/style polishing of author-written paragraphs.* No Generative AI tools were used to generate ideas, design the study, conduct analyses/experiments, produce results, or create figures/tables. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Ming-Zher Poh, Daniel J. McDuff, Rosalind W. Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation., Opt. Express 18 (2010) 10762–10774. URL: https://opg.optica.org/oe/abstract.cfm?URI=oe-18-10-10762. doi:10.1364/OE.18.010762.

[2] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, William Freeman, Eulerian video magnification for revealing subtle changes in the world, ACM Trans. Graph. 31 (2012). URL: https://doi.org/10.1145/2185520.2185561. doi:10.1145/2185520.2185561.

[3] Guha Balakrishnan, Frédo Durand, John V. Guttag, Detecting pulse from head motions in video, 2013 IEEE Conference on Computer Vision and Pattern Recognition (2013) 3430–3437. URL: https://api.semanticscholar.org/CorpusID:17407827.

[4] Gerard de Haan, Vincent Jeanne, Robust pulse rate from chrominance-based rppg, IEEE Transactions on Biomedical Engineering 60 (2013) 2878–2886. doi:10.1109/TBME.2013.2266196.

[5] Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, Gerard de Haan, Algorithmic principles of remote ppg, IEEE Transactions on Biomedical Engineering 64 (2017) 1479–1491. doi:10.1109/TBME.2016.2609282.

[6] Xiaobai Li, Jie Chen, Guoying Zhao, Matti Pietikainen, Remote heart rate measurement from face videos under realistic situations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[7] Weixuan Chen, Daniel McDuff, Deepphys: Video-based physiological measurement using convolutional attention networks, in: Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, Yair Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 356–373.

[8] Zitong Yu, Xiaobai Li, Guoying Zhao, Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks, 2019. URL: https://arxiv.org/abs/1905.02419. arXiv:1905.02419.

[9] Xin Liu, Josh Fromm, Shwetak Patel, Daniel McDuff, Multi-task temporal shift attention networks for on-device contactless vitals measurement, in: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 19400–19411. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/e1228be46de6a0234ac22ded31417bc7-Paper.pdf.

[10] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip H.S. Torr, 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 (CVPR), 2022, pp. 4186–4196.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, Illia Polosukhin, Attention is all you need, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[12] Xuesong Niu, Shiguang Shan, Hu Han, Xilin Chen, Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation, Trans. Img. Proc. 29 (2020) 2409 – 2423. URL: https://doi.org/10.1109/TIP.2019.2947204. doi:10.1109/TIP.2019.2947204.

[13] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, Ruslan Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Anna Korhonen, David Traum, Lluís Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6558–6569. URL: https://aclanthology.org/P19-1656/. doi:10.18653/v1/P19-1656.

[14] Mirco Ravanelli, Yoshua Bengio, Speaker recognition from raw waveform with sincnet, in: 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 1021–1028. doi:10.1109/SLT.2018.8639585.

[15] Albert Gu, Karan Goel, Christopher Ré, Efficiently modeling long sequences with structured state spaces, in: Proceedings of the International Conference on Learning Representations (ICLR), 2022.

URL: https://openreview.net/forum?id=uYLFoz1vlAC.

[16] Albert Gu, Tri Dao, Mamba: Linear-time sequence modeling with selective state spaces, in: Proceedings of the International Conference on Learning Representations (ICLR), 2024. URL: https://openreview.net/forum?id=AL1fq05o7H.

[17] Alex Kendall, Yarin Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

[18] Xuesong Li, Kun Peng, Xiaobai Li, Hu Han, Shiguang Shan, Xilin Chen, The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 242–249. doi:10.1109/FG.2018.00043.