Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TPA-Flow: Physiological Tensor Interactions and Rectified-Flow Decoding for Remote Photoplethysmography

Huixiang Wen

huixiangwen@bbmu.edu.cn

Bengbu Medical University

Ruixue Chen

Bengbu Medical University

Luo Zhou

Hong Kong University of Science and Technology

Shunfeng Gao

The Ninth People's Hospital Affiliated to Soochow University

Research Article

Keywords: remote photoplethysmography, physiological monitoring, tensor-product attention, Rectified Flow, heart-rate estimation, waveform reconstruction

Posted Date: October 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11099271/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

TPA-Flow: Physiological Tensor Interactions and Rectified-Flow Decoding for Remote Photoplethysmography

Anonymous Author(s)[1*]

[1*]Affiliation withheld for double-blind review.

Abstract

Remote photoplethysmography (rPPG) enables contact-free physiological monitoring with commodity cameras, providing an enabling technology for telemedicine and intelligent health systems. Existing approaches remain limited by weak physiological periodicity priors, inadequate modeling of multidimensional spatiotemporal interactions, and waveform distortion introduced by one-step deterministic decoding. We propose TPA-Flow, a unified framework that combines quality-aware physiological representation, Physiological Tensor-Product Attention (TPA), and direct-anchored Rectified Flow decoding. TPA injects a confidence-weighted periodic prior into attention and uses low-rank tensor products to capture high-order interactions across facial regions, cardiac cycles, and feature subspaces. The Direct branch first produces a stable waveform anchor. Conditioned on this anchor and a consensus of physiological candidates, the Flow branch learns a short continuous transport path that corrects residual phase and local morphological errors. The two paths are fused to preserve both periodic stability and waveform fidelity. On PURE, UBFC-rPPG, and MMPD, TPA-Flow achieves MAE/RMSE values of 0.04/0.05, 0.47/0.62, and 4.13/7.01 bpm, respectively. It also obtains the lowest RMSE in three of four zero-shot cross-dataset settings with directly comparable baselines. These results demonstrate strong accuracy, robustness, and cross-domain generalization.

Keywords: remote photoplethysmography, physiological monitoring, tensor-product attention, Rectified Flow, heart-rate estimation, waveform reconstruction

1 Introduction

Remote photoplethysmography (rPPG) recovers pulse signals from ordinary camera videos by analyzing subtle variations in skin reflectance caused by cardiac activity. It therefore requires neither electrodes nor optical sensors attached to the body Verkruysse et al. (2008); Wang et al. (2017). This contact-free, low-cost, and readily deployable technology is increasingly used for home health monitoring, telemedicine, driver-state sensing, human–computer interaction, and exercise assessment Wang et al. (2024); Gong et al. (2024). However, pulse-induced color variations are usually much weaker than head motion, illumination fluctuations, specular reflections, and compression noise. Reliable reconstruction of accurate and physiologically meaningful rPPG waveforms from unconstrained videos thus remains an important problem in intelligent health sensing and ubiquitous computing.

Early rPPG methods relied on skin-reflection models and extracted periodic pulse components through color-space projection, blind source separation, or spectral analysis Verkruysse et al. (2008); Wang et al. (2017). Their physical interpretation is clear, but they often assume stable illumination, limited motion, and predefined skin-color statistics. Deep learning subsequently enabled physiological representations to be learned directly from video. DeepPhys separates physiological variation from motion through appearance-guided convolutional attention Chen and McDuff (2018). PhysNet uses three-dimensional convolution for end-to-end waveform reconstruction Yu et al. (2019), while PulseGAN improves signal quality through generative waveform modeling Song et al. (2021). EfficientPhys further simplifies video preprocessing and network computation for efficient camera-based measurement Liu et al. (2023). Although convolutional models effectively capture local color changes and short-term motion, their receptive fields remain constrained by network depth and kernel scale. Fixed local operators also adapt poorly to variations in heart rate, frame rate, and sequence length.

Transformers, temporal attention, and state-space models have therefore been introduced to improve long-range temporal modeling. PhysFormer and PhysFormer++ aggregate local and global dynamics through temporal-difference attention and a SlowFast architecture Yu et al. (2022, 2023). Time–frequency learning enhances periodic representations in wavelet or spectral domains Das et al. (2023). RhythmFormer removes redundant temporal interactions through periodic sparse attention Zou et al. (2025), whereas RhythmMamba models long sequences with linear complexity Zou et al. (2025). Contrastive learning, low-light enhancement, occlusion modeling, and multi-color-space fusion have further improved adaptation under weak supervision and challenging environments Sun and Li (2024); Cao et al. (2024); Wang et al. (2024); Liu et al. (2024). Despite this progress, three limitations remain across physiological representation and waveform decoding.

First, physiological priors are still weakly coupled with visual attention. Standard self-attention relates features mainly through content-based dot products. Even when temporal differences, frequency supervision, or periodic sparsity are introduced, cardiac periodicity is usually treated as an auxiliary constraint rather than a confidence-controlled attention relation. Strong periodic interference can therefore be mistaken for a true pulse rhythm.

Second, high-order physiological interactions among time, facial regions, and color channels remain underexplored. Conventional self-attention models pairwise token similarity. Facial blood-volume changes, however, involve synchronized responses across regions, repeated patterns over cardiac cycles, and differential modulation across color channels. Flattening these dimensions into a single sequence captures their joint structure only implicitly and does not directly describe multiplicative coupling among feature subspaces.

Third, stronger encoders have not fully resolved fine-grained waveform decoding. Most end-to-end methods generate rPPG waveforms through one-step deterministic regression. This strategy recovers the dominant rhythm but can attenuate local peaks, descents, and transient phase changes. Motion- or illumination-induced encoding errors also propagate directly to the output. Generative adversarial learning has demonstrated the value of waveform modeling Song et al. (2021), yet a stable and efficient mechanism for refining direct predictions remains lacking.

To address these issues, we propose TPA-Flow, a remote pulse estimation framework that combines Physiological Tensor-Product Attention Zhang et al. (2025) with direct-anchored Rectified Flow decoding Liu et al. (2023). TPA-Flow first partitions each temporally aligned face into an 8×8 ROI grid. It integrates RGB observations, frame differences, and CHROM/POS projections, and applies continuous quality gating according to skin confidence, brightness, motion, and temporal stability. The resulting representation strengthens weak pulse information in reliable facial regions. Confidence-weighted physiological candidates then produce a periodic attention bias. TPA uses low-rank tensor products to model high-order interactions across regions, cardiac cycles, and feature subspaces. Multiscale temporal exchange and local convolution complement long-range rhythmic modeling with short-term waveform details.

The Direct branch generates an initial waveform with a stable dominant period and global phase. The Flow branch constructs a structured source endpoint from the direct estimate and physiological candidates. A conditional velocity field and a few Heun steps then correct residual phase and morphological errors. The two decoding paths are fused to combine periodic stability with waveform fidelity. Training uses negative Pearson correlation and Rectified Flow matching objectives. Direct-anchor learning is followed by flow alignment and joint fine-tuning. During inference, overlapping windows are aggregated to reconstruct the complete waveform, from which the final heart rate is estimated.

We evaluate TPA-Flow within and across PURE, UBFC-rPPG, and MMPD; MMPD spans multiple mobile devices, motion states, and illumination conditions Tang et al. (2023). The method achieves MAE/RMSE values of 0.04/0.05, 0.47/0.62, and 4.13/7.01 bpm on the three datasets, respectively. It also produces the lowest RMSE in three of four zero-shot transfer directions with directly comparable baselines. Bland–Altman analysis, waveform and spectral comparisons, robustness evaluation, and physiological attribution further show that the model preserves meaningful spatiotemporal evidence under changes in device, motion, and illumination.

The main contributions of this work are summarized as follows.

- We develop a quality-aware multi-source physiological representation that combines color, frame differences, and analytical physiological projections. Continuous ROI gating strengthens weak pulse evidence in reliable facial regions.
- We propose period-guided Physiological Tensor-Product Attention. It couples a confidence-controlled cardiac prior with low-rank tensor-product interactions to model long-range cross-region and cross-cycle dependencies together with local waveform characteristics.
- We introduce a direct-anchored Rectified Flow decoder. Conditional transport refines the Direct prediction, and dual-path fusion combines dominant-period stability with high-fidelity waveform reconstruction.

2 Related Work

2.1 Physiologically Informed Spatiotemporal Representation and Periodic Modeling

The primary challenge in rPPG estimation is to isolate weak blood-volume variations from motion, illumination, and compression noise. CHROM suppresses specular-reflection interference through chrominance combinations de Haan and Jeanne (2013), while POS projects RGB trajectories onto a plane orthogonal to the skin-tone direction to enhance pulsatile components Wang et al. (2017). Bobbia *et al.* select skin regions according to local pulsatility and fuse the resulting candidate signals Bobbia et al. (2019). These methods have clear physiological interpretations, but fixed projections and statistical rules do not readily adapt to dynamic pose, local occlusion, or nonuniform illumination.

Deep models learn spatiotemporal physiological representations directly from data. DeepPhys uses appearance attention to guide motion-feature extraction Chen and McDuff (2018); PhysNet models continuous video with three-dimensional convolution Yu et al. (2019); and RhythmNet organizes multi-region color traces into a spatiotemporal representation Niu et al. (2020). EfficientPhys uses temporal differences for lightweight measurement Liu et al. (2023). Cross-resolution constraints and optical-flow alignment improve robustness to subject distance and head motion Li et al. (2023), whereas ND-DeeprPPG separates environmental noise using background information Liu and Yuen (2024). Scalogram-based time–frequency representation Das et al. (2023), RGB–Retinex fusion Liu et al. (2024), frequency-aware self-supervision Yue et al. (2023), and spatiotemporal contrastive learning Sun and Li (2024) further demonstrate the value of multi-source color information and physiological priors.

For long-range dynamics, PhysFormer and PhysFormer++ aggregate multiscale temporal information through temporal-difference attention and a SlowFast architecture, respectively Yu et al. (2022, 2023). RhythmFormer filters redundant interactions through periodic sparse attention Zou et al. (2025), while RhythmMamba models long sequences with linear complexity Zou et al. (2025). Nevertheless, most existing designs rely on pairwise token relations or single-axis state propagation. They do not explicitly represent the high-order coupling among region quality, color response, and synchronization across cardiac cycles. Tensor-product attention offers a compact basis for

such interactions Zhang et al. (2025), but it has not been adapted to the physiological periodicity and spatial reliability requirements of rPPG.

2.2 rPPG Waveform Reconstruction and Generative Decoding

After spatiotemporal features have been extracted, most methods use deterministic regression heads to output an rPPG waveform. PhysNet reconstructs a continuous signal with a convolutional decoder Yu et al. (2019). PhysFormer and PhysFormer++ generate pulse sequences from globally encoded features Yu et al. (2022, 2023). These methods are compact and stable to train, and they recover the dominant period effectively. Their one-step mappings, however, provide no explicit refinement stage. Interference retained in the encoded features can propagate directly to the output and appear as smoothed peaks, phase shifts, or incorrect selection among competing spectral modes.

Generative modeling provides an alternative route to waveform fidelity. PulseGAN takes a CHROM signal as input and combines adversarial learning with temporal and spectral constraints to generate waveforms closer to contact PPG Song et al. (2021). Dual-GAN jointly models blood-volume-pulse prediction and noise while aligning multiple facial ROIs Lu et al. (2021). Adversarial learning, however, depends on a delicate balance between generator and discriminator. ConDiff-rPPG uses conditional diffusion to complete spatiotemporal representations under heterogeneous occlusion and suppresses background noise with low-rank constraints Wang et al. (2024). CodePhys formulates rPPG recovery as a query over a clean PPG codebook, thereby constraining the output with physiological prototypes Chu et al. (2025). Diffusion-based contact-PPG denoising further shows that generative priors can suppress complex motion artifacts Xia et al. (2025). Diffusion models nevertheless require repeated noising and reverse sampling, and they seldom exploit reliable periodic structure already recovered by a direct predictor.

Flow Matching learns a continuous probability path by regressing a conditional velocity field Lipman et al. (2023). Rectified Flow further encourages near-straight transport trajectories and thereby reduces the number of numerical integration steps Liu et al. (2023). Motivated by this property, TPA-Flow constructs a structured source endpoint from the Direct prediction and the consensus of physiological candidates. The Flow branch then concentrates on residual frequency, phase, and local morphological errors. Final fusion retains the periodic stability of direct regression and the waveform-correction capability of continuous transport.

3 Method

Given a facial video of length T, denoted by $\mathcal{V} = \{\mathbf{I}_t\}_{t=1}^{T}$, the objective is to recover a remote photoplethysmographic (rPPG) waveform $\hat{\mathbf{y}} \in \mathbb{R}^T$ that is temporally synchronized with the contact reference $\mathbf{y} \in \mathbb{R}^T$. This problem is challenging because the pulse-induced variations in skin reflectance are substantially weaker than interference caused by head motion, illumination changes, video compression, and inter-subject appearance variation. Rather than accumulating a large set of spectral or heart-rate auxiliary losses, TPA-Flow introduces physiological structure at the representation,

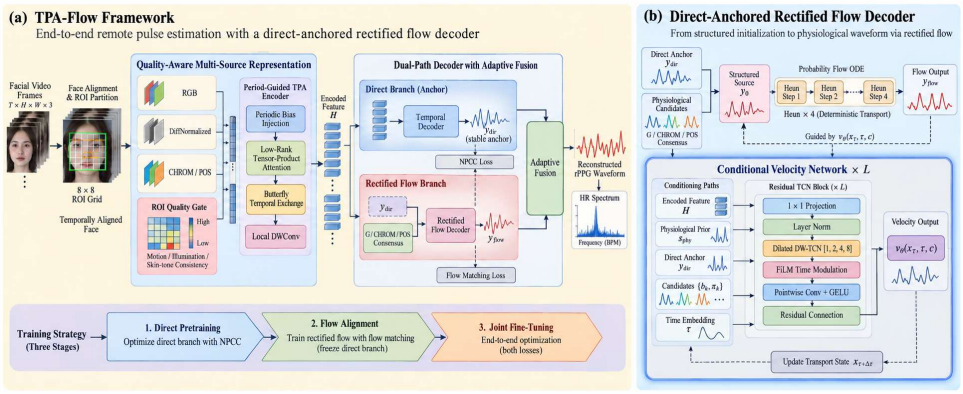


Fig. 1 Overview of the proposed TPA-Flow framework, including physiological encoding, period-guided tensor-product attention, and dual-path decoding. The Direct-anchored Rectified Flow progressively refines the rPPG waveform from a structured source.

encoding, and decoding levels. It first constructs a quality-aware facial representation and multiple physiological candidates, then models short- and long-range pulsatile dependencies using periodicity-guided Tensor-Product Attention (TPA), and finally refines a direct waveform estimate through conditional rectified-flow transport. The final output is

$$\hat{\mathbf{y}} = \mathcal{N}\big[\alpha\hat{\mathbf{y}}^{\mathrm{dir}} + (1-\alpha)\hat{\mathbf{y}}^{\mathrm{flow}}\big], \qquad \mathcal{N}(\mathbf{z}) = \frac{\mathbf{z}-\mu(\mathbf{z})}{\sigma(\mathbf{z})+\varepsilon}, \tag{1}$$

where $\hat{\mathbf{y}}^{\mathrm{dir}}$ and $\hat{\mathbf{y}}^{\mathrm{flow}}$ are the direct and flow-refined reconstructions, respectively, and α controls the contribution of the direct branch. Both waveforms are produced by jointly optimized decoding paths within one network; Eq. (1) is therefore not a test-time ensemble of independently trained models. An overview of the proposed framework is presented in Fig. 1, and its four principal components are detailed in the following subsections.

3.1 Physiology-Aware Spatiotemporal Representation

For each frame, the face is detected and temporally smoothed, resized to 72×72 pixels, and partitioned into an 8×8 regular grid. Let Ω_n denote the nth region of interest (ROI), and let $M_t(p)\in\{0,1\}$ be the skin mask at pixel p. The RGB observation of ROI n in frame t is

$$\mathbf{c}_{t,n} = \frac{\sum_{p\in\Omega_n} M_t(p)\mathbf{I}_t(p)}{\sum_{p\in\Omega_n} M_t(p)+\varepsilon} \in \mathbb{R}^3. \tag{2}$$

When the skin proportion within an ROI is lower than $r_{\min}=0.03$, the unmasked RGB mean of that grid cell replaces Eq. (2), thereby avoiding numerical instability caused by a nearly empty mask. The skin quality is obtained by normalizing the within-ROI skin proportion $r_{t,n}^{\mathrm{skin}}$ by its 85th percentile over the video:

$$q_{t,n}^{\mathrm{skin}} = \mathrm{clip}\left(\frac{\max(r_{t,n}^{\mathrm{skin}}, r_{\min})}{P_{85}(\mathbf{r}^{\mathrm{skin}}) + \varepsilon}, 0.05, 1\right). \tag{3}$$

Abrupt motion and exposure changes are further approximated from the frame-to-frame variation in local luminance $l_{t,n} = \frac{1}{3}\sum_{c\in\{R,G,B\}} c_{t,n,c}$:

$$\begin{aligned} q_{t,n}^{\mathrm{mot}} &= \frac{1}{1 + 10|l_{t,n} - l_{t-1,n}|}, \\ q_{t,n} &= \mathrm{clip}\big(0.65 q_{t,n}^{\mathrm{skin}} + 0.35 q_{t,n}^{\mathrm{mot}}, 0.03, 1\big). \end{aligned} \tag{4}$$

Unlike hard ROI rejection, $q_{t,n}$ acts as a continuous reliability signal for input rescaling, spatial gating, and subsequent temporal attention. The network can consequently redistribute its support among the forehead, nose, and bilateral cheek regions as their reliability changes.

Each ROI is represented by eight channels that jointly preserve raw color, fine-grained temporal variation, and analytical pulse projections. The normalized frame difference is defined as

$$\Delta\mathbf{c}_{t,n} = \frac{\mathbf{c}_{t+1,n} - \mathbf{c}_{t,n}}{\mathbf{c}_{t+1,n} + \mathbf{c}_{t,n} + \varepsilon}, \qquad \widetilde{\Delta\mathbf{c}}_{t,n} = \frac{\Delta\mathbf{c}_{t,n}}{\mathrm{Std}(\Delta\mathcal{C}) + \varepsilon}. \tag{5}$$

Let $Z(\cdot)$ denote standardization along the temporal dimension. Given the ROI color traces r_n, g_n, and b_n, the CHROM and POS projections are

$$\begin{aligned} s_n^{\mathrm{C}} &= Z(3r_n - 2g_n) \\ &\quad - \frac{\sigma(3r_n - 2g_n)}{\sigma(1.5r_n + g_n - 1.5b_n) + \varepsilon} Z(1.5r_n + g_n - 1.5b_n), \\ s_n^{\mathrm{P}} &= Z(g_n - b_n) \\ &\quad + \frac{\sigma(g_n - b_n)}{\sigma(g_n + b_n - 2r_n) + \varepsilon} Z(g_n + b_n - 2r_n). \end{aligned} \tag{6}$$

The model input is therefore

$$\begin{aligned} \mathbf{x}_{t,n} &= \left[\mathbf{c}_{t,n}; \widetilde{\Delta\mathbf{c}}_{t,n}; s_n^{\mathrm{C}}(t); s_n^{\mathrm{P}}(t)\right] \in \mathbb{R}^8, \\ \mathbf{X} &\in \mathbb{R}^{B\times T\times 64\times 8}. \end{aligned} \tag{7}$$

Raw RGB retains the slowly varying color baseline, the DiffNormalized channels emphasize weak inter-frame changes, and CHROM/POS provide analytical projections that are comparatively robust to color scaling. These channels are complementary inputs to one encoder rather than independent estimators.

To prevent a low-quality ROI from bypassing the attention gate through the flattened input, relative quality recalibration is performed before feature projection:

$$\widetilde{q}_{t,n} = \mathrm{clip}\left(\frac{q_{t,n}}{\frac{1}{N}\sum_j q_{t,j} + \varepsilon}, 0.1, 1.5\right), \qquad \widetilde{\mathbf{x}}_{t,n} = \widetilde{q}_{t,n}\mathbf{x}_{t,n}. \tag{8}$$

A two-layer multilayer perceptron (MLP) supplies a learnable spatial gate. By incorporating ROI motion, temporal smoothing, and clip-level stability, its logits and normalized weights are expressed as

$$\begin{aligned} g_{t,n}^{(0)} &= \mathbf{w}_2^\top \,\mathrm{GELU}(\mathbf{W}_1 \widetilde{\mathbf{x}}_{t,n} + \mathbf{b}_1) \\ &\quad + b_2, \\ g_{t,n} &= (1-\lambda_g)\mathcal{S}_{K_g}\left(g_{t,n}^{(0)} - \lambda_m m_n\right) \\ &\quad + \lambda_g \frac{1}{T}\sum_u \mathcal{S}_{K_g}\left(g_{u,n}^{(0)} - \lambda_m m_n\right) \\ &\quad + \log\left(\mathcal{S}_{K_g}(q)_{t,n} + \varepsilon\right), \\ w_{t,n} &= \mathrm{softmax}_n(g_{t,n}), \\ \bar{\mathbf{x}}_t &= \sum_n w_{t,n}\widetilde{\mathbf{x}}_{t,n}, \end{aligned} \tag{9}$$

where $\mathcal{S}_{K_g}$ denotes moving-average smoothing with kernel length K_g, and m_n is the normalized root-mean-square motion magnitude of ROI n over the three DiffNormalized channels. Equation (8) attenuates unreliable regions before network projection, whereas Eq. (9) learns their time-varying spatial contributions.

Three candidate physiological signals are then constructed from the green, CHROM, and POS traces. For $k \in \{G, C, P\}$, each ROI trace is temporally standardized and band-pass filtered over $[f_l, f_h] = [0.66, 3.0]$ Hz, after which ROIs are re-aggregated according to their phase consistency with an initial global signal:

$$\begin{aligned} \mathbf{b}_n^{(k)} &= \mathcal{B}_{[f_l,f_h]}\left(Z(\mathbf{u}_n^{(k)})\right), \\ \rho_n^{(k)} &= \frac{1}{T}\left\langle Z(\mathbf{b}_n^{(k)}), Z\left(\sum_j \bar{w}_j \mathbf{b}_j^{(k)}\right)\right\rangle, \\ \omega_n^{(k)} &= \frac{\bar{w}_n[\mathrm{ReLU}(\rho_n^{(k)}) + \delta_c]}{\sum_j \bar{w}_j[\mathrm{ReLU}(\rho_j^{(k)}) + \delta_c] + \varepsilon}, \\ \mathbf{s}^{(k)} &= Z\left(\sum_n \omega_n^{(k)} \mathbf{b}_n^{(k)}\right), \end{aligned} \tag{10}$$

where $\bar{w}_n$ is the temporally averaged and renormalized gate weight, and δ_c prevents an unstable correlation estimate from entirely suppressing an ROI. This aggregation favors regions that are in phase with the global facial pulse while attenuating locally phase-inverted motion.

The reliability of each candidate is measured by the log-ratio between its in-band and out-of-band energies:

$$
\begin{aligned}
e_k &= \log\left(\sum_{f\in[f_l,f_h]} P_k(f)+\varepsilon\right) \\
&\quad - \log\left(\sum_{f\notin[f_l,f_h]} P_k(f)+\varepsilon\right), \\
\pi_k &= \mathrm{softmax}_k(e_k),
\end{aligned}
\tag{11}
$$

where $P_k(f) = |\mathcal{F}(\mathbf{s}^{(k)})(f)|^2$. The resulting physiological prior is $\mathbf{s}^{\mathrm{phy}} = Z(\sum_k \pi_k \mathbf{s}^{(k)})$. For the long-sequence configuration, the complete in-band probability spectrum of each candidate is retained:

$$
\begin{aligned}
p_k(f) &= \mathrm{softmax}_f\left(\frac{\log(P_k(f)+\varepsilon)}{\tau_s}\right), \\
p(f) &= \sum_k \pi_k p_k(f).
\end{aligned}
\tag{12}
$$

Let $\hat{f} = \arg\max_f p(f)$ and $\hat{f}_k = \arg\max_f p_k(f)$. The confidence of the physiological prior is jointly determined by candidate agreement and spectral concentration:

$$
\begin{aligned}
A_f &= \exp\left[-\frac{\sqrt{\sum_k \pi_k (60\hat{f}_k - 60\hat{f})^2 + \varepsilon}}{15}\right], \\
C_f &= \mathrm{clip}\left(1 - \frac{-\sum_f p(f)\log[p(f)+\varepsilon]}{\log F}, 0, 1\right), \\
c_{\mathrm{phy}} &= \mathrm{clip}\left(C_f \frac{1+A_f}{2}, 0, 1\right).
\end{aligned}
\tag{13}
$$

Here, A_f measures whether the three candidates indicate similar heart rates, whereas C_f measures whether the mixture spectrum contains a well-defined peak. The frequency-band prior is thus a confidence-controlled cue; it does not force the network to interpret every periodic motion as a cardiac rhythm.

3.2 Periodicity-Guided Tensor-Product Attention Encoding

At every time step, the flattened ROI features, the quality-weighted global feature, and four physiological traces are combined into one token:

$$
\begin{aligned}
\mathbf{z}_t &= \left[\mathrm{vec}(\widetilde{\mathbf{X}}_{t,:,:}); \bar{\mathbf{x}}_t; s_t^G; s_t^C; s_t^P; s_t^{\mathrm{phy}}\right], \\
\mathbf{h}_t^{(0)} &= \mathbf{W}_{\mathrm{in}}\mathbf{z}_t + \mathbf{b}_{\mathrm{in}} + \mathbf{p}_t,
\end{aligned}
\tag{14}
$$

where $\mathbf{h}_t^{(0)} \in \mathbb{R}^{128}$ and $\mathbf{p}_t$ is a learnable temporal positional embedding. The flattened representation preserves the identity of individual ROIs, the weighted global feature provides stable color statistics, and the physiological traces supply explicit low-dimensional temporal anchors.

To make attention directly sensitive to plausible cardiac periods, a band-limited autocorrelation is constructed from the full probability spectrum in Eq. (12):

$$\begin{aligned} R(\ell) &= \mathrm{ReLU}\left[\sum_{f\in[f_l,f_h]} p(f)\cos\left(2\pi f\frac{\ell}{f_s}\right)\right], \\ B^{\mathrm{per}}_{t,s} &= \lambda_p c_{\mathrm{phy}} R(|t-s|), \\ B^{\mathrm{per}}_{t,t} &= 0. \end{aligned} \tag{15}$$

This bias strengthens attention between tokens separated by one or more plausible cardiac periods while preserving multiple frequency hypotheses. It therefore avoids averaging two spectral modes into a physiologically unsupported intermediate heart rate. For the short-sequence, single-mode configuration, Eq. (15) reduces to a Gaussian periodic bias centered on the estimated period $P = 1/\hat{f}$:

$$B^{\mathrm{per}}_{t,s} = \lambda_p c_{\mathrm{phy}} \exp\left[-\frac{d_P(|t-s|/f_s)^2}{2(0.12P)^2+\varepsilon}\right], \tag{16}$$

where $d_P(\Delta) = \min(\Delta \bmod P, P - \Delta \bmod P)$.

Before TPA, the lth encoder block exchanges temporal information at different scales through an exclusive-or (XOR) butterfly connection. Let $\pi_l(t) = t \,\mathrm{XOR}\, o_l$, where $o_l \in \{1, 2, 4, 8\}$, and project the partner feature and current context to $\mathbf{a}_t$ and $\mathbf{c}_t$, respectively. The component parallel to the current context is removed from the partner feature:

$$\begin{aligned} \mathbf{a}_t^{\perp} &= \mathbf{a}_t - \left\langle \mathbf{a}_t, \frac{\mathbf{c}_t}{\|\mathbf{c}_t\|_2+\varepsilon} \right\rangle \\ &\quad \times \frac{\mathbf{c}_t}{\|\mathbf{c}_t\|_2+\varepsilon}, \\ \widetilde{\mathbf{h}}_t &= \mathbf{h}_t + \sigma(\gamma_l)\mathbf{W}_o\mathbf{a}_t^{\perp}. \end{aligned} \tag{17}$$

The exponentially increasing butterfly offsets explicitly connect local and distant temporal positions. The orthogonal residual emphasizes information in a partner token that is not already explained by the current context, thereby limiting the direct replication of motion or appearance components.

Global temporal modeling is subsequently performed by TPA. In contrast to a single linear projection that directly produces complete query, key, and value vectors, TPA expresses each attention vector as a context-dependent low-rank tensor product. For the query,

$$\begin{aligned} \mathbf{q}_{t,h} &= \sum_{r=1}^{R} a^{Q}_{t,h,r}\mathbf{b}^{Q}_{t,r}, \\ a^{Q}_{t,h,r} &= [\mathbf{W}_{aQ}\mathbf{h}_t]_{h,r}, \\ \mathbf{b}^{Q}_{t,r} &= [\mathbf{W}_{bQ}\mathbf{h}_t]_{r,:}. \end{aligned} \tag{18}$$

The key and value use the same decomposition:

$$\mathbf{k}_{t,h} = \sum_{r=1}^{R} a^{K}_{t,h,r}\mathbf{b}^{K}_{t,r}, \qquad \mathbf{v}_{t,h} = \sum_{r=1}^{R} a^{V}_{t,h,r}\mathbf{b}^{V}_{t,r}. \tag{19}$$

The coefficients $a_{t,h,r}$ select low-rank components for each temporal position and attention head, whereas $\mathbf{b}_{t,r}$ defines a shared, context-dependent subspace. This multiplicative composition provides a structured representation of weak, multi-source periodic variations. It does not, however, replace standard quadratic attention with a linear-attention approximation.

The logit of the hth attention head is

$$\begin{aligned} L^{(h)}_{t,s} &= \kappa \left\langle \bar{\mathbf{q}}_{t,h}, \bar{\mathbf{k}}_{s,h} \right\rangle + B^{\mathrm{per}}_{t,s} + \log(\bar{q}_s + \varepsilon), \\ A^{(h)}_{t,s} &= \mathrm{softmax}_s(L^{(h)}_{t,s}), \end{aligned} \tag{20}$$

where $\bar{\mathbf{q}}$ and $\bar{\mathbf{k}}$ denote optionally ℓ_2-normalized vectors, and $\bar{q}_s = \frac{1}{N}\sum_n q_{s,n}$ is the mean frame quality. Equation (20) integrates the data-driven TPA similarity, analytical periodic bias, and frame-quality prior within a single attention distribution.

Each encoder block sequentially applies butterfly exchange, pre-normalized TPA, depthwise-separable temporal convolution with kernel size 7, and an MLP with an expansion ratio of 4:

$$\begin{aligned} \mathbf{H}' &= \mathcal{B}_l(\mathbf{H}), \\ \mathbf{H}'' &= \mathbf{H}' + \mathrm{TPA}(\mathrm{LN}(\mathbf{H}'); \mathbf{B}^{\mathrm{per}}, \mathbf{q}), \\ \mathbf{H}''' &= \mathbf{H}'' + \mathrm{PWConv}(\mathrm{DWConv}_7(\mathbf{H}'')), \\ \mathbf{H}^{\mathrm{out}} &= \mathbf{H}''' + \mathrm{MLP}(\mathrm{LN}(\mathbf{H}''')). \end{aligned} \tag{21}$$

Four encoder blocks produce $\mathbf{H} \in \mathbb{R}^{B\times T\times 128}$. Global TPA captures cross-period dependencies, while the local convolution complements it with short-term morphology and local pulse-wave transitions. Together, they form the conditional representation used by the flow decoder.

3.3 Direct-Anchored Rectified-Flow Decoding

The direct branch first produces a stable discriminative waveform anchor from the encoded features:

$$\begin{aligned} \widetilde{y}^{\mathrm{dir}}_t &= \mathbf{w}_2^{\top}\,\mathrm{GELU}(\mathbf{W}_1\,\mathrm{LN}(\mathbf{h}_t) + \mathbf{b}_1) + b_2, \\ \hat{\mathbf{y}}^{\mathrm{dir}} &= \mathcal{N}(\widetilde{\mathbf{y}}^{\mathrm{dir}}). \end{aligned} \tag{22}$$

Instead of requiring rectified flow to generate an entire waveform from isotropic Gaussian noise, TPA-Flow starts from a source endpoint close to a plausible rPPG manifold and learns a structured correction to the direct prediction. Let $\mathbf{d} = \hat{\mathbf{y}}^{\mathrm{dir}}$. For each standardized physiological candidate $\mathbf{b}_k$, its correlation with the direct waveform is $\rho_k = T^{-1}\mathbf{d}^{\top}\mathbf{b}_k$. After polarity alignment, the multi-candidate reliability and consensus prior are

$$
\begin{aligned}
r_k &= \frac{\pi_k(0.25 + 0.75|\rho_k|)}{\sum_j \pi_j(0.25 + 0.75|\rho_j|) + \varepsilon}, \\
\mathbf{b}^{\mathrm{con}} &= \sum_k r_k \,\mathrm{sign}(\rho_k)\mathbf{b}_k.
\end{aligned}
\tag{23}
$$

The unified source endpoint is

$$
\begin{aligned}
\beta &= \beta_{\max} c_{\mathrm{phy}} \left(0.5 + 0.5 \sum_k r_k |\rho_k|\right), \\
\mathbf{y}_0 &= \mathcal{N}\left[(1-\beta)\,\mathrm{sg}(\mathbf{d}) + \beta \mathbf{b}^{\mathrm{con}} + \sigma_0 \boldsymbol{\epsilon}\right], \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),
\end{aligned}
\tag{24}
$$

where sg denotes stop-gradient. If the physiological prior is diffuse or inconsistent with the direct branch, β decreases automatically and the source returns toward the direct prediction. Equations (23)–(24) describe the multi-candidate consensus source. In the configuration that uses only the mixed prior $\mathbf{b} = Z(\mathbf{s}^{\mathrm{phy}})$, we set $\rho = T^{-1}\mathbf{d}^{\top}\mathbf{b}$, $\widetilde{\mathbf{b}} = \mathrm{sign}(\rho)\mathbf{b}$, $\beta = \beta_{\max} c_{\mathrm{phy}} |\rho|$, and $\mathbf{b}^{\mathrm{con}} = \widetilde{\mathbf{b}}$. The direct-noise configuration further sets $\beta = 0$. Small source noise during training improves robustness of local transport. Its magnitude is set to zero during the reported inference procedure, making prediction deterministic.

Given the standardized target $\mathbf{y}_1 = \mathbf{y}$, a straight path between source and target is defined by

$$
\begin{aligned}
\mathbf{y}_\tau &= (1-\tau)\mathbf{y}_0 + \tau \mathbf{y}_1, \\
\tau &\sim \mathcal{U}(0,1), \\
\mathbf{v}_\tau^* &= \frac{\mathrm{d}\mathbf{y}_\tau}{\mathrm{d}\tau} = \mathbf{y}_1 - \mathbf{y}_0.
\end{aligned}
\tag{25}
$$

The conditional velocity network learns

$$
\mathbf{v}_\theta = v_\theta\left(\mathbf{y}_\tau, \tau \mid \mathbf{H}, \mathbf{s}^{\mathrm{phy}}, \mathbf{d}, c_{\mathrm{phy}}, \{\mathbf{b}_k, \pi_k\}\right). \tag{26}
$$

In implementation, the current state, physiological prior, time channel, and optional direct waveform, candidate waveforms, and candidate weights are concatenated with $\mathbf{H}$ along the channel dimension. A 1×1 convolution projects the concatenated feature before it enters a residual temporal convolutional network (TCN). Continuous time τ is encoded by a 64-dimensional sinusoidal embedding and modulates every TCN block through feature-wise linear modulation (FiLM). Dilation rates cycle through $\{1, 2, 4, 8\}$, allowing the velocity field to exploit both local waveform structure and longer temporal context.

After training, the waveform is obtained from $\mathbf{y}(0) = \mathbf{y}_0$ by solving

$$
\frac{\mathrm{d}\mathbf{y}(\tau)}{\mathrm{d}\tau} = v_\theta(\mathbf{y}(\tau), \tau \mid \mathcal{C}). \tag{27}
$$

TPA-Flow uses a four-step Heun solver at inference. For $\Delta\tau = 1/K$,

$$
\begin{aligned}
\mathbf{k}_1 &= v_\theta(\mathbf{y}_i, \tau_i \mid \mathcal{C}), \\
\widetilde{\mathbf{y}}_{i+1} &= \mathbf{y}_i + \Delta\tau \mathbf{k}_1, \\
\mathbf{k}_2 &= v_\theta(\widetilde{\mathbf{y}}_{i+1}, \tau_i + \Delta\tau \mid \mathcal{C}), \\
\mathbf{y}_{i+1} &= \mathbf{y}_i + \frac{\Delta\tau}{2}(\mathbf{k}_1 + \mathbf{k}_2).
\end{aligned}
\tag{28}
$$

The flow output is $\hat{\mathbf{y}}^{\text{flow}} = \mathcal{N}(\mathbf{y}_K)$ and is fused with the direct branch according to Eq. (1). The direct branch supplies a stable anchor, whereas the flow branch models conditional residual transport from that anchor to the reference waveform. This division of labor avoids burdening the flow model with recovery of an entire weak physiological signal from unstructured noise.

3.4 Stage-Wise Optimization and Sequence-Level Inference

The effective training objective of TPA-Flow contains only rectified-flow matching and a negative Pearson correlation coefficient (NPCC) loss on the final waveform. Define the velocity residual as

$$
\mathbf{e} = \mathbf{v}_\theta(\mathbf{y}_\tau, \tau \mid \mathcal{C}) - (\mathbf{y}_1 - \mathbf{y}_0). \tag{29}
$$

Flow matching uses the Huber penalty

$$
\mathcal{L}_{\text{FM}} = \frac{1}{BT} \sum_{b,t} \ell_\delta(e_{b,t}), \qquad \ell_\delta(e) = \begin{cases} \dfrac{e^2}{2\delta}, & |e| < \delta, \\ |e| - \dfrac{\delta}{2}, & |e| \geq \delta. \end{cases} \tag{30}
$$

The Huber form limits extreme gradients induced by difficult clips or local reference-signal artifacts. The time variable is antithetically sampled: each mini-batch pairs τ with $1 - \tau$ to cover both ends of the transport path more uniformly.

The final fused waveform is supervised by

$$
\mathcal{L}_{\text{NPCC}} = \frac{1}{B} \sum_{b=1}^{B} \left[1 - \frac{\langle \hat{\mathbf{y}}_b - \bar{\hat{y}}_b, \mathbf{y}_b - \bar{y}_b \rangle}{\|\hat{\mathbf{y}}_b - \bar{\hat{y}}_b\|_2 \|\mathbf{y}_b - \bar{y}_b\|_2 + \varepsilon} \right]. \tag{31}
$$

Because NPCC is invariant to global offset and linear amplitude scaling, it is well suited to constraining the phase and morphology of a standardized rPPG waveform. The total loss is

$$
\mathcal{L} = \lambda_{\text{FM}} \mathcal{L}_{\text{FM}} + \lambda_{\text{NPCC}} \mathcal{L}_{\text{NPCC}}. \tag{32}
$$

During direct-branch pretraining, $\lambda_{\text{FM}} = 0$, $\lambda_{\text{NPCC}} = 1$, and $\alpha = 1$. Both loss weights are then set to one for flow alignment followed by joint fine-tuning. This schedule first establishes a reliable direct anchor and only then learns local transport, preventing rapid drift of the flow source distribution early in training. A two-step Euler solver is used when evaluating NPCC during training to control computational cost, whereas the four-step Heun scheme in Eq. (28) is used for final inference. Apart from Eqs. (30) and (31), the weights assigned to FFT, heart-rate classification, smoothness, signal-to-noise-ratio, and harmonic losses are all zero in the final configuration.

Table 1 Dataset-Level Instantiations of Unified TPA-Flow

Setting	PURE	UBFC-rPPG	MMPD
Input length T	360	180	360
TPA rank R	4	2	4
Period prior	Multi	Single-HR	Multi
ROI-gate kernel K_g	15	1	21
Gate mixing λ_g	0.70	0	0.80
Motion penalty λ_m	0.25	0	0.40
Flow width/blocks	128/6	128/6	160/8
Flow source	Conf. prior	Noise	Consensus
Direct fusion α	0.85	0.80	0.50
Inference solver	Heun-4	Heun-4	Heun-4

The model is optimized with AdamW using a weight decay of 10^{-4}, gradient norm clipping at 1.0, and automatic mixed precision. Data augmentation comprises horizontal ROI flipping, spatial translation by at most one grid cell, random ROI dropout, per-ROI and per-channel amplitude perturbation, additive feature noise, and synchronized temporal reversal of the input and reference. These transformations vary appearance, cropping, and local signal quality without changing the underlying heart rate. Checkpoints are selected exclusively from complete validation sequences using the normalized score

$$S_{\mathrm{val}} = \sqrt{\left(\frac{\mathrm{MAE}_{\mathrm{val}}}{m^*}\right)^2 + \left(\frac{\mathrm{RMSE}_{\mathrm{val}}}{r^*}\right)^2}. \tag{33}$$

The test set is never used for checkpoint selection, early stopping, or determination of the fusion coefficient. All three datasets use the same TPA-Flow computational framework, with dataset-level hyperparameters chosen on validation data to accommodate differences in sequence length and motion complexity (Table 1).

Accordingly, TPA-Flow should be described as a unified framework with dataset-level instantiations, rather than as either a system with identical hyperparameters for every dataset or an oracle combination of different historical model versions.

For a long video, overlapping clip predictions are returned to their original temporal positions and averaged. Let $\mathcal{K}(t)$ be the set of clips that cover time index t and let s_k denote the starting index of clip k. Then

$$\hat{y}_t^{\mathrm{seq}} = \frac{1}{|\mathcal{K}(t)|} \sum_{k \in \mathcal{K}(t)} \hat{y}_{t-s_k}^{(k)}. \tag{34}$$

The reconstructed sequence is standardized again and multiplied by a Hann window. Heart rate is estimated from the dominant frequency within the physiological band:

$$\widehat{\mathrm{HR}} = 60 \arg \max_{f \in [0.66, 3.0]} \left| \mathcal{F}\left(\mathbf{w}_{\mathrm{Hann}} \odot \hat{\mathbf{y}}^{\mathrm{seq}}\right)(f) \right|^2. \tag{35}$$

Parabolic interpolation over the spectral peak and its two neighboring bins improves frequency resolution. In cross-dataset evaluation, a model is trained only on the source domain and selected using its source-domain validation set. It is then applied directly

Table 2 Dataset Splits and Input Protocols

Dataset	Train subjects	Train sequences	Validation subjects	Validation sequences	Test subjects/ sequences	Clips and window/stride
PURE	5	30	1	5	4/24	649/108/520; 12 s/3 s
UBFC-rPPG	28	28	6	6	8/8	561/120/161; 6 s/3 s
MMPD	23	460	5	100	5/100	4600/1000/ 1000; 12 s/6 s

to the target domain under a zero-shot protocol, without target-domain fine-tuning, calibration, or re-selection of the fusion coefficient.

4 Experiments and Performance Evaluation

4.1 Experimental Setup

4.1.1 Datasets

Experiments were conducted on PURE, UBFC-rPPG, and MMPD. PURE contains 60 nominal recordings from 10 subjects under controlled head-motion protocols. One recording was unavailable in the local copy. We therefore used 59 sequences in the actual split. UBFC-rPPG DATASET 2 contains 42 subject videos with synchronized contact references. MMPD contains 660 mobile videos from 33 subjects. It covers five motion states and four illumination conditions. MMPD has the largest within-domain variation because motion, illumination, and subject appearance change jointly. Table 2 summarizes the actual splits and input protocols.

4.1.2 Implementation Details

All models were trained and evaluated on one NVIDIA GeForce RTX 5090 GPU. The software environment used PyTorch 2.11.0 with CUDA 12.8 and OpenCV 4.13.0.92. Automatic mixed precision was enabled. Face boxes were updated every 10 frames and smoothed with a factor of 0.75. Each face was resized to 72×72 pixels and divided into an 8×8 ROI grid.

Each ROI contains eight input channels. They comprise raw RGB values, normalized frame differences, and CHROM and POS projections. ROIs with a skin proportion below 0.03 were down-weighted. The estimated ROI quality was used for both input recalibration and spatial gating.

The encoder contains four TPA blocks with a hidden dimension of 128. Each block uses four attention heads with 32 dimensions per head. The Direct branch regresses the rPPG waveform. The Rectified Flow branch conditions on the Direct output and physiological candidates. It performs nonlinear refinement with a four-step Heun solver. The network topology and training objectives remain unchanged across the

Table 3 Training and Model Settings

Setting	PURE	UBFC-rPPG	MMPD
Maximum epochs	14	22	46
Batch size	8	16	8
Initial learning rate	2×10^{-5}	3×10^{-5}	2×10^{-4}
TPA rank	4	2	4
Flow width/blocks	128/6	128/6	160/8
Flow solver steps	4	4	4
Direct weight α	0.85	0.80	0.50
Early-stop patience	6	8	12

three datasets. Dataset-level hyperparameters were selected on the corresponding validation set. These settings account for differences in sequence length, sample size, and motion complexity.

AdamW was used with a weight decay of 1×10^{-4}. The gradient norm was clipped at 1.0. Training followed a stage-wise schedule. For PURE and UBFC-rPPG, the long-sequence TPA module was first adapted from a pretrained Direct checkpoint. The Rectified Flow decoder was then optimized. For MMPD, Direct pretraining was followed by Flow alignment and joint fine-tuning.

We report mean absolute error (MAE), root mean square error (RMSE), and Pearson's correlation coefficient r. MAE and RMSE are measured in beats per minute (bpm). The 90th-percentile absolute error is also reported in the decoder-path study. Lower error values and a higher r indicate better performance.

4.2 Heart-Rate Estimation and Agreement Analysis

4.2.1 Overall Performance

Table 4 compares TPA-Flow with representative rPPG methods. On PURE, our method achieves an MAE of 0.04 bpm, an RMSE of 0.05 bpm, and $r = 1.00$. On UBFC-rPPG, it obtains an RMSE of 0.62 bpm and $r = 0.99$. Its MAE is 0.47 bpm. On MMPD, our method yields an MAE of 4.13 bpm and the lowest RMSE of 7.01 bpm. The correlation coefficient is 0.79. These results show that the joint TPA and Rectified Flow design improves estimation accuracy while limiting large errors.

4.2.2 Agreement Analysis

Figure 2 evaluates sequence-level agreement. Panels (a)–(c) compare predicted and reference heart rates. The dashed diagonal denotes ideal agreement. Panels (d)–(f) show Bland–Altman plots. Their horizontal axis is the mean of the predicted and reference heart rates. Their vertical axis is the prediction difference. The solid line denotes the mean bias. The dashed lines show the 95% limits of agreement, defined as the bias plus or minus 1.96 standard deviations.

On PURE, the 24 test points nearly overlap the identity line. The mean bias is $+0.017$ bpm, and the limits of agreement are $[-0.072, +0.106]$ bpm. Quality gating preserves informative color changes from the forehead and cheeks. The periodic TPA

Table 4 Within-Dataset Heart-Rate Estimation Results

Method	PURE MAE	PURE RMSE	PURE r	UBFC-rPPG MAE	UBFC-rPPG RMSE	UBFC-rPPG r	MMPD MAE	MMPD RMSE	MMPD r
DeepPhys	0.83	1.54	0.99	6.27	10.82	0.65	22.27	28.92	-0.03
PhysFormer	1.10	1.75	0.99	0.50	0.71	0.99	11.99	18.41	0.18
Contrast-Phys+	0.48	0.98	0.99	**0.21**	0.80	0.99	–	–	–
DiffPhy	1.46	5.88	0.90	1.05	1.63	0.99	–	–	–
RhythmFormer	0.27	0.47	0.99	0.50	0.78	0.99	4.69	11.31	0.60
RhythmMamba	0.23	0.34	0.99	0.50	0.75	0.99	**3.16**	7.27	**0.84**
STFPNet	0.47	0.67	0.99	0.41	0.95	0.99	–	–	–
PhysMamba	0.25	0.40	0.99	0.54	0.76	0.99	–	–	–
TPA-Flow	**0.04**	**0.05**	**1.00**	0.47	**0.62**	**0.99**	4.13	**7.01**	0.79

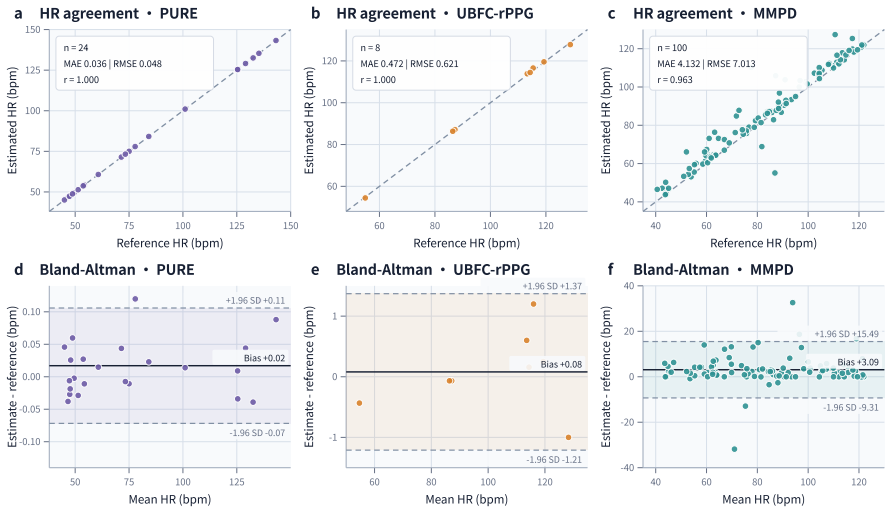


Fig. 2 Heart-rate estimation and agreement analysis on the three within-domain test sets. The upper row compares predicted and reference heart rates. The lower row shows the corresponding Bland–Altman plots. Each point represents one complete test sequence.

bias then tracks the dominant cardiac frequency. Sequence-level overlap averaging further reduces window-level phase jitter.

The UBFC-rPPG points also remain close to the identity line. The mean bias is +0.08 bpm, with limits of $[-1.211, +1.369]$ bpm. Video compression, local highlights, and natural motion introduce small frequency shifts. The model nevertheless preserves the relative heart-rate ordering across subjects. Because the test split contains only eight sequences, r should be interpreted together with MAE and RMSE.

MMPD contains stronger motion, pose, and illumination changes. Its mean bias is +3.092 bpm, and the limits are $[-9.313, +15.486]$ bpm in the supplied agreement plot. Most points remain near the identity line. The wider limits mainly arise from a small set of difficult sequences. Periodic mechanical motion can enter the physiological

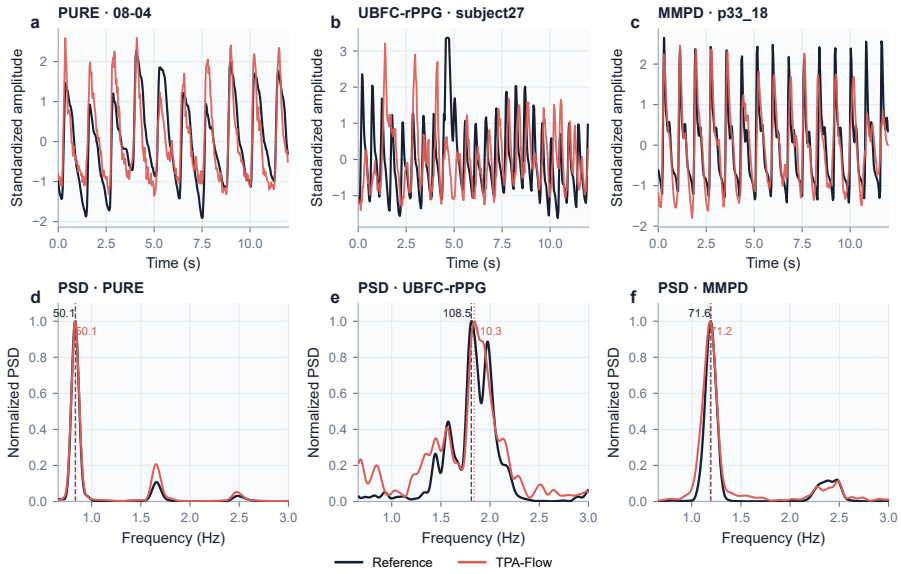


Fig. 3 Waveform reconstruction and spectral consistency for the test sequence with the minimum absolute heart-rate error in each dataset. The upper row shows 12-s standardized waveforms. The lower row shows normalized Welch spectra over the complete sequence.

frequency band in these cases. The model still retains competitive overall performance under this stronger interference.

4.3 Waveform Reconstruction and Spectral Consistency

The waveform analysis uses one low-error example from each test set: PURE 08-04, UBFC-rPPG subject27, and MMPD p33_18. Panels (a)–(c) in Fig. 3 show 12-s standardized reference and predicted waveforms on the original time axis. No phase shift or polarity inversion is applied. Panels (d)–(f) show normalized Welch spectra over 0.66–3.00 Hz.

Sequence-level heart rate is computed with a windowed FFT over the complete reconstructed waveform. The displayed spectrum uses overlapped Welch segments. These estimators have different window lengths and frequency resolutions. Their peak values may therefore differ slightly. They should remain within the same physiological-frequency neighborhood.

For PURE 08-04, the reference and predicted heart rates are 49.515 and 49.513 bpm. The absolute error is 0.002 bpm. Both Welch spectra peak at 50.1 bpm. The principal systolic peaks also occur at similar times. Small differences in peak width and decay may reflect propagation delay, ROI motion, and different sensor bandwidths.

For UBFC-rPPG subject27, the sequence-level error is 0.064 bpm. The reference and predicted Welch peaks are 108.5 and 110.3 bpm. The prediction preserves the rapid dominant rhythm. Local waveform morphology is less consistent because nearby spectral peaks and time-varying phase can reduce pointwise correlation without changing the estimated heart rate.

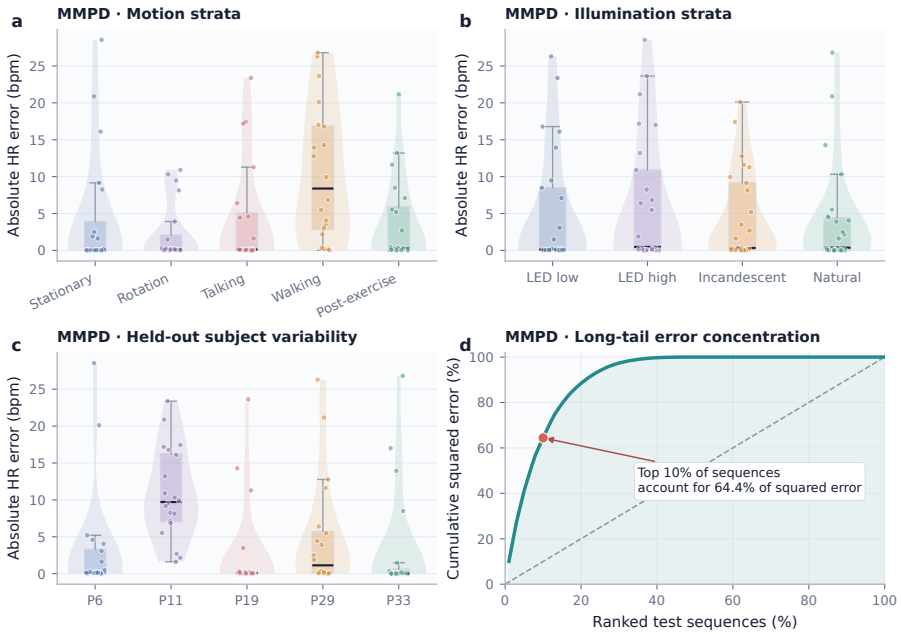


Fig. 4 Robustness and long-tail error analysis on MMPD. The panels show absolute heart-rate errors across motion states, illumination conditions, and held-out subjects. The final panel reports the cumulative squared-error contribution after sorting sequences by error.

For MMPD p33_18, the sequence-level error is 0.003 bpm. The reference and predicted Welch peaks are 71.6 and 71.2 bpm, respectively. The full-sequence waveform correlation is 0.691. Peak timing is consistent in the first half of the segment. Moderate amplitude and phase drift appears later. Overall, TPA-Flow reliably identifies the dominant physiological frequency across the three datasets. The remaining local differences mainly reflect nonstationary pulse morphology and cross-sensor discrepancies.

4.4 Robustness Under Challenging Conditions

We stratify the 100 MMPD test sequences by motion, illumination, and held-out subject. Each motion group contains 20 sequences. Each illumination group contains 25 sequences. Each of the five test subjects contributes 20 sequences. Figures 4(a)–(c) combine violin densities, box statistics, and jittered observations. Figure 4(d) ranks sequences by squared error and reports the cumulative contribution to total squared error.

Walking is the most difficult motion condition. Its MAE and median absolute error are 8.499 and 8.389 bpm. Rotation yields an MAE of 2.280 bpm and a median error of 0.125 bpm. Stationary and talking sequences have median errors of 0.061 and 0.127 bpm. Their MAEs are 3.539 and 3.423 bpm. Thus, most samples in these groups are estimated accurately, while a few large errors increase the mean.

Walking jointly changes face position, apparent scale, occlusion, and body motion. These effects can produce coherent periodic interference across ROIs. TPA strengthens

stable periodic evidence, which is useful in most cases. However, a mechanical rhythm close to the heart-rate band can also become a competing spectral peak. The observed distribution is therefore dominated by a limited number of peak-selection errors rather than uniform degradation.

Among illumination conditions, LED-high has the largest MAE at 6.484 bpm. Natural light gives the lowest MAE at 4.086 bpm. The median error remains below 1 bpm in all four groups. High-intensity LED lighting can produce local saturation and specular reflections on the forehead and cheeks. This effect distorts the approximate linear relation among RGB channels. The results also show that illumination alone does not determine performance. Its effect depends on motion, pose, and individual reflectance.

The subject-level results show similar concentration. P11 has an MAE of 10.855 bpm and a median error of 9.712 bpm. P19 has an MAE of 2.710 bpm and a median error of 0.068 bpm. These differences may reflect pose range, motion amplitude, reflectance, camera distance, and local illumination. The five held-out subjects are insufficient for a causal demographic interpretation. The most difficult 10% of MMPD sequences contribute 64.4% of the total squared error. This confirms that the remaining performance gap is driven mainly by rare large errors.

4.5 Cross-Dataset Generalization and Physiological Attribution

4.5.1 Zero-Shot Cross-Dataset Performance

Each model was trained on one source dataset and directly evaluated on the other datasets. Source checkpoints and fusion weights were selected only with source-domain training and validation data. No target-domain fine-tuning, threshold search, parameter update, or post-hoc calibration was allowed. Evaluation covered 59 available PURE sequences, 42 UBFC-rPPG sequences, and 660 MMPD sequences. The source model retained its original 6-s or 12-s context during inference.

Table 5 reports the four transfer directions for which directly comparable baseline results are available. TPA-Flow obtains the lowest RMSE in three directions. Its RMSE is 1.44 bpm for PURE→UBFC-rPPG, which is 21.3% lower than the next-best result. For PURE→MMPD, it achieves the lowest RMSE of 15.76 bpm and the highest r of 0.39. For UBFC-rPPG→MMPD, the RMSE is 16.24 bpm, a reduction of 5.3% over the next-best value. On UBFC-rPPG→PURE, the method obtains the lowest MAE of 1.12 bpm and a tied-best r of 0.97.

Transfer to PURE and UBFC-rPPG is generally more accurate because these target domains have more stable acquisition conditions. Transfer to MMPD is harder for every method. MMPD combines stronger motion, illumination variation, and device differences. These factors enlarge the source-target distribution gap. Despite this shift, TPA-Flow retains the lowest RMSE in both reported transfers to MMPD. This result indicates improved control of large cross-domain errors.

Table 5 Strict Zero-Shot Cross-Dataset Heart-Rate Estimation Results

Method	PURE→UBFC-rPPG			PURE→MMPD			UBFC-rPPG→PURE			UBFC-rPPG→MMPD		
	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r	MAE	RMSE	r
DeepPhys	1.21	2.90	**0.99**	16.92	24.61	0.05	5.54	18.51	0.66	17.50	25.00	0.05
PhysFormer	1.44	3.77	0.98	14.57	20.71	0.15	12.92	24.36	0.47	12.10	17.79	0.17
RhythmMamba	**0.95**	1.83	**0.99**	10.44	16.70	0.36	1.98	6.51	0.96	**10.63**	17.14	**0.34**
PhysMamba	0.97	1.93	**0.99**	**10.31**	16.02	0.34	1.20	**5.99**	**0.97**	11.96	17.69	0.29
TPA-Flow	0.98	**1.44**	**0.99**	10.56	**15.76**	**0.39**	**1.12**	6.19	**0.97**	11.05	**16.24**	0.26

4.5.2 Physiological Attribution and Waveform Visualization

Figure 5 presents representative cross-dataset cases. The left panel of each case is a physiological Grad-CAM map. The attribution target is the log power at the predicted dominant frequency. Temporal gradient aggregation is applied to the learned 8×8 ROI-gate responses. The resulting map is projected back to the original image using the face box from preprocessing.

A continuous skin-confidence field attenuates responses on hair, eyes, lips, glasses, and the background. This operation only suppresses existing activations. It cannot create a hotspot where the original model response is zero. The right panel shows a corresponding 6-s reference and predicted waveform segment.

Across the six cases, high responses are concentrated on visible skin over the forehead, nose, and bilateral cheeks. The visible-skin concentration ranges from 0.738 to 0.887. Leakage into occluded regions does not exceed 0.078. The UBFC-rPPG and PURE targets show close waveform periodicity. This behavior agrees with their lower cross-domain errors. For the two MMPD targets, local amplitude and morphology vary more strongly. The dominant frequency is still recovered accurately.

These observations support the intended division of labor. ROI gating selects reliable color variations from several physiological regions. TPA models periodic relationships across time and facial regions. The Rectified Flow decoder constrains the reconstructed waveform distribution. Together, these components preserve physiologically meaningful evidence under changes in device, illumination, appearance, and motion.

4.6 Ablation Study and Decoder-Path Complementarity

The path-level ablation uses one complete TPA-Flow checkpoint per dataset. Direct, Rectified Flow, and Fused outputs are read from the same model and evaluated on the same test sequences. The Direct path regresses the waveform from TPA features. The Flow path performs nonlinear refinement with a four-step Heun solver. The Fused path combines both outputs using the validation-selected weight.

The two decoding paths provide complementary estimates. On PURE, Fused reduces MAE and RMSE to 0.04 and 0.05 bpm. It is the best path for 17 of 24 samples. Relative to Flow, the reductions are 69.2% in MAE and 66.7% in RMSE. On UBFC-rPPG, Fused achieves an MAE of 0.47 bpm and an RMSE of 0.62 bpm. Its 90th-percentile error is 0.95 bpm. These values are 45.3%, 44.6%, and 50.8% lower than the corresponding Flow results.

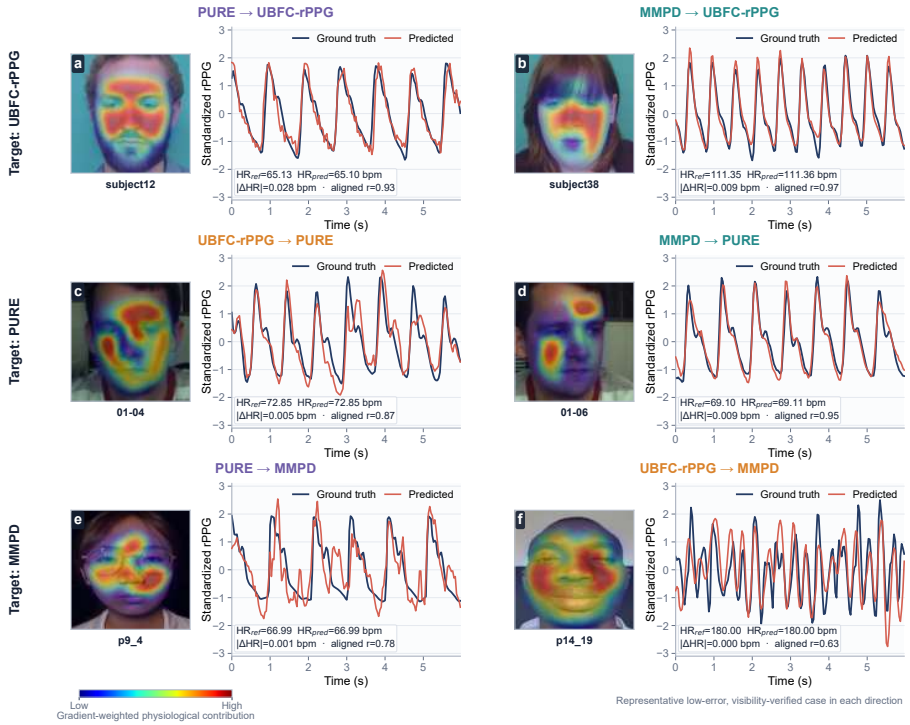


Fig. 5 Cross-dataset physiological attribution and waveform reconstruction. Each row fixes one target domain, and the two columns use models trained on the other source domains. Each case contains a physiological Grad-CAM map and a corresponding 6-s reference and predicted waveform segment.

Table 6 Sequence-Level Performance of the Direct, Flow, and Fused Output Paths

Dataset	Output path	MAE (bpm)	RMSE (bpm)	90th-percentile error (bpm)	Sample-wise wins
PURE	Direct	0.24	0.25	0.19	2
	Flow	0.13	0.15	0.11	5
	Fused	**0.04**	**0.05**	**0.08**	**17**
UBFC-rPPG	Direct	0.87	1.15	2.02	1
	Flow	0.86	1.12	1.93	2
	Fused	**0.47**	**0.62**	**0.95**	**5**
MMPD	Direct	9.57	18.64	34.76	24
	Flow	8.99	17.41	28.18	31
	Fused	**4.13**	**7.01**	**19.05**	**45**

The benefit remains substantial on MMPD. Fused obtains an MAE of 4.13 bpm, an RMSE of 7.01 bpm, and a 90th-percentile error of 19.05 bpm. It is the best path for 45 of 100 sequences. The Direct path provides a stable periodic and phase anchor. The Rectified Flow path corrects nonlinear morphology and limits large errors on difficult

samples. Their errors are not identical. Fusion can therefore retain their shared physiological component while attenuating path-specific motion, illumination, and local noise.

5 Conclusion

This paper presented TPA-Flow, a unified framework for remote pulse estimation. The proposed method employs quality-aware multi-source physiological representations to extract reliable facial cues, while period-guided Tensor-Product Attention captures higher-order dependencies across facial regions and cardiac cycles. Using the Direct prediction as a stable anchor, the Rectified Flow decoder further corrects phase deviations and local waveform distortions. Within-dataset and cross-dataset experiments demonstrated that TPA-Flow achieves a favorable balance among heart-rate estimation accuracy, waveform fidelity, and cross-domain generalization. Ablation and physiological attribution analyses further confirmed the effectiveness of its core components and dual-path decoding design.

References

Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., Dubois, J.: Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognition Letters **124**, 82–90 (2019)

Cao, M., Cheng, X., Liu, X., Jiang, Y., Yu, H., Shi, J.: ST-Phys: Unsupervised spatio-temporal contrastive remote physiological measurement. IEEE Journal of Biomedical and Health Informatics **28**(8), 4613–4624 (2024)

Chen, W., McDuff, D.: DeepPhys: Video-based physiological measurement using convolutional attention networks. In: European Conference on Computer Vision, pp. 349–365 (2018)

Chu, S., Xia, M., Yuan, M., *et al.*: CodePhys: Robust video-based remote physiological measurement through latent codebook querying. IEEE Journal of Biomedical and Health Informatics **29**(7), 4932–4945 (2025)

Das, M., Bhuyan, M.K., Sharma, L.N.: Time-frequency learning framework for rPPG signal estimation using scalogram-based feature map of facial video data. IEEE Transactions on Instrumentation and Measurement **72** (2023). Art. no. 4007710

Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rPPG. IEEE Transactions on Biomedical Engineering **60**(10), 2878–2886 (2013)

Gong, Z., Yang, X., Song, R., *et al.*: Heart rate estimation in driver monitoring system using quality-guided spectrum peak screening. IEEE Transactions on Instrumentation and Measurement **73**, 1–14 (2024). Art. no. 5008214

Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (2023)

Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: International Conference on Learning Representations (2023)

Liu, X., Hill, B., Jiang, Z., Patel, S., McDuff, D.: EfficientPhys: Enabling simple, fast and accurate camera-based cardiac measurement. In: IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5008–5017 (2023)

Lu, H., Han, H., Zhou, S.K.: Dual-GAN: Joint BVP and noise modeling for remote physiological measurement. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12404–12413 (2021)

Liu, L., Xia, Z., Zhang, X., Feng, X., Zhao, G.: Illumination variation-resistant network for heart rate measurement by exploring RGB and MSR spaces. IEEE Transactions on Instrumentation and Measurement **73** (2024). Art. no. 5026613

Liu, S.-Q., Yuen, P.C.: Robust remote photoplethysmography estimation with environmental noise disentanglement. IEEE Transactions on Image Processing **33**, 27–41 (2024)

Li, J., Yu, Z., Shi, J.: Learning motion-robust remote photoplethysmography through arbitrary resolution videos. In: AAAI Conference on Artificial Intelligence, vol. 37, pp. 1334–1342 (2023)

Niu, X., Shan, S., Han, H., Chen, X.: RhythmNet: End-to-end heart rate estimation from face via spatial-temporal representation. IEEE Transactions on Image Processing **29**, 2409–2423 (2020)

Song, R., Chen, H., Cheng, J., Li, C., Liu, Y., Chen, X.: PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography. IEEE Journal of Biomedical and Health Informatics **25**(5), 1373–1384 (2021) doi:10.1109/JBHI.2021.3051176

Sun, Z., Li, X.: Contrast-Phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5835–5851 (2024)

Tang, J., Chen, K., Wang, Y., *et al.*: MMPD: Multi-domain mobile video physiology dataset. In: IEEE Engineering in Medicine and Biology Society Conference, pp. 1–5 (2023)

Verkruysse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. Optics Express **16**(26), 21434–21445 (2008)

Wang, W., Brinker, A.C., Stuijk, S., Haan, G.: Algorithmic principles of remote-PPG. IEEE Transactions on Biomedical Engineering **64**(7), 1479–1491 (2017) doi:10.1109/TBME.2016.2609282

Wang, H., Huang, J., Wang, G., Lu, H., Wang, W.: Contactless patient care using hospital IoT: CCTV-camera-based physiological monitoring in ICU. IEEE Internet of Things Journal **11**(4), 5781–5797 (2024)

Wang, J., Wei, X., Lu, H., Chen, Y., He, D.: ConDiff-rPPG: Robust remote physiological measurement to heterogeneous occlusions. IEEE Journal of Biomedical and Health Informatics **28**(12), 7090–7102 (2024)

Xia, Z., Luo, Z., Chen, C.-H., Shen, X.: An effective photoplethysmography denoising method based on diffusion probabilistic model. IEEE Journal of Biomedical and Health Informatics **29**(6), 4071–4080 (2025)

Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In: British Machine Vision Conference (2019)

Yue, Z., Shi, M., Ding, S.: Facial video-based remote physiological measurement via self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 13844–13859 (2023)

Yu, Z., Shen, Y., Shi, J., *et al.*: PhysFormer: Facial video-based physiological measurement with temporal difference Transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4186–4196 (2022)

Yu, Z., Shen, Y., Shi, J., *et al.*: PhysFormer++: Facial video-based physiological measurement with SlowFast temporal difference Transformer. International Journal of Computer Vision **131**, 1307–1330 (2023)

Zou, B., Guo, Z., Chen, J., Zhuo, J., Huang, W., Ma, H.: RhythmFormer: Extracting patterned rPPG signals based on periodic sparse attention. Pattern Recognition **164**, 111511 (2025)

Zou, B., Guo, Z., Hu, X., Ma, H.: RhythmMamba: Fast, lightweight, and accurate remote physiological measurement. In: AAAI Conference on Artificial Intelligence, vol. 39, pp. 11077–11085 (2025)

Zhang, Y., Liu, Y., Yuan, H., Qin, Z., Yuan, Y., Gu, Q., Yao, A.C.-C.: Tensor product attention is all you need. In: Advances in Neural Information Processing Systems (2025)

Statements and Declarations

Funding. Details are provided in the separate title page and in the submission system to preserve double-blind review.

Competing interests. Details are provided in the separate title page and in the submission system to preserve double-blind review.

Author contributions. The author-contribution statement is provided in the separate title page.

Data availability. The PURE, UBFC-rPPG, and MMPD benchmark datasets are available from their respective providers subject to the original access conditions. Derived evaluation records and implementation details supporting the findings are available from the corresponding author upon reasonable request.

Ethics approval and consent to participate. This study is a secondary analysis of publicly released benchmark datasets and involved no new recruitment or collection of human-participant data. Ethical approval and informed consent were handled by the original dataset providers.

Generative AI statement. ChatGPT was used to assist with language refinement. All technical content, analyses, interpretations, and the final manuscript were reviewed and approved by the authors, who take full responsibility for the work.