READING THE PULSE ACROSS SCALES: ADAPTIVE WAVELET DECOMPOSITION AND INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote photoplethysmography (rPPG) recovers physiological signals from facial videos, where subtle pulse responses are entangled with motion, illumination variations, and imaging noise. These mixed dynamics evolve over different temporal scales and vary across recordings. Existing methods primarily learn from mixed spatiotemporal representations, while predefined wavelet bases offer limited flexibility to accommodate diverse physiological rhythms and interference patterns. To read the pulse across scales, we propose PhysWave, a framework that couples adaptive wavelet decomposition with hierarchical component interaction. PhysWave adaptively combines multiple learnable wavelet experts at each decomposition level to organize local video dynamics into components at different temporal scales. Since temporal decomposition alone does not isolate physiological information, we model intra-component dependencies through sparse attention with a shared, orthogonally initialized projection and learnable gating to refine query–key matching. Temporal spectra then guide the exchange of spatiotemporal features across components, integrating complementary cues across scales for waveform estimation. Experiments on five public rPPG benchmarks demonstrate competitive performance under intra-dataset and cross-dataset evaluation, while ablation studies support the effectiveness of adaptive decomposition and component interaction. The code is available at the `https://anonymous.4open.science/r/JSNISE-BEBB`.

1 INTRODUCTION

Remote photoplethysmography (rPPG) recovers pulse waveforms and estimates heart rate from subtle skin-color variations in facial videos (Verkruysse et al., 2008), offering a contactless approach to health monitoring (Gedam & Paul, 2020) and stress assessment (Huang et al., 2023). Reliable recovery remains challenging because blood-volume changes contribute only a small fraction of the observed video dynamics. Motion, illumination variations, facial expressions, and imaging noise introduce additional changes that can obscure the physiological response. Moreover, both pulse patterns and interference evolve during a recording, making their mixture difficult to characterize with a uniform temporal analysis.

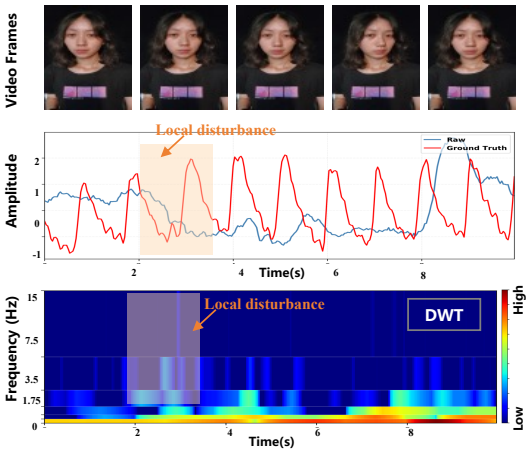


Figure 1: Visualization of the facial signal in the time and wavelet domains. Physiological responses and interference coexist in the temporal trace, while the wavelet characterizes their variations across scales and temporal locations.

Early approaches addressed this problem through handcrafted color processing and signal separation (Poh et al., 2010; de Haan & Jeanne, 2013; Wang et al., 2017). Recently, deep models subsequently shifted the focus toward learning spatiotemporal representations from video, using appearance-guided attention (Chen & McDuff, 2018), 3D convolutions (Yu et al., 2019), and temporal attention mechanisms (Yu et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2022; Zou et al., 2025a). These advances strengthen the modeling of physiological dynamics, but their feature extraction primarily operates on mixed spatiotemporal representations. Without explicit temporal decomposition, variations associated with different processes must be interpreted jointly within the same representation. Our analysis of the skin reflection model in Proposition 1 clarifies this difficulty: ***temporal differencing suppresses stationary appearance but retains a mixture of physiological and non-physiological dynamics.*** As illustrated in Fig. 1, these dynamics can extend over different durations, from sustained pulse oscillations and gradual illumination changes to brief motion disturbances. A useful representation should therefore characterize temporal structure at multiple scales while retaining its location in time. Fortunately, wavelet analysis provides localized responses across temporal scales (Mallat, 1989; Daubechies, 1992), motivating its use for organizing mixed video dynamics.

However, applying wavelets to rPPG raises two related considerations. (*i*) The choice of wavelet basis influences which temporal structures are emphasized during decomposition. Existing wavelet-based rPPG methods employ predefined bases for signal analysis or feature extraction (Bousefsaf et al., 2013; Jaiswal & Meenpal, 2022; Dey et al., 2026), applying the same analysis filters across recordings. Yet wavelet bases exhibit different responses to smooth trends, oscillatory patterns, and abrupt changes, while physiological rhythms, pulse morphology, and interference vary across subjects and imaging conditions. A fixed basis therefore imposes a common analysis preference on videos with different temporal characteristics and may not represent them equally well. This motivates adapting the analysis through combinations of learnable wavelet filters that accommodate the dynamics of each video. (*ii*) Temporal decomposition organizes the observed dynamics by scale, but scale alone does not distinguish physiological responses from interference. Both may coexist within a component, while useful physiological cues may be distributed across multiple components. Decomposition thus provides a structured representation of the mixture whose components still require further interpretation. Fig. 2 compares the original stem feature and individual wavelet components under a RhythmFormer (Zou et al., 2025a) inference pipeline. Their performance varies across noise conditions, suggesting that the usefulness of each component depends on the recording environment. This observation motivates dedicated refinement within each component and information exchange across components for joint physiological signal estimation.

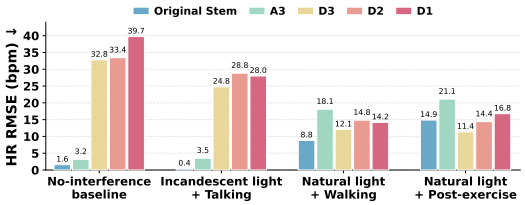


Figure 2: Performance of original stem feature and individual wavelet components by RhythmFormer (Zou et al., 2025a) across different noise scenes on MMPD (Tang et al., 2023), illustrating variation in component effectiveness across recording conditions.

To address these issues, we propose **PhysWave**, which combines adaptive temporal wavelet decomposition with hierarchical component interaction. **PhysWave** learns a bank of wavelet filters and combines them according to the current video features at each decomposition level. It retains the approximation and detail components alongside the original stem feature, providing multiple representations of the observed dynamics. Within each component, sparse attention with gated query–key projection learns spatiotemporal dependencies. Across components, temporal spectra determine affinities that guide the exchange of spatiotemporal features. Waveform predictions from the refined components are then fused to recover the rPPG signal. Our main contributions are:

- We propose **PhysWave** that combines adaptive temporal decomposition with hierarchical component interaction to recover physiological signals from mixed video dynamics.
- We introduce *Adaptive Multi-Expert Wavelet Decomposition (AMWD)*, which adaptively combines learnable wavelet experts for each video and decomposition level to accommodate diverse physiological rhythms and interference patterns.
- We develop *Sparse Orthogonal Component Attention* for intra-component refinement through gated, orthogonally initialized query–key projections, and *Temporal-Spectral Component Interaction (TSCI)* for spectrum-guided cross-component feature exchange.
- We evaluate PhysWave on five public rPPG benchmarks, demonstrating competitive intra- and cross-dataset performance. Ablation studies and visualizations support the effectiveness of adaptive decomposition and component interaction.

2 Related Work

Remote Physiological Measurement. Early studies established the feasibility of recovering physiological signals from facial videos through color analysis and signal separation. GREEN extracts pulse signals from green-channel variations (Verkruysse et al., 2008), while blind source separation identifies pulsatile components from multiple color channels (Poh et al., 2010). To improve robustness to motion and illumination changes, CHROM (de Haan & Jeanne, 2013) combines chrominance signals, and POS (Wang et al., 2017) projects color variations onto a plane orthogonal to the skin-tone direction. Deep learning subsequently enabled joint learning of spatial and temporal representations. DeepPhys (Chen & McDuff, 2018) combines normalized frame differences with appearance-guided attention, whereas PhysNet (Yu et al., 2019) employs 3D convolutions to recover pulse waveforms. EfficientPhys (Liu et al., 2023a) integrates frame differencing and spatiotemporal modeling into an efficient end-to-end framework. Transformer-based approaches further exploit temporal dependencies and physiological periodicity. PhysFormer (Yu et al., 2022) incorporates temporal-difference attention and frequency-domain supervision, while RhythmFormer (Zou et al., 2025a) combines periodic sparse attention with hierarchical temporal modeling. However, these methods primarily focus on spatiotemporal feature learning, without explicitly decomposing mixed video representations into multi-scale temporal components.

Wavelet for Remote Physiological Measurement. Wavelet analysis provides localized time-frequency representations for characterizing non-stationary signals at multiple resolutions (Mallat, 1989). Traditional rPPG approaches primarily apply wavelets to one-dimensional color traces or pulse signals. Bousefsaf et al. (2013) use continuous wavelet analysis for heart-rate estimation, while Finžgar & Podržaj (2018) reconstruct pulse signals through weighted wavelet subbands. More recent approaches incorporate wavelet representations into deep models. Jaiswal & Meenpal (2022) extract spatial-frequency features from video frames using 2D wavelet decomposition, and SFTT (Dey et al., 2026) jointly models spatial, frequency, and temporal information through a pyramidal transformer. Our approach differs from prior wavelet-based rPPG methods (Jaiswal & Meenpal, 2022; Dey et al., 2026), which rely on predefined wavelet bases. We learn a bank of wavelet filters and adaptively combine them for each video, allowing temporal analysis to accommodate variations in physiological rhythms and interference patterns across recordings. Besides, we further incorporate intra- and inter-component modeling to refine the decomposed representations.

3 Methodology

Problem Definition Given a facial video clip $\mathbf{X} \in \mathbb{R}^{3\times T\times H\times W}$, rPPG aims to recover a blood volume pulse waveform $\mathbf{y} \in \mathbb{R}^{T}$, where T denotes the number of frames and $H \times W$ the spatial resolution. We learn a mapping f_θ that predicts $\hat{\mathbf{y}} = f_\theta(\mathbf{X})$ from subtle pulse-related variations entangled with motion, illumination changes, and imaging noise.

Model Overview As illustrated in Fig. 3, a standard Fusion Stem (Zou et al., 2025a) first extracts shallow spatiotemporal features $\mathbf{X}_{\text{stem}} \in \mathbb{R}^{C\times T\times H/8\times W/8}$ from raw frames and their temporal differences. *Adaptive Multi-Wavelet Decomposition* then decomposes $\mathbf{F}_0$ into components at multiple temporal scales. Together with the original stem features $\mathbf{F}_0$, these components are independently refined by *Sparse Orthogonal Component Attention*, followed by cross-component exchange through *Spectrum-Guided Component Interaction*. Finally, we fuse their waveform predictions with learnable weights to produce the rPPG signal.

3.1 Adaptive Multi-Expert Wavelet Decomposition (AMWD)

The Fusion Stem (Zou et al., 2025a) encodes facial appearance and temporal changes, but these changes reflect both physiological activity and external disturbances. To analyze how these sources contribute to the observed dynamics, we derive a time-varying mixture formulation from the skin reflection model in Proposition 1 as follows,

Proposition 1. *Under the first-order skin reflection model (Chen & McDuff, 2018), with locally constant illumination and color conditions, temporal differencing suppresses stationary skin appearance. The remaining RGB dynamics admit the continuous-time representation:*

$$\mathbf{x}(t) = \mathbf{a}_p(t)p'(t) + \mathbf{a}_m(t)m'(t) + \boldsymbol{\epsilon}(t), \tag{1}$$

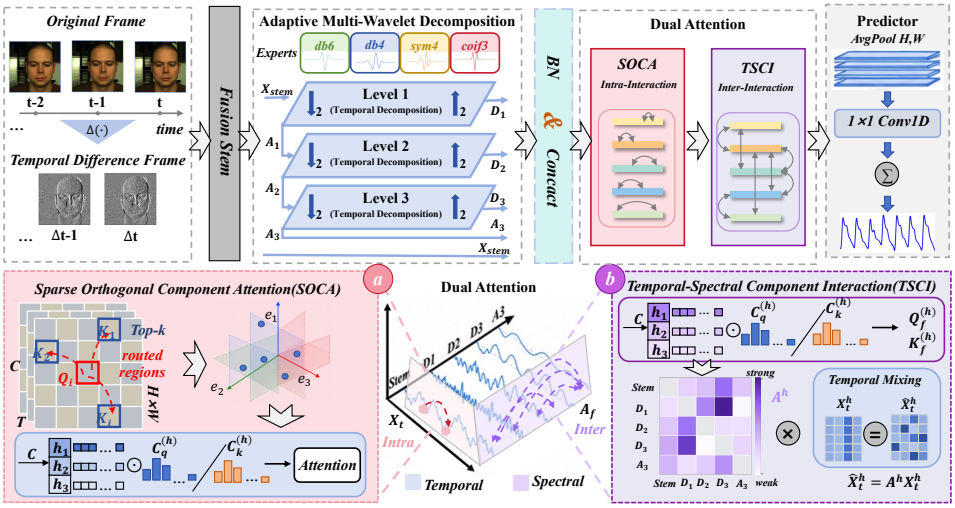


Figure 3: **The framework overview of PhysWave.** Given a facial video, the *Fusion Stem* jointly encodes RGB frames and their temporal differences. *Adaptive Multi-Wavelet Decomposition* then adaptively combines learnable wavelet experts to recursively decompose these features into three detail components $\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3$ and a final approximation $\mathbf{A}_3$. Together with the original stem feature, they form five temporally aligned and independently normalized streams. *Sparse Orthogonal Component Attention* refines each stream through sparse region attention with gated, orthogonally initialized query–key projection. *Spectrum-Guided Component Interaction* estimates component affinities from temporal spectra to guide spatiotemporal feature exchange. Finally, spatial pooling and temporal projection produce a waveform from each stream, and learnable fusion weights combine the five predictions into the estimated rPPG signal.

where $p'(t)$ and $m'(t)$ denote physiological and non-physiological dynamics, respectively; $\mathbf{a}_p(t)$ and $\mathbf{a}_m(t)$ are their effective response coefficients that depend on physiological and imaging conditions; and $\boldsymbol{\epsilon}(t)$ collects measurement and approximation errors. Consequently, the dynamic observation remains a mixture whose source contributions and temporal patterns can vary over time. The detailed proof is provided in ***Appendix A****.*

These mixed dynamics motivate wavelet analysis for its localized multi-scale representation. Prior rPPG methods (Jaiswal & Meenpal, 2022; Dey et al., 2026) have incorporated standard wavelet decomposition but rely on fixed wavelet bases. Across subjects and imaging conditions, such fixed bases may not adequately characterize diverse physiological rhythms and interference patterns. We therefore introduce *Adaptive Multi-Expert Wavelet Decomposition*, which adaptively combines learnable wavelet experts at each decomposition level.

Specifically, given the shallow video feature $\mathbf{X}_{\text{stem}} \in \mathbb{R}^{C\times T\times H/8\times W/8}$, we perform adaptive temporal wavelet analysis using $M = 4$ learnable experts initialized from db4, db6, sym4, and coif3. For expert m, let b_m^q denote the original wavelet coefficients, where $q \in L, H$ indexes the low-pass and high-pass filters. We interpolate these coefficients to a common kernel length K and rescale them to preserve their original L_1 norms:

$$\widetilde{b}_m^q = \text{Interp}(b_m^q, K), \qquad h_{m,c}^q \leftarrow \widetilde{b}_m^q \frac{|b_m^q|_1}{|\widetilde{b}_m^q|_1}, \tag{2}$$

where $h_{m,c}^q$ denotes the learnable temporal filter for channel c. Each channel is initialized with the same rescaled coefficients, which are subsequently optimized during training. At decomposition level i, let $\mathbf{A}_{i-1}$ denote the current approximation, with $\mathbf{A}_0 = \mathbf{X}_{\text{stem}}$. Global average pooling over time and space produces a channel descriptor, which a two-layer MLP maps to expert weights:

$$\mathbf{z}_i = \text{AvgPool}_{t,h,w}(\mathbf{A}_{i-1}), \qquad \boldsymbol{\alpha}_i = \text{Softmax}\left(\text{MLP}(\mathbf{z}_i)\right), \qquad \sum_{m=1}^{M} \alpha_{i,m} = 1. \tag{3}$$

For each sample and decomposition level, we compute routing weights shared across channels and spatiotemporal positions. These weights combine the expert filters into an effective low-pass/high-pass analysis pair:

$$g_i^L = \sum_{m=1}^{M} \alpha_{i,m} h_m^L, \qquad g_i^H = \sum_{m=1}^{M} \alpha_{i,m} h_m^H. \tag{4}$$

Using the same weights for both branches maintains a consistent expert combination for approximation and detail extraction. The resulting filters perform multi-channel convolution at every spatial location. Omitting spatial indices for clarity, filtering and two-fold temporal downsampling yield

$$\mathbf{A}_i[n] = \sum_{u=0}^{K-1} g_i^L[u], A_{i-1}[2n+u], \; \mathbf{D}_i[n] = \sum_{u=0}^{K-1} g_i^H[u], A_{i-1}[2n+u], \tag{5}$$

where $\mathbf{A}_i$ and $\mathbf{D}_i$ are the low-frequency-guided approximation and high-frequency-guided detail responses at level i, respectively. Temporal padding handles boundary positions. The approximation $\mathbf{A}_i$ is recursively decomposed, while the detail response $\mathbf{D}_i$ is retained. For subsequent component modeling, we restore the responses to temporal length T through nearest-neighbor upsampling and length alignment. Let $\uparrow_2$ denote two-fold upsampling, defined along time by $(\uparrow_2 f)[n] = f[\lfloor n/2 \rfloor]$. A lightweight spatial refinement block $\mathcal{R}$ then produces

$$\mathbf{A}_i = \mathcal{R}\left(\uparrow_2 \mathbf{A}_i\right), \qquad \mathbf{D}_i = \mathcal{R}\left(\uparrow_2 \mathbf{D}_i\right). \tag{6}$$

After three decomposition levels, the three detail components and final approximation are retained alongside the undecomposed stem feature:

$$\mathbf{X}_{\text{comp}} = \{\mathbf{X}_{\text{stem}}, \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3, \mathbf{A}_3\}. \tag{7}$$

All components are aligned to the original temporal resolution while retaining responses at different temporal scales. They are independently normalized and forwarded to the subsequent component modeling module

3.2 Sparse Orthogonal Component Attention (SOCA)

While *Adaptive Multi-Expert Wavelet Decomposition* exposes temporal structure at multiple scales, physiological responses and interference can coexist within each component. Based on this consideration, we introduce *Sparse Orthogonal Component Attention* to refine these representations through selective spatiotemporal interaction.

Specifically, each component is independently normalized and spatially embedded. Following RhythmFormer (Zou et al., 2025a), we adopt a three-stage Temporal Periodic Transformer(TPT) with temporal patch sizes $2, 4, 8$ and two blocks per stage. Within each component, temporal center-difference convolutions generate queries $\mathbf{Q}$ and keys $\mathbf{K}$, while a $1 \times 1 \times 1$ projection produces values $\mathbf{V}$. Queries and keys are partitioned into non-overlapping spatiotemporal regions, whose average descriptors select the K most relevant key regions for each query region:

$$\bar{\mathbf{Q}}_r = \text{AvgPool}(\mathbf{Q}_r), \qquad \bar{\mathbf{K}}_s = \text{AvgPool}(\mathbf{K}_s), \qquad \mathcal{N}_K(r) = \text{TopK}_s\left(\bar{\mathbf{Q}}_r \bar{\mathbf{K}}_s^\top\right). \tag{8}$$

Routing determines which regions interact; token matching within those regions further depends on the channel representation. To adapt the query–key representation, we apply a shared projection $\mathbf{W}_o$ to queries and keys. We initialize $\mathbf{W}_o$ orthogonally, satisfying $\mathbf{W}_o^\top \mathbf{W}_o = \mathbf{I}$, so that the initial projection preserves the Euclidean norm of each query and key vector before head-wise scaling. The projection is subsequently optimized without an orthogonality constraint. Learnable scaling then adjusts the contribution of the projected channels within each attention head:

$$\begin{aligned} \mathbf{P}_q^h &= \mathbf{S}_q^h \odot \text{Split}_h\left(\mathbf{W}_o \mathbf{Q}_i\right), \quad \widetilde{\mathbf{Q}}_i^h = \mathbf{Q}_i^h + g\left(\mathbf{P}_q^h - \mathbf{Q}_i^h\right), \\ \mathbf{P}_k^h &= \mathbf{S}_k^h \odot \text{Split}_h\left(\mathbf{W}_o \mathbf{K}_i\right), \quad \widetilde{\mathbf{K}}_i^h = \mathbf{K}_i^h + g\left(\mathbf{P}_k^h - \mathbf{K}_i^h\right), \end{aligned} \tag{9}$$

where $\mathbf{S}_q^h$ and $\mathbf{S}_k^h$ are learnable scaling factors and $g = \tanh(\gamma)$. We initialize $\gamma = 0$, allowing attention to start from the original queries and keys and learn the contribution of the projected features.

Orthogonality applies only to initialization; $\mathbf{W}_o$ is subsequently optimized without this constraint. The refined features determine attention over tokens in the selected regions:

$$\mathbf{O}_r^h = \mathrm{Softmax}\left(\frac{\widetilde{\mathbf{Q}}_r^h \left(\widetilde{\mathbf{K}}^h_{\mathcal{N}_K(r)}\right)^\top}{\sqrt{d_h}}\right)\mathbf{V}^h_{\mathcal{N}_K(r)}, \tag{10}$$

where d_h denotes the head dimension. The head outputs are concatenated, combined with a depth-wise convolution branch for local positional enhancement, and projected back to the component feature space. The detailed illustration of the SOCA pipeline is provided in **Appendix E.1.**

3.3 Temporal-Spectral Component Interaction (TSCI)

Independent refinement captures dependencies within each component but leaves relationships across components unexplored. Since different components may contain complementary physiological cues, we compare their temporal spectra to characterize rhythmic content across the clip. The resulting affinities guide information exchange between their spatiotemporal features.

Let $\mathbf{X} \in \mathbb{R}^{B\times N\times C\times T\times H'\times W'}$ denote the refined features at the interaction stage, where $N = 5$. Spatial average pooling yields a temporal sequence $\mathbf{s}_n \in \mathbb{R}^{B\times C\times T}$ for each component. We remove its temporal mean, apply a real-input Fourier transform, and construct a normalized log-magnitude descriptor:

$$\mathbf{P}_n = \mathrm{Norm}_F\left[\log\left(1 + \left|\mathrm{rfft}\left(\mathbf{s}_n - \mathrm{Mean}_T(\mathbf{s}_n)\right)\right|\right)\right], \tag{11}$$

Here, Norm_F divides each channel's descriptor by its mean over frequency bins. Temporal centering suppresses the DC component, while logarithmic compression and normalization reduce the influence of large spectral peaks and magnitude differences across components. We split the descriptors along the channel dimension into attention heads and apply separate learnable scaling factors to form spectral queries and keys. Within each head, the channel and frequency dimensions are flattened, followed by L_2 normalization of each component descriptor. Omitting the batch index, the component affinities are

$$\mathbf{A}^{(h)} = \mathrm{Softmax}_m\left(\bar{\mathbf{Q}}_f^{(h)}\bar{\mathbf{K}}_f^{(h)\top} + \mathrm{softplus}(\delta)\mathbf{I}_N\right), \tag{12}$$

where $\bar{\mathbf{Q}}_f^{(h)}$ and $\bar{\mathbf{K}}_f^{(h)}$ contain the normalized descriptors, and $\mathbf{A}^{(h)} \in \mathbb{R}^{N\times N}$. Softmax operates over source components m. The learnable nonnegative diagonal bias favors each component's own features while permitting exchange with other components. These spectral affinities weight the original spatiotemporal features at corresponding positions:

$$\widetilde{\mathbf{X}}_n^{(h)} = \sum_{m=1}^{N} A_{n,m}^{(h)}\mathbf{X}_m^{(h)},\ \mathbf{X}'_n \quad = \mathbf{X}_n + \tanh(g)\left(\mathrm{Concat}_h\left[\widetilde{\mathbf{X}}_n^{(h)}\right] - \mathbf{X}_n\right). \tag{13}$$

Here, $\mathbf{X}_m^{(h)}$ denotes the channel group assigned to head h, and g is a learnable scalar initialized to zero. The update therefore starts as an identity mapping and learns the contribution of cross-component exchange during training. Each head uses one attention matrix per clip, shared across temporal and spatial positions. Spectral descriptors determine the exchange weights, while the original features supply the information for waveform prediction. Each output component is then mapped to a candidate rPPG waveform, and learnable softmax-normalized weights fuse the five predictions into the final signal. The detailed illustration of the TSCI pipeline is provided in **Appendix E.2.**

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We evaluate PhysWave on six public rPPG benchmarks: **UBFC-rPPG** (Bobbia et al., 2019), **PURE** (Stricker et al., 2014), **MMPD** (Tang et al., 2023), **BUAA** (Xi et al., 2020), **VIPL-HR** (Niu et al., 2019), and **MR-NIRP-Car** (Nowara et al., 2020). We report mean absolute error (MAE), root mean square error (RMSE), and Pearson's correlation coefficient (r) for HR estimation. Dataset details are provided in **Appendix B**.

Table 1: **Intra-dataset HR estimation results.** Models are trained and tested on the PURE, UBFC-rPPG, BUAA, MMPD, and MR-NIRP-Car datasets. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	PURE			UBFC-rPPG			BUAA			MMPD			MR-NIRP-Car		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
GREEN (Verkruysse et al., 2008)	7.23	17.05	0.69	7.50	14.41	0.62	6.89	10.39	0.60	21.68	27.69	-0.01	11.88	14.46	0.17
CHROM (de Haan & Jeanne, 2013)	2.07	2.50	0.99	2.37	4.91	0.89	/	/	/	13.66	18.76	0.08	7.16	9.09	0.35
ICA (Poh et al., 2010)	3.76	12.60	0.85	5.17	11.76	0.65	/	/	/	18.60	24.30	0.01	10.87	13.62	0.02
PhysNet (Yu et al., 2019)	2.10	2.60	0.99	2.95	3.67	0.97	10.89	11.70	-0.04	4.80	11.80	0.60	11.18	13.53	0.17
DeepPhys (Chen & McDuff, 2018)	0.83	1.54	0.99	6.27	10.82	0.65	/	/	/	22.27	28.92	-0.03	11.98	14.18	0.05
PhysFormer (Yu et al., 2022)	1.10	1.75	0.99	0.50	**0.71**	0.99	8.46	10.17	-0.06	11.99	18.41	0.18	10.17	12.57	-0.20
EfficientPhys (Liu et al., 2023a)	4.75	9.39	0.99	1.14	1.81	0.99	16.09	16.80	0.14	13.47	21.32	0.21	18.36	20.01	-0.04
Contrast-Phys+ (Sun & Li, 2024)	0.48	0.98	0.99	**0.21**	0.80	0.99	/	/	/	4.36	10.76	0.65	/	/	/
RhythmFormer (Zou et al., 2025a)	0.27	0.47	0.99	0.50	0.78	0.99	9.19	11.93	-0.10	3.07	6.81	0.86	/	/	/
RhythmMamba (Zou et al., 2025b)	0.23	0.34	0.99	0.50	0.75	0.99	11.90	14.22	0.16	3.16	7.27	0.84	8.07	10.03	0.24
PhysWave (Ours)	**0.20**	**0.33**	**0.99**	0.50	**0.71**	**0.99**	**1.91**	**2.02**	**0.99**	**2.81**	**6.33**	**0.88**	**7.10**	**8.59**	**0.39**

Table 2: **Cross-dataset HR estimation results.** Models are trained on PURE/UBFC-rPPG and tested on UBFC-rPPG/PURE/MMPD.

Method	PURE → UBFC-rPPG			UBFC-rPPG → PURE			PURE → MMPD			UBFC-rPPG → MMPD		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
GREEN (Verkruysse et al., 2008)	19.73	21.00	0.37	10.09	23.85	0.34	21.68	27.69	-0.01	21.68	27.69	-0.01
CHROM (de Haan & Jeanne, 2013)	4.06	8.83	0.89	5.77	14.93	0.81	13.66	18.76	0.08	13.66	18.76	0.08
PhysNet (Yu et al., 2019)	1.63	3.79	0.98	9.36	20.63	0.62	13.22	19.61	0.23	10.24	16.54	0.29
PhysFormer (Yu et al., 2022)	1.44	3.77	0.98	12.92	24.36	0.47	14.57	20.71	0.15	12.20	17.79	0.17
EfficientPhys (Liu et al., 2023a)	2.13	3.00	**0.99**	5.47	17.04	0.71	14.03	21.62	0.17	13.78	22.25	0.09
RhythmFormer (Zou et al., 2025a)	0.89	1.83	**0.99**	0.97	3.36	**0.99**	8.98	14.85	**0.51**	9.08	15.07	**0.53**
RhythmMamba (Zou et al., 2025b)	0.95	1.83	**0.99**	1.98	6.51	0.96	10.44	16.70	0.36	10.63	17.14	0.34
rPPG-TFCL (Zhao et al., 2025)	**0.78**	1.64	**0.99**	0.60	1.11	**0.99**	9.67	15.07	0.43	8.36	14.43	**0.53**
PhysWave (Ours)	0.84	**1.60**	**0.99**	**0.43**	**0.74**	**0.99**	**8.80**	**14.67**	0.47	**8.01**	**13.77**	0.49

Implementation Details. PhysWave is implemented in PyTorch using rPPG-Toolbox (Liu et al., 2023b). For each video clip, the facial region detected in the first frame is cropped, resized to 128×128, and fixed throughout the clip. We use $L = 3$ wavelet decomposition levels, Top-$K = 40$ for sparse attention, batch size $B = 4$, and train PhysWave for 30 epochs on NVIDIA RTX 4090 GPUs. The model is optimized using a composite loss consisting of a time-domain negative Pearson correlation loss, a frequency-domain cross-entropy loss, and a KL-divergence-based heart-rate distribution consistency loss. Post-processing applies a second-order Butterworth band-pass filter with passband $[0.75, 2.5]$ Hz, followed by Welch PSD for heart-rate estimation.

4.2 Intra-Dataset HR Evaluation

We conduct intra-dataset training and evaluation on five benchmark datasets, including PURE, UBFC-rPPG, BUAA, MMPD, and MR-NIRP-Car. Detailed dataset splits are provided in **Appendix B**. As shown in Tab. 1, PhysWave achieves competitive performance across all five benchmarks, with particularly strong results under challenging acquisition conditions. Specifically, it reduces RMSE by 80.1% on BUAA and 7.0% on MMPD relative to the previous best results. PhysWave also achieves the best overall performance on MR-NIRP-Car under realistic driving conditions. These results indicate that PhysWave can more effectively preserve physiological dynamics while suppressing interference from illumination, motion, and other non-physiological disturbances.

4.3 Cross-Dataset HR Evaluation

Laboratory-to-Laboratory Generalization. We evaluate cross-dataset transfer between laboratory datasets. As shown in Tab. 2, existing methods often degrade under domain shifts, particularly when MMPD is the target. PhysWave shows more stable transfer performance, reducing RMSE by 78.0% for UBFC-rPPG→PURE and RMSE by 8.6% for UBFC-rPPG→MMPD. These results indicate better robustness to dataset-specific variations in motion, illumination, and acquisition conditions.

Laboratory-to-Real-World Generalization. We further evaluate generalization from laboratory datasets to the real-world MR-NIRP-Car dataset. As shown in Tab. 3, the transition to realistic driving conditions introduces stronger non-stationary variations in motion, illumination, and acquisition

Table 3: **Cross-dataset generalization from different laboratory settings to the real-world MR-NIRP-Car dataset.**

Method	UBFC-rPPG→MR-NIRP-Car			MMPD→MR-NIRP-Car			VIPL-HR→MR-NIRP-Car			Average		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
GREEN Verkruysse et al. (2008)	11.01	14.60	0.19	11.01	14.60	0.19	11.01	14.60	0.19	11.01	14.60	0.19
CHROM de Haan & Jeanne (2013)	9.70	13.71	0.28	9.70	13.71	0.28	9.70	13.71	0.28	9.70	13.71	0.28
PhysNet Yu et al. (2019)	14.51	20.63	0.16	13.63	16.86	0.13	10.37	15.05	0.16	12.84	17.51	0.15
PhysFormer Yu et al. (2022)	10.95	13.51	0.02	11.29	13.81	0.10	12.32	14.26	0.01	11.52	13.86	0.04
EfficientPhys Liu et al. (2023a)	6.80	11.09	0.48	7.69	13.00	0.39	11.69	14.53	0.01	8.73	12.87	0.29
RhythmMamba Zou et al. (2025b)	6.71	11.48	0.48	15.99	18.92	-0.21	10.05	13.03	0.14	10.92	14.48	0.14
PhysWave (Ours)	**4.89**	**8.17**	**0.71**	**6.67**	**10.97**	**0.42**	**6.32**	**9.89**	**0.50**	**5.96**	**9.68**	**0.54**

Table 4: Ablation of the main components in PhysWave.

AMWD	SOCA	TSCI	MAE↓	RMSE↓	r ↑
✓	✓	✓	**2.81**	**6.30**	**0.88**
×	✓	✓	3.00	7.06	0.85
✓	×	✓	4.01	8.23	0.81
✓	✓	×	2.88	6.43	0.88

Table 5: Adaptive vs. fixed wavelet decomposition.

Wavelet	MAE↓	RMSE↓	r ↑
Adaptive	**2.81**	**6.33**	**0.88**
db6	3.04	6.96	0.86
sym4	3.47	7.82	0.81
db4	3.48	7.89	0.81
coif3	3.48	7.88	0.81

Table 6: Ablation of different wavelet expert sets.

Wavelet Expert Set	MAE↓	RMSE↓	r ↑
db6 + sym4 + db4 + coif3	**2.81**	**6.33**	**0.88**
db6 + sym4	3.56	8.12	0.81
db6 + sym4 + db4	3.01	7.17	0.84
db2 + sym6 + coif2 + db8	3.40	7.88	0.81
db6 + sym4 + db4 + coif2	3.02	7.16	0.85

environments, leading to substantial performance degradation for many existing methods. Notably, greater motion and illumination diversity in the source domain does not necessarily translate into better real-world generalization, indicating that source diversity alone is insufficient to overcome large scene shifts. In contrast, PhysWave maintains stable performance across all three source domains and achieves the lowest average MAE and RMSE. Relative to the strongest competing averages, these errors are reduced by 31.7% and 24.8%, respectively. This indicates that PhysWave is less sensitive to scene variations and can preserve reliable physiological dynamics while suppressing non-stationary interference in realistic environments. Additional cross-dataset evaluations under different source-target settings are provided in **Appendix C**.

4.4 Ablation Study

All ablation studies are conducted on the MMPD dataset. Additional analysis of orthogonal initialization is provided in **Appendix C.3.**

Effect of Main Components. As shown in Tab. 4, replacing adaptive wavelet decomposition with a conventional hierarchical multi-scale construction degrades performance, supporting the effectiveness of the proposed temporal decomposition. Removing intra-component interaction also reduces accuracy, highlighting the importance of refinement within each component. Likewise, the decline without spectral inter-component interaction supports using spectral relationships to exchange complementary cues across temporal scales.

Effect of Adaptive Wavelet Decomposition. We further evaluate the design of adaptive wavelet decomposition from two aspects: **the decomposition strategy** and **the choice of wavelet experts**. For the decomposition strategy, Tab. 5 compares the proposed adaptive decomposition with fixed-wavelet alternatives. Overall, the adaptive design achieves better performance than using a single fixed mother wavelet, indicating that a sample-independent decomposition strategy is insufficient to consistently handle varying input conditions. To further examine whether the adaptive mechanism indeed responds differently to different inputs, Fig. 4 visualizes the scene-wise routing deviations. Distinct routing patterns are observed across different scenes, with the relative contributions of individual wavelet experts changing noticeably with the input condition. This suggests that the router adjusts the expert combination according to the input rather than relying on a fixed expert preference. Beyond the routing strategy itself, the composition of the expert pool also affects the

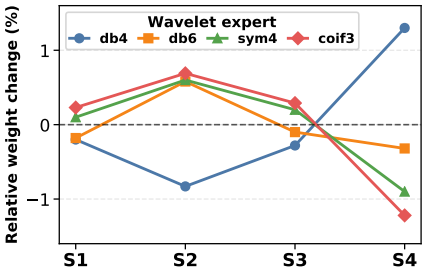
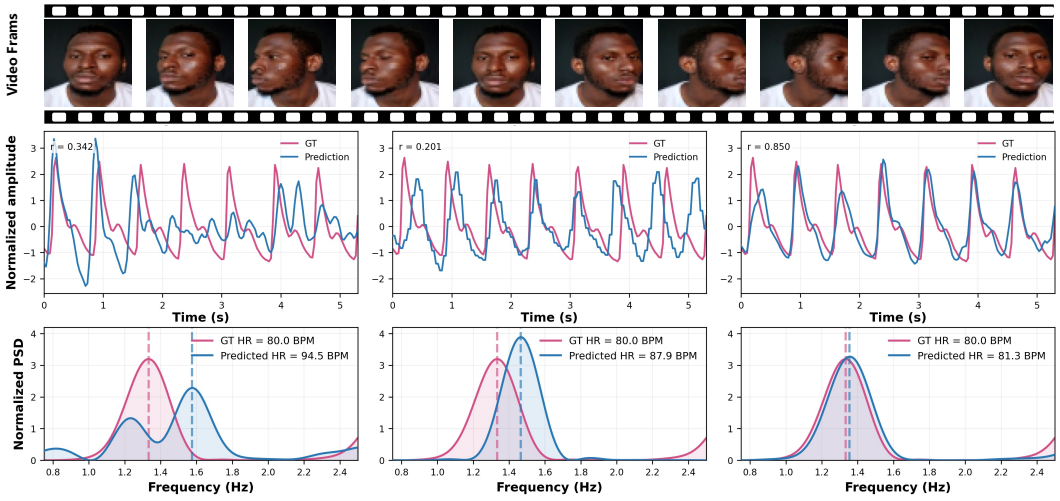


Figure 4: Scene-wise wavelet-expert routing deviations.



(a) RhythmFormer (b) RhythmMamba **(c) PhysWave (Ours)**

Figure 5: **Visualizations of rPPG predictions in the time and frequency domains on MMPD.**The top row shows representative video frames from the same input sequence. The middle and bottom rows show the predicted rPPG waveforms and corresponding power spectral densities (PSDs), respectively. Results are shown for (a) RhythmFormer, (b) RhythmMamba, and (c) PhysWave. Vertical dashed lines indicate the ground-truth and predicted heart rates.

final performance. As shown in Tab. 6, different combinations of mother wavelets lead to different estimation results, indicating that the choice of candidate wavelet experts also influences the adaptive decomposition.

4.5 Visualization

Fig. 5 provides a qualitative comparison of RhythmFormer, RhythmMamba, and PhysWave on a representative MMPD sample. The top row shows video frames from the same input sequence, illustrating the facial variations that may interfere with rPPG estimation. The middle row compares the predicted rPPG waveforms with the ground truth in the time domain, while the bottom row presents their corresponding power spectral densities to examine the dominant periodic components. Overall, the comparison reveals clear differences in how the three methods preserve the underlying physiological rhythm under facial dynamics. Compared with the baselines, PhysWave produces a temporal waveform that more closely follows the ground-truth variation and a spectral distribution whose dominant peak is better aligned with the reference. This qualitative result illustrates that PhysWave provides more consistent modeling of both temporal waveform evolution and spectral periodicity under challenging input variations. Additional attention visualizations are provided in **Appendix D.**

5 Conclusion

In this paper, we presented PhysWave, a framework for recovering physiological signals from facial videos through adaptive wavelet decomposition and hierarchical component interaction. By combining learnable wavelet experts for each video and decomposition level, PhysWave organizes mixed dynamics into temporal components while accommodating variations in physiological rhythms and interference patterns. Sparse attention with gated, orthogonally initialized query–key projections refines each component, while spectral affinities guide information exchange across scales. Experiments on five public rPPG benchmarks demonstrate competitive intra- and cross-dataset performance, with ablation studies supporting the contributions of the proposed modules. These findings highlight the value of adapting temporal decomposition to video dynamics and jointly modeling dependencies within and across components for physiological signal recovery.

6 LLM USAGE

Large Language Models (LLMs) were used solely to assist with writing and polishing this manuscript. In particular, we employed an LLM to refine language, improve readability, and enhance the overall clarity and flow of the text (e.g., grammar checking, sentence rephrasing). The LLM was not involved in ideation, research methodology, experimental design, or data analysis. All scientific concepts, contributions, and results presented in this paper were conceived and executed entirely by the authors. The role of the LLM was strictly limited to improving linguistic quality.

The authors take full responsibility for the entire content of the manuscript, including any sections refined with LLM assistance. All usage adhered to ethical standards, and no plagiarism or scientific misconduct was introduced.

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.

Frédéric Bousefsaf, Choubeila Maaoui, and Alain Pruski. Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate. *Biomedical Signal Processing and Control*, 8(6):568–574, 2013. doi: 10.1016/j.bspc.2013.05.010.

Weixuan Chen and Daniel McDuff. DeepPhys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 349–365, 2018. doi: 10.1007/978-3-030-01216-8_22.

Ingrid Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992. doi: 10.1137/1.9781611970104.

Gerard de Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013. doi: 10.1109/TBME.2013.2266196.

R. Dey, S. Palaiahnakote, S. Bhattacharya, U. Pal, and S. Chanda. SFTT: A spatial-frequency-temporal-based end-to-end transformer for heart rate estimation. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 10(1):454–467, 2026. doi: 10.1109/TETCI.2025.3582841.

Miha Finžgar and Primož Podržaj. A wavelet-based decomposition method for a robust extraction of pulse rate from video recordings. *PeerJ*, 6:e5859, 2018. doi: 10.7717/peerj.5859.

Shruti Gedam and Sanchita Paul. Automatic stress detection using wearable sensors and machine learning: A review. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–7. IEEE, 2020.

Bin Huang, Shen Hu, Zimeng Liu, Chun-Liang Lin, Junfeng Su, Changchen Zhao, Li Wang, and Wenjin Wang. Challenges and prospects of visual contactless physiological monitoring in clinical study. *npj Digital Medicine*, 6(1):231, 2023. doi: 10.1038/s41746-023-00973-x.

Kokila Bharti Jaiswal and T. Meenpal. Heart rate estimation network from facial videos using spatiotemporal feature image. *Computers in Biology and Medicine*, 151:106307, 2022. doi: 10.1016/j.compbiomed.2022.106307.

Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. EfficientPhys: Enabling simple, fast and accurate camera-based cardiac measurement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5008–5017, 2023a.

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 *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023b. URL https://openreview.net/forum?id=q4XNX15kSe.

Stephane G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. doi: 10.1109/34.192463.

Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. RhythmNet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29: 2409–2423, 2019.

Ewa M. Nowara, Tim K. Marks, Hassan Mansour, and Ashok Veeraraghavan. Near-infrared imaging photoplethysmography during driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(4):3589–3600, 2020.

Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762–10774, 2010. doi: 10.1364/OE.18.010762.

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

Zhaodong Sun and Xiaobai Li. Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Jiankai Tang, Ke Chen, Yuntao Wang, Yuzhe Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. MMPD: Multi-domain mobile video physiology dataset. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–5. IEEE, 2023.

Wim Verkruysse, Lars O. Svaasand, and J. Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434–21445, 2008. doi: 10.1364/OE.16.021434.

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. doi: 10.1109/TBME.2016.2609282.

Long Xi, Weixuan Chen, Changchen Zhao, Xue Wu, and Jian Wang. Image enhancement for remote photoplethysmography in a low-light environment. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 1–7. IEEE, 2020.

Yiping Xie, Bo Zhao, Mingtong Dai, Jian-Ping Zhou, Yue Sun, Tao Tan, Weicheng Xie, Linlin Shen, and Zitong YU. PhysLLM: Harnessing large language models for cross-modal remote physiological sensing. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=aR43t8OEeW.

Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *British Machine Vision Conference (BMVC)*, pp. 277, 2019.

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

Ziyi Zhao, Kailing Guo, Fang Liu, Xiaofen Xing, Lin Wang, Xiangmin Xu, and Zhanpeng Jin. rppg-tfcl: Time-frequency consistency learning for robust remote physiological measurement. *Knowledge-Based Systems*, 330:114607, 2025. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2025.114607. URL https://www.sciencedirect.com/science/article/pii/S0950705125016466.

Bochao Zou, Zizheng Guo, Jiansheng Chen, Junbao Zhuo, Weiran Huang, and Huimin Ma. RhythmFormer: Extracting patterned rPPG signals based on periodic sparse attention. *Pattern Recognition*, 164:111511, 2025a. doi: 10.1016/j.patcog.2025.111511.

Bochao Zou, Zizheng Guo, Xinyu Hu, and Huimin Ma. RhythmMamba: Fast remote physiological measurement with arbitrary length videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025b.

A Proof of Proposition

Proposition 1. *Under the first-order skin reflection model (Chen & McDuff, 2018), with locally constant illumination and color conditions, temporal differencing suppresses stationary skin appearance. The remaining RGB dynamics admit the continuous-time representation*

$$\mathbf{x}(t) = \mathbf{a}_p(t)p'(t) + \mathbf{a}_m(t)m'(t) + \boldsymbol{\epsilon}(t), \tag{14}$$

where $p'(t)$ and $m'(t)$ denote physiological and non-physiological dynamics, respectively; $\mathbf{a}_p(t)$ and $\mathbf{a}_m(t)$ are their effective response coefficients that depend on physiological and imaging conditions; and $\boldsymbol{\epsilon}(t)$ collects measurement and approximation errors. Consequently, the dynamic observation remains a mixture whose source contributions and temporal patterns can vary over time.

Proof. Following the dichromatic reflection model (Chen & McDuff, 2018), substituting the diffuse reflection, specular reflection, and illumination components gives

$$C_k(t) = I_0\left(1 + \Psi(m(t), p(t))\right)\Big(u_c c_0 + u_s\Phi(m(t), p(t)) + u_p p(t)\Big) + v_n(t). \tag{15}$$

Neglecting higher-order products among the small time-varying terms, the retained smooth observation is

$$C_k(t) \approx u_c c_0 I_0 + I_0 u_p p(t) + u_c c_0 I_0 \Psi(m(t), p(t)) + I_0 u_s \Phi(m(t), p(t)) + v_n(t). \tag{16}$$

Over a short interval, I_0, c_0, u_c, u_s, and u_p are treated as locally constant. Temporal differentiation therefore removes the stationary term $u_c c_0 I_0$. Retaining the first-order dynamic terms gives

$$\frac{dC_k(t)}{dt} \approx I_0 u_p p'(t) + u_c c_0 I_0 \frac{d\Psi(m(t), p(t))}{dt} + I_0 u_s \frac{d\Phi(m(t), p(t))}{dt} + \frac{dv_n(t)}{dt}. \tag{17}$$

Here, temporal derivatives describe the continuous-time approximation to frame differences. Applying the chain rule to Ψ and Φ yields

$$\frac{dC_k(t)}{dt} \approx I_0 u_p p'(t) + u_c c_0 I_0 \left(\frac{\partial\Psi}{\partial p}p'(t) + \frac{\partial\Psi}{\partial m}m'(t)\right) + I_0 u_s \left(\frac{\partial\Phi}{\partial p}p'(t) + \frac{\partial\Phi}{\partial m}m'(t)\right) + \frac{dv_n(t)}{dt}, \tag{18}$$

where all partial derivatives are evaluated at $(m(t), p(t))$. Let $\mathbf{x}(t) = \frac{dC_k(t)}{dt}$. Grouping the terms associated with $p'(t)$ and $m'(t)$, we define

$$\mathbf{a}_p(t) = I_0\left(u_p + u_c c_0 \frac{\partial\Psi}{\partial p} + u_s \frac{\partial\Phi}{\partial p}\right), \mathbf{a}_m(t) = I_0\left(u_c c_0 \frac{\partial\Psi}{\partial m} + u_s \frac{\partial\Phi}{\partial m}\right). \tag{19}$$

Collecting the dynamic noise contribution $dv_n(t)/dt$ and the residual of the first-order temporal approximation into $\boldsymbol{\epsilon}(t)$ gives

$$\mathbf{x}(t) = \mathbf{a}_p(t)p'(t) + \mathbf{a}_m(t)m'(t) + \boldsymbol{\epsilon}(t), \tag{20}$$

which establishes Eq. equation 14. Although the stationary reflection is removed, physiological and non-physiological variations remain in the dynamic observation. Their effective coefficients depend on $(m(t), p(t))$ through the derivatives of Φ and Ψ, and may therefore vary with the physiological and imaging conditions. The resulting representation is thus a mixture with potentially time-varying source contributions. □

B Experimental Details

We conduct experiments on six public rPPG datasets: **PURE** Stricker et al. (2014), **UBFC-rPPG** Bobbia et al. (2019), **BUAA** Xi et al. (2020), **MMPD** Tang et al. (2023), **VIPL-HR** (Niu et al., 2019),and **MR-NIRP-Car** Nowara et al. (2020). These datasets cover a diverse range of rPPG acquisition scenarios, from relatively controlled indoor settings to more complex real-world environments, with substantial variations in illumination conditions, head motion, imaging devices, and application settings.

PURE Stricker et al. (2014). PURE is a widely used benchmark for remote pulse measurement under motion. It contains recordings from 10 subjects, each performing six tasks: sitting still,

talking, slow head translation, fast head translation, small head rotation, and moderate head rotation. The videos were recorded at 30 fps with a spatial resolution of 640×480, together with contact PPG signals as physiological reference. The original dataset contains approximately 60 one-minute videos, although one recording is unavailable in the public release and some studies therefore use 59 valid sequences. PURE mainly evaluates the influence of different head motions under relatively controlled conditions. For intra-dataset evaluation, we use the first 60% of the dataset for training and the remaining 40% for testing.

UBFC-rPPG Bobbia et al. (2019). UBFC-rPPG contains 42 videos from 42 subjects. The videos were captured using a Logitech C920 HD Pro camera at 30 fps and a resolution of 640×480. During recording, subjects performed a time-sensitive mathematical task while facing the camera, inducing natural heart-rate variations. Since motion and environmental variations are relatively limited, UBFC-rPPG is commonly used as a benchmark under controlled acquisition conditions. For intra-dataset evaluation, the first 30 subjects are used for training and the remaining 12 subjects for testing.

BUAA Xi et al. (2020). BUAA-MIHR was designed primarily to investigate remote physiological measurement under different illumination intensities. It contains 165 videos from 15 subjects. Recordings were collected in a dark room isolated from external ambient light, while the LED illumination intensity was varied from normal illumination to low-light conditions. Videos were recorded at 30 fps and a resolution of 640×480, with reference PPG signals acquired using a CMS50E device. Following prior experimental settings, we remove extremely dark sequences with illumination levels below 10 lux and retain only samples recorded at illumination levels of at least 10 lux. For intra-dataset evaluation, the dataset is divided sequentially, with the first 70% used for training and the remaining 30% used for testing.

MMPD Tang et al. (2023). MMPD is a multi-domain rPPG dataset designed for challenging mobile scenarios. It contains recordings from 33 subjects and approximately 11 hours of smartphone videos. Compared with PURE and UBFC-rPPG, MMPD exhibits substantially greater diversity in skin tone, illumination, and motion. It includes Fitzpatrick skin types III–VI, four illumination conditions, and activities including stationary recording, head rotation, talking, and walking. These variations make MMPD particularly suitable for evaluating robustness to complex interference and distribution shifts. For intra-dataset evaluation, the subjects are sequentially divided into training, validation, and testing sets with a ratio of $7 : 1 : 2$.

MR-NIRP-Car Nowara et al. (2020). MR-NIRP-Car targets contact-free physiological measurement in real-world driving environments. It contains recordings from 18 subjects and synchronously provides broad-spectrum RGB videos, narrow-band NIR videos, and reference signals from a fingertip pulse oximeter. Data were collected both while vehicles were parked in a garage and while driving on urban roads, with conditions including stationary behavior and small head movements. Vehicle acceleration, deceleration, turning, changing ambient illumination, and camera vibration introduce substantial disturbances even when subjects are instructed to remain relatively still. MR-NIRP-Car therefore provides a realistic benchmark for evaluating rPPG measurement under practical driving conditions. For intra-dataset evaluation, the dataset is divided sequentially into training, validation, and testing sets. The first 70% is used for training, the following 10% for validation, and the remaining 20% for testing.

VIPL-HR. Niu et al. (2019) VIPL-HR is a large-scale, less-constrained rPPG dataset containing recordings from 107 subjects under diverse illumination, motion, and acquisition-device conditions. In this work, VIPL-HR is used only as a source-domain dataset in the supplementary cross-dataset experiment and is not included in the main intra-dataset evaluation. All available VIPL-HR samples are used for source-domain training, and the trained model is evaluated directly on the target dataset without using any target-domain samples for training.

C Additional Experimental Results

C.1 Cross-Dataset Generalization with Two Source Domains

We first adopt a two-source training and single-target testing protocol to examine the sensitivity of PhysWave to different source-domain combinations. MMPD and BUAA are separately treated as

Table 7: **Cross-dataset generalization evaluation results on MMPD under different two-source training settings**. U=UBFC-rPPG, P=PURE, B=BUAA, and M=MMPD. Bold: best results.

Method	P+U→M			P+B→M			U+B→M			Average		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
GREEN Verkruysse et al. (2008)	21.68	27.69	-0.01	21.68	27.69	-0.01	21.68	27.69	-0.01	21.68	27.69	-0.01
CHROM de Haan & Jeanne (2013)	13.66	18.76	0.08	13.66	18.76	0.08	13.66	18.76	0.08	13.66	18.76	0.08
PhysNet Yu et al. (2019)	11.00	17.30	0.28	13.20	16.70	0.23	13.50	17.00	0.09	12.57	17.00	0.20
PhysFormer Yu et al. (2022)	11.40	17.50	0.23	13.90	18.60	0.21	13.20	16.50	0.12	12.83	17.53	0.19
EfficientPhys Liu et al. (2023a)	11.80	18.90	0.22	11.90	18.50	0.21	15.50	20.80	0.03	13.07	19.40	0.15
RhythmFormer Zou et al. (2025a)	10.50	16.72	0.28	13.98	19.46	0.12	12.57	17.45	0.15	12.35	17.88	0.18
PhysLLM Xie et al. (2026)	9.95	14.96	/	11.90	15.30	/	12.10	15.20	/	11.32	15.15	/
PhysWave (Ours)	**7.03**	**12.98**	**0.58**	**6.97**	**12.47**	**0.58**	**7.90**	**13.56**	**0.48**	**7.30**	**13.00**	**0.55**

Table 8: **Cross-dataset generalization evaluation results on BUAA under different two-source training settings**. U=UBFC-rPPG, P=PURE, B=BUAA, and M=MMPD. Bold: best results.

Method	P+U→B			P+M→B			U+M→B			Average		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
GREEN Verkruysse et al. (2008)	6.89	10.39	0.60	6.89	10.39	0.60	6.89	10.39	0.60	6.89	10.39	0.60
CHROM de Haan & Jeanne (2013)	6.09	8.29	0.51	6.09	8.29	0.51	6.09	8.29	0.51	6.09	8.29	0.51
PhysNet Yu et al. (2019)	15.34	21.48	-0.29	20.97	24.75	0.01	11.40	16.72	0.14	15.90	20.98	-0.05
PhysFormer Yu et al. (2022)	18.23	22.17	0.07	14.86	18.26	0.03	10.87	16.20	0.08	14.64	18.80	0.06
EfficientPhys Liu et al. (2023a)	4.60	8.06	0.72	4.15	7.14	0.77	3.00	5.18	0.89	3.92	6.79	0.79
RhythmFormer Zou et al. (2025a)	3.90	6.51	0.82	4.32	6.70	0.82	6.20	11.23	0.49	4.81	8.15	0.71
PhysWave (Ours)	**2.07**	**2.57**	**0.98**	**2.43**	**3.15**	**0.97**	**2.07**	**2.54**	**0.98**	**2.19**	**2.75**	**0.98**

unseen target domains, while two datasets from the remaining domains are jointly used for training. Unlike single-source transfer, this setting introduces heterogeneous training distributions and therefore tests whether the model remains stable when the composition of source domains changes.

Generalization to MMPD. MMPD serves as a challenging target with substantial variations in motion, illumination, skin tone, facial occlusion, and acquisition settings. As shown in Tab. 7, PhysWave achieves the lowest MAE and RMSE, together with the highest r, under all three source combinations: P+U→M, P+B→M, and U+B→M. while also obtaining the highest correlation coefficients. This indicates that PhysWave is less dependent on a particular source-domain combination and can better accommodate heterogeneous training distributions when transferring to a complex target domain.

Generalization to BUAA. Unlike MMPD, BUAA provides a different transfer challenge, where the domain variation is mainly associated with large changes in illumination intensity. As shown in Tab. 8, PhysWave consistently achieves the best results under P+U→B, P+M→B, and U+M→B. The stable performance across these source combinations indicates that PhysWave can effectively suppress illumination-induced interference and preserve reliable physiological information, leading to strong performance under the diverse lighting conditions of BUAA.

Overall, the larger errors on MMPD than on BUAA also reflect different levels of target-domain difficulty. Nevertheless, PhysWave remains consistently effective across source combinations for both target domains, demonstrating low sensitivity to source selection under heterogeneous multi-source training.

C.2 Cross-Dataset Generalization with Three Source Domains

We further adopt a three-source training and single-target testing protocol and perform leave-one-domain-out evaluation across PURE, UBFC-rPPG, BUAA, and MMPD. Compared with the two-source setting, introducing an additional source domain provides richer training variations but also increases distribution heterogeneity, making it more difficult to learn shared physiological patterns across domains.

As shown in Tab. 9, PhysWave achieves the best MAE, RMSE, and r on three of the four target domains, namely UBFC-rPPG, BUAA, and MMPD. The results also show that increasing the number of source domains does not eliminate the differences in transfer difficulty across target datasets, indicating that simply aggregating more heterogeneous training data is not sufficient for reliable domain generalization.

Table 9: **Cross-dataset generalization evaluation results on the PURE/BUAA/UBFC-rPPG/MMPD datasets under three-source training and single-target testing**. U=UBFC-rPPG, P=PURE, B=BUAA, and M=MMPD. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	P+B+M→U			U+B+M→P			P+U+M→B			P+U+B→M		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
GREEN Verkruysse et al. (2008)	19.73	31.00	0.37	10.09	23.85	0.34	6.89	10.39	0.60	21.68	27.69	-0.01
CHROM de Haan & Jeanne (2013)	7.23	8.92	0.51	9.79	**12.76**	0.37	6.09	8.29	0.51	13.66	18.76	0.08
PhysNet Yu et al. (2019)	13.83	23.66	0.35	33.23	35.25	-0.15	12.75	16.37	0.08	13.37	16.64	0.29
PhysFormer Yu et al. (2022)	10.29	18.13	0.60	19.75	24.30	0.24	22.09	26.21	0.03	13.90	19.30	0.06
EfficientPhys Liu et al. (2023a)	12.87	18.80	0.19	7.15	15.04	0.23	32.30	34.00	-0.03	12.87	18.80	0.19
RhythmFormer Zou et al. (2025a)	14.70	22.49	0.43	21.11	25.76	0.04	6.04	10.84	0.42	16.14	20.50	-0.11
PhysWave (Ours)	**0.79**	**1.45**	**0.99**	**5.63**	14.32	**0.82**	**1.95**	**2.43**	**0.98**	**6.84**	**12.34**	**0.60**

Table 10: **Ablation of orthogonal initialization on MMPD and UBFC-rPPG**.

Orthogonal	MMPD			UBFC-rPPG		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
✓	**2.81**	**6.33**	**0.88**	**0.50**	**0.71**	**0.99**
×	3.00	7.16	0.84	0.63	0.90	0.99

A comparison between the two-source and three-source settings reveals that adding more source domains can provide complementary information, but does not always lead to consistent improvements. For MMPD and BUAA, PhysWave benefits from the additional source domain, with MAE/RMSE/r improving from $7.30/13.00/0.55$ to $6.84/12.34/0.60$ on MMPD and MAE/RMSE decreasing from $2.19/2.75$ to $1.95/2.43$ on BUAA. However, the weaker performance under U+B+M→P indicates that additional source domains may also introduce distribution conflicts and source-specific interference. These results suggest that multi-source diversity is beneficial only when its complementary information can be effectively exploited.

C.3 Effect of Orthogonal Initialization.

Finally, we examine the effect of orthogonal initialization in the intra-component interaction. As shown in Tab. 10, removing orthogonal initialization consistently degrades performance on both MMPD and UBFC-rPPG. This indicates that orthogonally initialized projections help reduce redundant correlations within each component and encourage more diverse feature subspaces, leading to more effective intra-component modeling.

C.4 Computational Cost Analysis

We evaluate computational efficiency in terms of parameters, MACs, and peak GPU memory. All methods are evaluated at a spatial resolution of 128×128 on a single NVIDIA GeForce RTX 4090 GPU (24 GB). Peak memory is measured during inference, while parameters and MACs are computed using THOP. Since different rPPG methods adopt different temporal and channel input formats, we retain their original input configurations.

As shown in Tab. 11, PhysWave contains only 0.631M parameters, the smallest among all compared methods. It also requires 29.590G MACs, achieving the second-lowest computational cost, while its peak memory usage is only 530 MiB. Compared with RhythmMamba, PhysWave reduces the parameter count by 40.8% while maintaining a comparable memory footprint, although RhythmMamba has lower MACs.

Overall, PhysWave achieves a favorable balance among model size, computational cost, and memory consumption. This indicates that the proposed multi-scale component modeling improves representational capacity without relying on a large parameter budget, while keeping the overall computational overhead moderate.

Table 11: **Computational cost comparison with representative rPPG methods.** The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Params (M)↓	MACs (G)↓	Memory (MB)↓
DeepPhys (Chen & McDuff, 2018)	1.975	119.112	2572
PhysNet (Yu et al., 2019)	0.769	70.119	**485**
EfficientPhys (Liu et al., 2023a)	1.909	59.796	1700
RhythmFormer (Zou et al., 2025a)	3.251	38.641	611
RhythmMamba (Zou et al., 2025b)	1.065	**12.943**	477
PhysWave (Ours)	**0.631**	29.590	530

D Additional Visualization Analysis

To complement the visualization results in the main paper, we provide additional analyses of the internal behavior of PhysWave. Specifically, we visualize the spatial responses of different components, their temporal attention patterns, and the inter-component attention relationships. These results provide further evidence that the proposed multi-component modeling does not merely generate redundant representations, but instead captures complementary physiological cues from different temporal resolutions.

D.1 Intra-Component Spatial Attention

Fig. 6 presents the spatial attention maps of the stem feature and different wavelet components. We observe that different components exhibit distinct spatial attention patterns rather than consistently focusing on identical facial regions. For example, the high-response regions vary across Stem, D1, D2, D3, and A3, covering different facial regions and, in some cases, areas near the facial boundary. Moreover, the spatial distributions of the same component also vary across different input samples. These observations show that the spatial attention patterns are component-dependent and sample-dependent, rather than being shared uniformly across all decomposed components.

D.2 Intra-Component Temporal Attention

Fig 7 visualizes the refined temporal attention within the stem feature and different wavelet components after the preceding region-level attention. The vertical and horizontal axes denote query time and key time, respectively. Near-diagonal responses indicate attention to temporally nearby positions, while vertical structures indicate that multiple query positions repeatedly attend to the same key time.

For the PURE samples in (a) and (b), Stem and D1 exhibit relatively localized and diagonal responses, whereas D2, D3, and A3 show more pronounced vertical structures at selected temporal positions. In contrast, the MMPD samples in (c) and (d) present more heterogeneous patterns, with localized responses, scattered off-diagonal activations, and vertical structures appearing differently across components. These results show that the refined temporal attention varies across both components and input samples rather than following a single shared temporal pattern.

D.3 Inter-Component Attention Analysis

Fig. 8visualizes the attention weights between different components. Since a self-component bias is explicitly introduced in the interaction module, larger diagonal responses are expected by design. We therefore focus on how the remaining attention weights are distributed across components and how these distributions vary across input samples.

For the PURE samples in (a) and (b), the weights are relatively more distributed across components, with noticeable off-diagonal responses. In contrast, the MMPD samples in (c) and (d) exhibit more concentrated attention on the corresponding components and weaker cross-component responses. Although all samples share the same self-bias, their attention distributions remain different, showing that the inter-component interaction varies with the input.

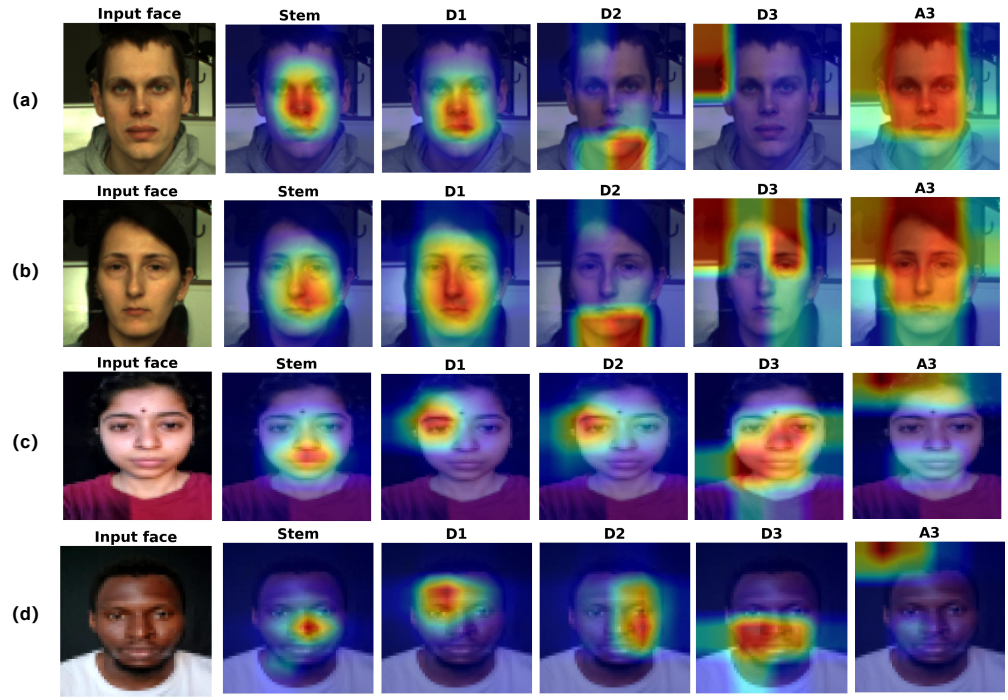


Figure 6: **Spatial attention maps of the stem feature and different wavelet components.** Different components exhibit distinct spatial attention patterns across input samples. Samples (a) and (b) are from the PURE dataset, while (c) and (d) are from the MMPD dataset.

E Detailed Module Illustration

To provide a clearer view of the proposed architecture, we present detailed workflow illustrations of the intra-component and inter-component modules used in PhysWave. Specifically, Fig. 9 shows the detailed pipelines of Sparse Orthogonal Component Attention (SOCA) and Temporal-Spectral Component Interaction (TSCI), which correspond to the intra-component and inter-component modeling processes, respectively.

E.1 Sparse Orthogonal Component Attention (SOCA)

As shown in Fig. 9(a), SOCA independently models the spatio-temporal dependencies within the five components, Stem, D1, D2, D3, and A3. For each component, temporal central difference convolution is first employed to generate the query Q and key K, enhancing sensitivity to local temporal variations, while the value V is produced by a standard $1 \times 1 \times 1$ convolution to preserve the content information used for feature aggregation. Subsequently, Q and K are projected into a shared orthogonal feature space and modulated using different head-wise scaling parameters. The orthogonal features are then fused with the original Q and K through learnable residual coefficients, preserving the original representations while enhancing discrimination across different channel directions. Finally, the model selects the Top-K correlated spatio-temporal regions according to region-level correlations and performs fine-grained sparse attention within the selected regions. This mechanism reduces the computational cost of global attention while focusing the modeling on local periodic variations associated with the pulse signal within each component.

E.2 Temporal-Spectral Component Interaction (TSCI)

As shown in Fig. 9(b), TSCI models the interactions among these five components. For each component, spatial average pooling is first applied to obtain a multi-channel temporal sequence, followed

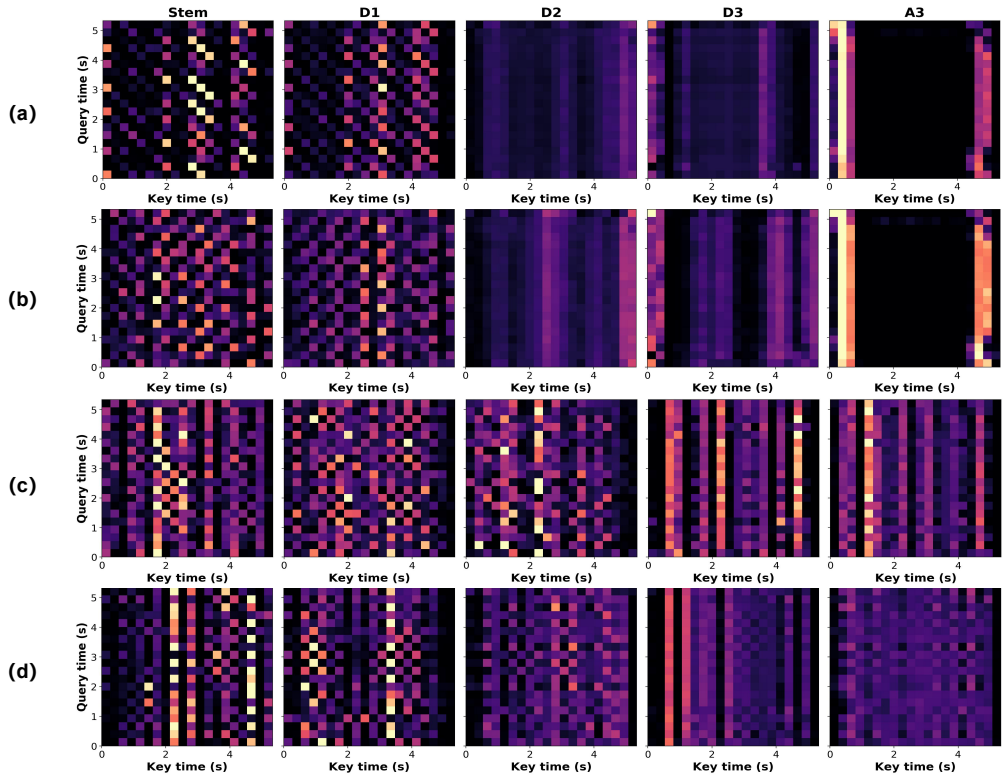


Figure 7: **Refined temporal attention maps of the stem feature and different wavelet components.** Near-diagonal responses indicate attention to temporally nearby positions, while vertical structures indicate that multiple query positions attend to the same key time. The attention patterns vary across components and input samples. Samples (a) and (b) are from the PURE dataset, while (c) and (d) are from the MMPD dataset.

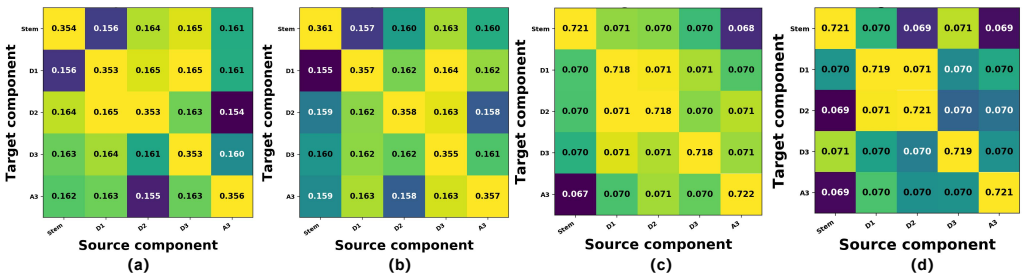
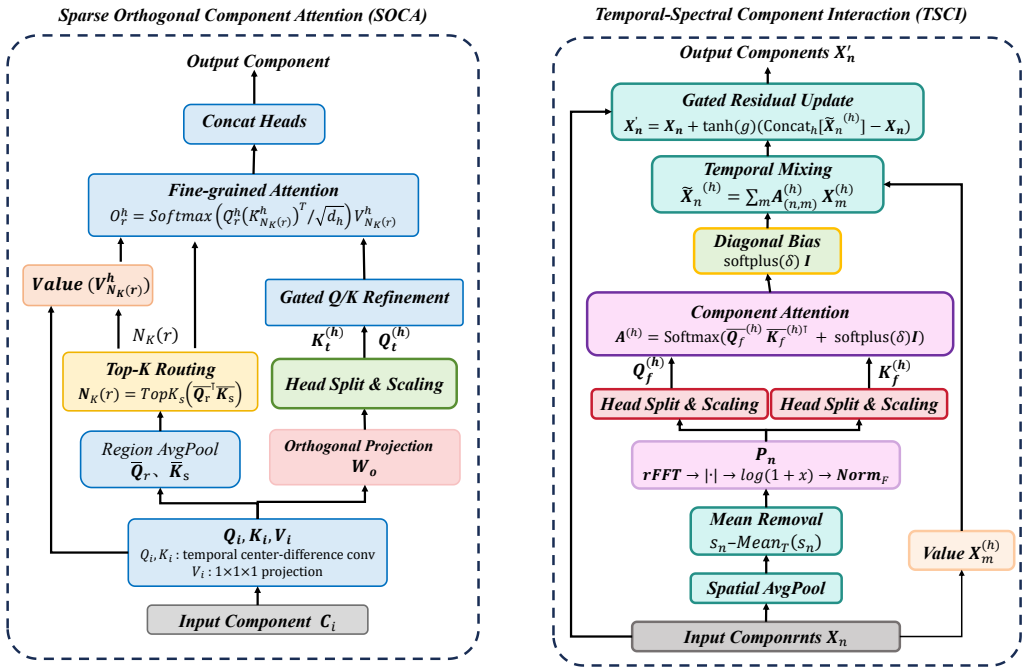


Figure 8: **Inter-component attention matrices.** A self-component bias is explicitly introduced in the interaction module, resulting in stronger diagonal responses. The distributions of the remaining cross-component weights vary across input samples. Samples (a) and (b) are from the PURE dataset, while (c) and (d) are from the MMPD dataset.

by temporal mean removal to suppress static components. The resulting sequence is then transformed into the frequency domain along the temporal dimension using the real-valued fast Fourier transform (rFFT). After logarithmic compression and mean normalization, the spectral representations are split into multiple attention heads and modulated by learnable head-wise scaling parameters to construct the queries and keys. TSCI is subsequently computed according to the similarity be-



(a) (b)

Figure 9: **Detailed architectures of the proposed component modeling modules**. (a) SOCA performs intra-component modeling through region-level routing, orthogonal projection, and fine-grained attention. (b) TSCI performs inter-component interaction through spectral representation, multi-head scaled Q/K construction, component attention, temporal mixing, and gated residual update.

tween the normalized spectral representations, enabling information exchange among components with related spectral responses.

To avoid excessive cross-component mixing that may weaken the original characteristics of individual components, a learnable diagonal bias is introduced into the attention matrix to encourage self-component retention. The resulting attention weights are then applied to the original spatio-temporal features for weighted aggregation, and the interacted features are returned through a residual connection. In the current implementation, this module is activated only in the last Transformer block of the final network stage, allowing adaptive interaction among different wavelet components at a high-level feature representation while limiting the additional computational cost.