FreqPhys: Repurposing Implicit Physiological Frequency Prior for Robust Remote Photoplethysmography

Wei Qian[1], ✉Dan Guo[1,2,3], Jinxing Zhou[4], Bochao Zou[5], Zitong Yu[6,7], and Meng Wang[1,2,3]

[1]Hefei University of Technology
[2]Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[4]MBZUAI, [5]University of Science and Technology Beijing, [6]Great Bay University
[7]Dongguan Key Laboratory for Intelligence and Information Technology

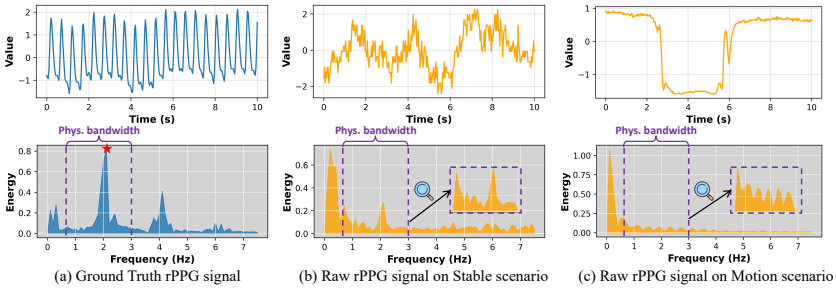
Abstract. Remote photoplethysmography (rPPG) enables contactless physiological monitoring by capturing subtle skin-color variations from facial videos. However, most existing methods predominantly rely on time-domain modeling, making them vulnerable to motion artifacts and illumination fluctuations, where weak physiological clues are easily overwhelmed by noise. To address these challenges, we propose **FreqPhys**, a frequency-guided rPPG framework that explicitly leverages physiological frequency priors for robust signal recovery. Specifically, **FreqPhys** first applies a *Physiological Bandpass Filtering* module to suppress out-of-band interference, and then performs *Physiological Spectrum Modulation* together with *adaptive spectral selection* to emphasize pulse-related frequency components while suppress residual in-band noise. A *Cross-domain Representation Learning* module further fuses these spectral priors with deep time-domain features to capture informative spatial–temporal dependencies. Finally, a frequency-aware conditional diffusion process progressively reconstructs high-fidelity rPPG signals. Extensive experiments on six benchmarks demonstrate that **FreqPhys** yields significant improvements over state-of-the-art approaches, particularly under challenging motion conditions. It highlights the importance of explicitly modeling physiological frequency priors. Our code is available at the `https://github.com/WeiQian98/FreqPhys`.

Keywords: Remote photoplethysmography · Physiological frequency priors · Conditional diffusion

1 Introduction

Physiological signals such as heart rate (HR), heart rate variability (HRV), and respiration frequency (RF) are fundamental indicators of physical and mental

✉: Corresponding authors (`guodan@hfut.edu.cn`).



(a) Ground Truth rPPG signal (b) Raw rPPG signal on Stable scenario (c) Raw rPPG signal on Motion scenario

Fig. 1: Visualization of the differences between ground-truth and raw rPPG signals in time and frequency domains. (a) Ground-truth rPPG signal, whose spectrum exhibits clear physiological priors: (*i*) **Physiological Band Constraint**, where the spectral energy is concentrated within the physiological bandwidth of [0.66, 3.0] Hz corresponding to normal heart rate ranges, and (*ii*) **Dominant Peak Property**, where a prominent spectral peak within this band reflects the periodic cardiac rhythm, while other in-band components remain comparatively low. The dominant frequency peak (marked by ★) corresponds to heart rate and is converted to beats per minute by multiplying the frequency by 60. (b)(c) Raw rPPG signals extracted from facial videos under stable and motion conditions, computed by averaging green-channel pixel intensities over time [49]. In the time domain, physiological signals are heavily entangled with noise. In the frequency domain, noise manifests as both out-of-band interference and residual in-band components, predominantly concentrated at lower frequencies. While the stable scenario preserves a relatively distinct spectral peak, motion artifacts disperse the in-band energy distribution, making reliable denoising considerably more challenging.

health. Traditional electrocardiogram (ECG) and photoplethysmography (PPG) systems require direct skin contact, which may cause discomfort and limit their applicability. Remote photoplethysmography (rPPG) [43] has therefore emerged as a promising non-invasive alternative optical technique, enabling applications in health monitoring [10], face anti-spoofing [42, 52], psychological stress assessment [6], and affective computing [8, 12, 13]. Despite its potential, reliably extracting rPPG signals, subtle skin color changes caused by blood volume fluctuations in facial videos captured by commodity cameras, remains challenging, particularly under motion and illumination variation.

Early rPPG studies [4,27,43,49] mainly relied on handcrafted signal processing pipelines, which depended heavily on specific assumptions. With the rapid development of deep learning [44–48, 57], numerous data-driven rPPG models have been proposed [14–16, 20, 23, 25, 28, 30, 31, 33, 38, 53, 60]. Although these methods achieve strong performance under stable conditions, their performance often degrades in complex scenarios due to noise sources such as motion artifacts and illumination variations [32, 34]. More recently, diffusion models have been introduced for rPPG estimation [2, 32], leveraging their powerful capability to model complex noise distributions and progressively recover clean signals. Nevertheless, these approaches predominantly operate in the time domain, where noise

appears in irregular and highly entangled forms (Fig. 1(b)(c)), making it inherently challenging to disentangle physiological components from interference. Several recent studies have attempted to incorporate frequency-domain information into deep rPPG models. Typical strategies include introducing auxiliary spectral losses [40, 54, 60], embedding Fourier-based modules to enrich feature representations [60], or synthesize frequency-modulated augmentations [56]. While these efforts demonstrate the potential value of spectral clues, they do not explicitly leverage the intrinsic physiological priors that characterize cardiac-driven rPPG signals, nor do they fully address the structured spectral distortions introduced by motion artifacts.

In this work, we revisit the frequency domain and show that rPPG signals, driven by quasi-periodic cardiac rhythms, exhibit strong and well-defined spectral priors [7, 36], as illustrated in Fig. 1(a): (*i*) **Physiological Band Constraint:** the spectral energy is largely concentrated within a physiological frequency band, typically [0.66, 3.0] Hz, corresponding to the normal heart-rate range; and (*ii*) **Dominant Peak Property:** a prominent spectral peak emerges within this band, reflecting the periodic cardiac rhythm, while other in-band components remain comparatively weak. To further illustrate this phenomenon, Fig. 1(b)(c) presents raw rPPG signals extracted under stable and motion conditions by computing the mean green-channel intensity over time [49]. While time-domain signals exhibit severe entanglement between physiological patterns and noise, their frequency-domain representations reveal two distinct types of interference: (*i*) **out-of-band components dominated by low-frequency drift**, and (*ii*) **residual in-band disturbances that blur the cardiac peak under motion**. These observations naturally raise a central question: ***How can we simultaneously suppress out-of-band and in-band noise while preserving physiologically meaningful spectral structures?***

To address this challenge, we propose **FreqPhys**, a diffusion-based rPPG framework that repurposes implicit physiological frequency priors to guide robust signal recovery. Specifically, to remove out-of-band interference, we first apply a *Physiological Bandpass Filter* that retains only cardiac-relevant spectral components. To further emphasize true cardiac harmonics and mitigate residual in-band noise, we introduce *Physiological Spectrum Modulation* together with *Adaptive Spectral Selection*, enabling dynamic enhancement and suppression within the physiological frequency range. Finally, we design a *Cross-domain Representation Learning* module that employs cross-attention to fuse spectral priors with deep time-domain representations, providing the diffusion model with explicit physiological regularities and informative temporal dependencies. Our main contributions are summarized as follows:

- We highlight the importance of explicitly leveraging implicit physiological frequency priors for robust rPPG estimation, revealing the structured spectral characteristics induced by cardiac rhythms.
- We propose **FreqPhys**, a frequency-aware diffusion framework that integrates physiological frequency denoising with cross-domain representation learning for robust rPPG signal reconstruction.

- Unlike prior diffusion-based rPPG methods that operate solely in the time domain, our method incorporates frequency-domain conditioning to better capture the quasi-periodic structure of rPPG signals.
- Extensive experiments on six public datasets demonstrate that our method achieves superior performance in both robustness and generalization, especially under challenging motion conditions.

2 Related Work

Traditional rPPG Methods. Early rPPG works mainly relied on classical signal processing techniques, including spatial averaging or color–separation–based approaches such as GREEN [43], ICA [27], CHROM [4], and POS [49]. These methods are computationally efficient and easy to implement; however, they typically depend on strong assumptions (*e.g.*, stable illumination conditions or linear color reflection models). As a result, their performance degrades considerably in unconstrained complex scenarios where motion artifacts and illumination variations are prevalent.

Deep Learning–based rPPG Methods. The success of deep learning has led to a wide range of data-driven rPPG models. CNN-based architectures, such as DeepPhys [3], PhysNet [53], CVD [23], and TS-CAN [18], focus on learning local spatial–temporal representations from facial videos. To capture longer temporal dependencies, Transformer-based models, including PhysFormer [55], EfficientPhys [19], and Dual-TL [28], have been proposed. More recently, state-space and Mamba-style architectures (*e.g.*, PhysMamba [21] and RhythmMamba [60]) have been explored to enlarge the temporal receptive field while maintaining efficient sequential modeling. In addition, diffusion-based frameworks, such as DiffPhys [2] and PhysDiff [32], leverage generative denoising processes to progressively recover clean physiological signals from noisy temporal input.

Frequency-domain rPPG Methods. Beyond purely time-domain modeling, several studies have explored incorporating spectral clues into rPPG estimation. Existing approaches generally fall into three categories: (*i*) introducing frequency-based loss functions to regularize training, as in PhysFormer++ [54], ContrastPhys [40], and RhythmMamba [60]; (*ii*) designing frequency-aware representations, such as Fourier-based blocks or spectral embeddings [60]; and (*iii*) leveraging frequency-guided augmentation or synthesis strategies, for example, generating negative samples through spectral manipulation [56]. However, these approaches mainly incorporate spectral information as auxiliary supervision or representation enhancement, without explicitly addressing the structured spectral distortions introduced by motion and illumination. In practice, these disturbances manifest as both out-of-band interference and residual in-band noise that overlap with cardiac rhythms, making it difficult for existing methods to reliably isolate physiological components. A more comprehensive discussion of related work is provided in **Appendix A**.

Remark. Our work revisits the role of frequency information in rPPG modeling. Existing approaches typically use spectral cues either (*i*) during post-processing

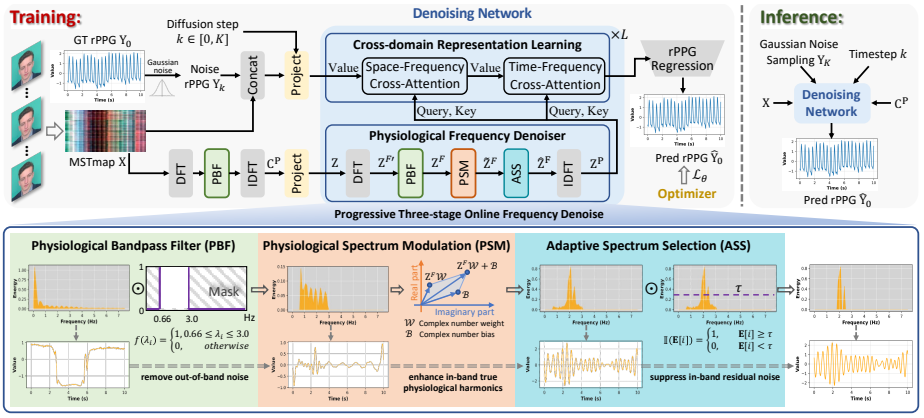


Fig. 2: The pipeline of proposed FreqPhys. Given a facial video, we first construct MSTmap $\mathbf{X}$ as the temporal condition and generate the frequency condition $\mathbf{C^P}$ by applying the PBF. During training, we initially generate noise rPPG $\mathbf{Y}_k$ by adding Gaussian noise to Ground Truth rPPG $\mathbf{Y}_0$ for the k-th step. Then, we input $\mathbf{Y}_k$, $\mathbf{X}$, k, and $\mathbf{C^P}$ into the *Denoising Network*. Specifically, the frequency condition $\mathbf{C^P}$ is fed into the *Physiological Frequency Denoiser* module to enhance physiological spectral clues through three key steps: (*i*) PBF removes out-of-band noise based on the physiological frequency bandwidth [0.66,3.0] Hz; (*ii*) PSM emphasizes valid physiological harmonics by modeling interactions between real and imaginary components; (*iii*) ASS dynamically suppresses in-band noise using data-driven energy thresholds. Next, with *Cross-domain Representation Learning*, our FreqPhys includes frequency-domain denoised information into space and time dependencies modeling to estimate the high-fidelity rPPG signal. During inference, the initial rPPG $\mathbf{Y}_K$ is randomly sampled from Gaussian noise, with frequency condition and denoising network processes mirroring those used in training.

for HR estimation [19, 20, 23, 28, 55], or (*ii*) as auxiliary training components, such as Fourier blocks [60], spectral augmentations [56], or frequency-based losses [40, 54]. In contrast, FreqPhys incorporates implicit physiological frequency priors directly into the online denoising process as an internal structural constraint. Specifically, we introduce a progressive three-stage frequency-guided diffusion denoising framework consisting of (*i*) out-of-band suppression, (*ii*) cardiac harmonic enhancement, and (*iii*) adaptive in-band selection, which is applied at every diffusion step during both training and inference.

3 Methodology

Remote physiological measurement from facial videos can be regarded as a video sequence to signal sequence problem. Let $\mathbf{V} \in \mathbb{R}^{T\times 3\times H\times W}$ denote a raw facial video clip containing T frames with 3 color channels and spatial resolution $H \times W$. Following established rPPG preprocessing protocols [23, 32], we extract N facial regions of interest (ROI) through landmark alignment and pixel-level

average pooling of C color channels, constructing a multi-scale temporal map (MSTmap) $\mathbf{X} \in \mathbb{R}^{T\times N\times C}$ as the model input. The objective is to recover the clean periodic rPPG signal $\mathbf{Y} \in \mathbb{R}^{T}$ from $\mathbf{X}$, formulated as learning a denoising function $f_\theta : \mathbf{X} \mapsto \mathbf{Y}$, where θ denotes trainable parameters. The overview of our method is illustrated in Fig. 2, where the details are described as follows.

3.1 Physiological Frequency Denoiser

Physiological Bandpass Filter (PBF). Inspired by the fact that true cardiac activities mainly fall within a fixed frequency bandwidth, typically [0.66,3.0] Hz [49], we devise a *Physiological Bandpass Filter* that directly isolates cardiac frequency components in the spectral space. Specifically, we first project the frequency condition $\mathbf{C^P} \in \mathbb{R}^{T\times N\times C}$ into a D-dimensional latent space $\mathbf{Z} \in \mathbb{R}^{T\times N\times D}$ by a linear layer, then transform it to frequency domain via the Discrete Fourier Transform $\mathcal{F}$,

$$\mathbf{Z}^{F'} = \mathcal{F}(\mathbf{Z}) \in \mathbb{C}^{(\lfloor T/2\rfloor+1)\times N\times D}. \tag{1}$$

Then, the noisy frequency components outside the physiological bandwidth range can be discarded using an ideal bandpass filter:

$$\mathbf{Z}^{F} = \mathbf{Z}^{F'} \odot f(\lambda_i), \quad \text{for } i = 0, \ldots, \lfloor T/2\rfloor, \tag{2}$$

where $\odot$ denotes the Hadamard product. λ_i denotes the physical frequency corresponding to the i-th frequency bins and $\lambda_i = if_s/T$ Hz, with f_s denoting the sampling rate. $f(\lambda_i)$ is an indicator function, which outputs 1 when $(0.66 \leq \lambda_i \leq 3.0)$ and 0 otherwise.

Physiological Spectrum Modulation (PSM). While the *Physiological Bandpass Filter* is effective in removing noise outside the physiological frequency band, another challenge still exists where noise components may overlap or closely resemble physiological signals within this band. Such overlapping frequencies may severely distort the signal, making it difficult to accurately extract physiological features. We apply a learnable *Physiological Spectrum Modulation* in the frequency domain to emphasize true physiological harmonics while suppressing non-physiological components. The detailed implementation process of PSM is shown in Fig. 3. Specifically, given the physiological frequency representation $\mathbf{Z}^F = \mathbf{Z}^{F\to re} + j\cdot\mathbf{Z}^{F\to im} \in \mathbb{C}^{(\lfloor T/2\rfloor+1)\times N\times D}$, we denote its real and imaginary parts as $\mathbf{Z}^{F\to re}$ and $\mathbf{Z}^{F\to im}$, separately. To achieve more exhaustive spectrum modulation, we encode the real and imaginary parts separately to generate the modulation signals:

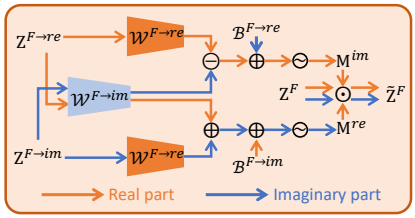


Fig. 3: The details of physiological spectrum modulation module.

$$\mathbf{M} = \mathbf{M}^{re} + j\cdot\mathbf{M}^{im} = \sigma(\mathbf{Z}^F\mathcal{W}^F + \mathcal{B}^F), \mathbf{M} \in \mathbb{C}^{(\lfloor T/2\rfloor+1)\times N\times D}, \tag{3}$$

where σ is the ReLU activation function, $\mathcal{W}^F = (\mathcal{W}^{F\to re} + j\cdot\mathcal{W}^{F\to im}) \in \mathbb{C}^{D\times D}$ is the trainable complex number weight matrix with $\{\mathcal{W}^{F\to re}, \mathcal{W}^{F\to im}\} \in \mathbb{R}^{D\times D}$, and $\mathcal{B}^F = (\mathcal{B}^{F\to re} + j\cdot\mathcal{B}^{F\to im}) \in \mathbb{C}^D$ is the trainable complex number biases with $\{\mathcal{B}^{F\to re}, \mathcal{B}^{F\to im}\} \in \mathbb{R}^D$. According to the rule of multiplication of complex numbers (details can be seen in **Appendix B.1**), we further unfold into real and imaginary parts as follows:

$$\begin{aligned}\mathbf{M}^{re} &= \sigma(\mathbf{Z}^{F\to re}\mathcal{W}^{F\to re} - \mathbf{Z}^{F\to im}\mathcal{W}^{F\to im} + \mathcal{B}^{F\to re}),\\ \mathbf{M}^{im} &= \sigma(\mathbf{Z}^{F\to re}\mathcal{W}^{F\to im} + \mathbf{Z}^{F\to im}\mathcal{W}^{F\to re} + \mathcal{B}^{F\to im}).\end{aligned} \tag{4}$$

Afterwards, the generated complex signal is used to modulate counterparts of the original frequency-domain feature, which can be written as,

$$\tilde{\mathbf{Z}}^F = \mathbf{M} \odot \mathbf{Z}^F \in \mathbb{C}^{(\lfloor T/2\rfloor+1)\times D}. \tag{5}$$

By Theorem 1, this spectral multiplication operation in Equation 5 is mathematically equivalent to a global circular convolution in time, endowing each sequence with a content-adaptive receptive field that is ideal for capturing periodic cardiac rhythms. By means of Equation 5, the physiological frequency components can be effectively enhanced via direct spectral modulation. The detailed proof of *theorem* 1 is illustrated in **Appendix C.1**.

Theorem 1. *(Frequency-domain Convolution Theorem) The multiplication of two signals in the frequency domain is equivalent to the frequency transformation of a circular convolution of these two signals in the temporal domain, which can be summarized as:*

$$\mathcal{F}[\mathbf{M}(v) \otimes \mathbf{Z}(v)] = \mathcal{F}(\mathbf{M}(v)) \odot \mathcal{F}(\mathbf{Z}(v)), \tag{6}$$

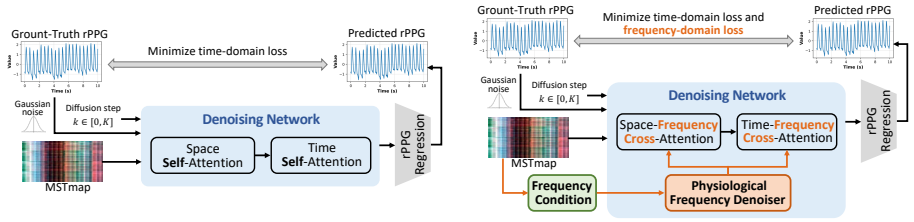
where $\otimes$ and $\odot$ represent circular convolutional operation and element multiplication operation, respectively, $\mathbf{M}(v)$ and $\mathbf{Z}(v)$ represent two signals for the time variable v, and $\mathcal{F}(\cdot)$ denotes the Discrete Fourier Transform.

Adaptive Spectrum Selection (ASS). Although the physiological spectrum modulation block enhances the cardiac frequency band, residual noise components whose frequencies lie close to the physiological range may still persist. To further isolate the true pulse periodicity, we introduce an *Adaptive Spectrum Selection* (ASS) module that performs learnable frequency-domain filtering via a data-driven threshold. Given the modulated spectrum $\widetilde{\mathbf{Z}}^F$, we first compute its per-frequency energy:

$$\mathbf{E}[i] = \sqrt{(\widetilde{\mathbf{Z}}^{F\to re}[i])^2 + (\widetilde{\mathbf{Z}}^{F\to im}[i])^2} = \left\|\widetilde{\mathbf{Z}}^F[i]\right\|_2, \quad i = 0, 1, \ldots, \lfloor T/2\rfloor. \tag{7}$$

We then introduce a learnable threshold τ to adaptively distinguish dominant cardiac frequencies from residual noise. A binary selection mask is defined as:

$$m[i] = \mathbb{I}(\mathbf{E}[i] \geq \tau), \tag{8}$$



(a) Existing **time-domain-based** rPPG diffusion framework (b) Our **frequency-aware** diffusion framework

Fig. 4: Architecture comparison with existing diffusion methods.

where $\mathbb{I}(\cdot)$ denotes the indicator function. The filtered spectrum is computed as:

$$\hat{\mathbf{Z}}^F[i] = \widetilde{\mathbf{Z}}^F[i] \odot m[i]. \tag{9}$$

Importantly, the hard selection mask is trained using the Straight-Through Estimator (STE) strategy described previously, allowing the adaptive threshold τ to be jointly optimized with the entire network. Finally, the selected spectrum is transformed back to the time domain via inverse DFT:

$$\mathbf{Z}^{\mathbf{P}} = \mathcal{F}^{-1}(\hat{\mathbf{Z}}^F), \tag{10}$$

yielding a pulse signal dominated by cardiac frequency components.

3.2 Cross-domain Representation Learning

Time domain modeling focuses on local dependencies and transient behaviors, while frequency domain analysis provides insights into the global correlations and periodicity of the data. Therefore, combining these two domains is a promising approach to recover high-fidelity rPPG signals. To effectively integrate intermediate representations $\mathbf{Z}$ in the time domain with frequency-domain priors $\mathbf{Z}^{\mathbf{P}}$, we propose a cross-domain representation learning module. Specifically, we perform L alternating cross-attention layers that can progressively learn the various input domains for representation learning. In each layer l, the initial physiological frequency representation is first obtained by applying PBF, followed by PSM and ASS modules. Next, the physiological frequency representation $\mathbf{Z}^{\mathbf{P}}$ and intermediate representations $\mathbf{Z}$ are integrated by cross-attention across the spatial and temporal axis, respectively. Formally, this process can be formulated as follows:

$$\begin{aligned}
\mathbf{Z}^{\mathbf{P},(l)} &= \mathrm{ASS}(\mathrm{PSM}(\mathrm{PBF}(\mathbf{Z}^{(l)}))), \\
\mathbf{Z}^{(l)'} &= \mathrm{CA}(\mathbf{Z}^{\mathbf{P},(l)}, \mathbf{Z}^{(l)}) + \mathbf{Z}^{(l)}, \\
\mathbf{Z}^{(l+1)} &= \mathrm{CA}(\mathbf{Z}^{\mathbf{P},(l)}, \mathbf{Z}^{(l)'}) + \mathbf{Z}^{(l)'},
\end{aligned} \tag{11}$$

where $\mathrm{CA}(a, b)$ refers to Cross-Attention, with a denotes query and key, and b denotes value.

3.3 Frequency-aware Diffusion Model

Recently, denoising diffusion probabilistic models (DDPMs) [9] have emerged as powerful generative frameworks that progressively refine noisy inputs through learned reverse Markov chains, capturing complex data distributions. Inspired by this, some diffusion-based rPPG methods [2, 32] for rPPG estimation have been proposed and achieved SOTA performance. They treat the rPPG estimation task as calculating the conditional rPPG signal probability distribution $q(\mathbf{Y}_0|\mathbf{C})$, where $q(\mathbf{Y}_0)$ is the clean rPPG distribution, and the condition $\mathbf{C}$ for probability distribution calculation is generally the input $\mathbf{X}$ in the time domain. However, these diffusion models mainly focus on time-domain conditioning and overlook the unique spectral prior. To alleviate this limitation, we introduce a novel frequency-aware diffusion model that explicitly incorporates physiological frequency priors to guide the generation of high-fidelity rPPG signals, as shown in Fig. 4. Specifically, our frequency-aware diffusion model fuses with physiological frequency condition $\mathbf{C}^{\mathbf{P}}$ to learn the conditional rPPG distribution $q(\mathbf{Y}_0|\mathbf{Y}, \mathbf{X}, \mathbf{C}^{\mathbf{P}})$, through two Markov chain processes of diffusion step K, *i.e.*, the forward process and the reverse process.

Forward Process. The forward process q incrementally adds Gaussian noise to the ground truth rPPG signal $\mathbf{Y}_0 \in \mathbb{R}^T$ via a fixed Markov chain $\mathbf{Y}_0, \dots \mathbf{Y}_K$ as follows:

$$\begin{aligned} q(\mathbf{Y}_{1:K}|\mathbf{Y}_0) &= \prod_{k=1}^{K} q(\mathbf{Y}_k|\mathbf{Y}_{k-1}), \\ q(\mathbf{Y}_k|\mathbf{Y}_{k-1}) &= \mathcal{N}(\mathbf{Y}_k; \sqrt{1-\beta_k}\mathbf{Y}_{k-1}, \beta_k\mathbf{I}), \end{aligned} \tag{12}$$

where β_k is a noise schedule, satisfying $\beta_k < \beta_{k-1}$. As K becomes large, $\mathbf{Y}_K \approx \mathcal{N}(0, \mathbf{I})$. Following DDPM [9], we sample $\mathbf{Y}_k$ from $\mathbf{Y}_0$ at any time step k in a closed form:

$$q(\mathbf{Y}_k|\mathbf{Y}_0) = \mathcal{N}(\mathbf{Y}_k; \sqrt{\bar{\alpha}_k}\mathbf{Y}_0, (1-\bar{\alpha}_k)\mathbf{I}), \tag{13}$$

where $\alpha_k = 1 - \beta_k$ and $\bar{\alpha}_k = \prod_{s=0}^{k} \alpha_s$. Utilizing the parameterization trick [11], we express $\mathbf{Y}_k$ as:

$$\mathbf{Y}_k = \sqrt{\bar{\alpha}_k}\mathbf{Y}_0 + \sqrt{1-\bar{\alpha}_k}\epsilon, \tag{14}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The detailed derivations are provided in **Appendix D.1**.

Reverse Process. The reverse process aims to estimate the posterior $q(\mathbf{Y}_{k-1}|\mathbf{Y}_k)$. Different from PhysDiff [32], in our frequency-aware diffusion model, this distribution is approximated by a neural network f_θ conditioned on both the time-domain condition $\mathbf{X}$ and the physiological frequency condition $\mathbf{C}^{\mathbf{P}}$:

$$p_\theta(\mathbf{Y}_{k-1}|\mathbf{Y}_k, \mathbf{X}, \mathbf{C}^{\mathbf{P}}) = \mathcal{N}(\mathbf{Y}_{k-1}; \mu_\theta(\mathbf{Y}_k, \mathbf{X}, \mathbf{C}^{\mathbf{P}}, k), \Sigma_\theta). \tag{15}$$

Next, we show that incorporating the physiological frequency prior $\mathbf{C}^{\mathbf{P}}$ can effectively reduce the uncertainty in the reverse diffusion process, leading to more accurate rPPG signal reconstruction. It can be formalized in Proposition 1.

Proposition 1. *The conditional entropy is satisfied:*

$$\mathbf{H}(\mathbf{Y}_{k-1}|\mathbf{Y}_k, \mathbf{X}, \mathbf{C}^{\mathbf{P}}) < \mathbf{H}(\mathbf{Y}_{k-1}|\mathbf{Y}_k, \mathbf{X}), \tag{16}$$

indicating that the inclusion of additional physiological frequency condition $\mathbf{C}^{\mathbf{P}}$ *in the reverse process reduces uncertainty. The detailed proof is provided in* ***Appendix C.2****.*

Training. Traditional DDPM-based training involves learning to predict the added Gaussian noise at each diffusion step, which can be inefficient [9]. Instead, our denoising network f_θ is designed to directly reconstruct the clean rPPG signal $\mathbf{Y}_0$ from the noisy input $\mathbf{Y}_0$, conditioned on $\mathbf{X}$, $\mathbf{C}^{\mathbf{P}}$, and the timestep k:

$$\hat{\mathbf{Y}}_0 = f_\theta(\mathbf{Y}_k, \mathbf{X}, \mathbf{C}^{\mathbf{P}}, k). \tag{17}$$

In practice, for Equation 15, the mean μ_θ and covariance σ_k^2 in reverse process are parameterized as $\mu_\theta(\mathbf{Y}_k, \mathbf{X}, \mathbf{C}^{\mathbf{P}}, k)$ and Σ_θ (details are presented in **Appendix D.2**). Furthermore, inspired by the Fourier-based loss term, which is beneficial for the accurate reconstruction of the signals [5], we propose to guide the diffusion training by applying it to the frequency domain with the Fourier transform. Formally, our training objective integrates both time and frequency-domain constraints:

$$\mathcal{L}_\theta(\hat{\mathbf{Y}}_0, \mathbf{Y}_0) = \underbrace{1 - Pearson(\hat{\mathbf{Y}}_0, \mathbf{Y}_0)}_{\text{time-domain loss}} + \underbrace{MSE(\mathcal{F}(\hat{\mathbf{Y}}_0), \mathcal{F}(\mathbf{Y}_0))}_{\text{frequency-domain loss}}, \tag{18}$$

where $Pearson$ represents Pearson correlation coefficient, MSE denotes Mean Square Error, and $\mathcal{F}$ denotes the Discrete Fourier Transform. For inference, we start from $\mathbf{Y}_K \sim \mathcal{N}(0, \mathbf{I})$, K, $\mathbf{X}$, and $\mathbf{C}^{\mathbf{P}}$. Then, we follow DDIM [32, 35] and perform the reverse process to obtain the final rPPG signal.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We evaluate our method using both intra-dataset and cross-dataset protocols on six widely used rPPG benchmark datasets, including UBFC-rPPG [1], PURE [37], MMPD [41], BUAA [50], VIPL-HR [22], and MR-NIRP-Car [26]. The detailed description of datasets and evaluation metrics is provided in **Appendix F.1** and **Appendix F.2**, respectively.

Implementation Details. Following prior works [20, 24, 32], facial videos are first transformed into MSTmap with a temporal length of $T = 300$ and a spatial dimension of $N = 63$. The proposed denoising network consists of $L = 4$ layers with a feature dimension of $D = 128$. The diffusion step is set to $K = 1000$, following PhysDiff [32]. During training, the model is optimized using Adam for 50 epochs with an initial learning rate of 1×10^{-3}, and all experiments are conducted on four NVIDIA RTX 4090 GPUs (24GB).

Table 1: Intra-dataset HR estimation results. Models are trained and tested on the UBFC-rPPG, PURE, BUAA, MMPD, VIPL-HR, and MR-NIRP-Car datasets. **Bold**: best results.

Method	Venue	UBFC-rPPG			PURE			BUAA			MMPD			VIPL-HR			MR-NIRP-Car		
		MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
Traditional rPPG Methods																			
GREEN [43]	Opt Express'08	7.50	14.41	0.62	7.23	17.05	0.69	6.89	10.39	0.60	21.68	27.69	-0.01	-	-	-	11.88	14.46	0.17
ICA [27]	Opt Express'10	5.17	11.76	0.65	3.76	12.60	0.85	-	-	-	18.60	24.30	0.01	-	-	-	10.87	13.62	0.02
CHROM [4]	TBE'13	2.37	4.91	0.89	2.07	2.50	0.99	-	-	-	13.66	18.76	0.08	11.40	16.90	0.28	7.16	9.09	0.35
Deep Learning-based rPPG Methods																			
DeepPhys [3]	ECCV'18	6.27	10.82	0.65	0.83	1.54	0.99	-	-	-	22.27	28.92	-0.03	11.0	13.8	0.72	11.98	14.18	0.05
PhysNet [53]	BMVC'19	2.95	3.67	0.97	2.10	2.60	0.99	10.89	11.70	-0.04	4.80	11.80	0.60	10.80	14.80	0.20	11.18	13.53	0.17
CVD [23]	ECCV'20	2.19	3.12	0.99	1.29	2.01	0.98	-	-	-	-	-	-	5.02	7.97	0.79	-	-	-
TS-CAN [18]	NeurIPS'20	1.70	2.72	0.99	2.48	9.01	0.92	-	-	-	9.71	17.22	0.44	-	-	-	13.62	15.46	0.19
Gideon et al. [7]	ICCV'21	1.85	4.28	0.93	2.30	2.90	0.99	-	-	-	-	-	-	9.01	14.02	0.58	-	-	-
Dual-GAN [20]	CVPR'21	0.44	0.67	0.99	0.82	1.31	0.99	-	-	-	-	-	-	4.93	7.68	0.81	-	-	-
PhysFormer [55]	CVPR'22	0.50	0.71	0.99	1.10	1.75	0.99	8.46	10.17	-0.06	11.99	18.41	0.18	4.97	7.79	0.78	10.17	12.57	-0.20
Contrast-Phys [39]	ECCV'22	0.64	1.00	0.99	1.00	1.40	0.99	-	-	-	-	-	-	32.1	36.1	0.04	-	-	-
EfficientPhys [19]	WACV'23	1.14	1.81	0.99	4.75	9.39	0.99	16.09	16.80	0.14	13.47	21.32	0.21	-	-	-	18.36	20.01	-0.04
Li et al. [17]	ICCV'23	0.48	0.64	**1.00**	0.64	1.16	0.99	-	-	-	-	-	-	4.97	7.79	0.78	-	-	-
Yue et al. [56]	TPAMI'23	0.58	0.94	0.99	1.23	2.01	0.99	-	-	-	-	-	-	-	-	-			
Dual-TL [28]	TCSS'24	0.17	0.41	0.99	0.37	0.68	0.99	-	-	-	-	-	-	4.36	6.92	0.69	-	-	-
Cluster-Phys [29]	ACM MM'24	0.25	0.50	1.00	0.37	0.61	1.00	-	-	-	-	-	-	4.07	6.79	0.84	-	-	-
Contrast-Phys+! [40]	TPAMI'24	0.21	0.80	0.99	0.48	0.98	0.99	-	-	-	-	-	-	-	-	-	-	-	-
RhythmMamba [60]	AAAI'25	0.50	0.75	0.99	0.23	0.34	0.99	11.99	14.22	0.16	**3.16**	7.27	0.84	4.30	7.49	0.81	8.07	10.03	0.24
PhysDiff [32]	AAAI'25	0.33	0.57	**1.00**	0.29	0.54	**1.00**	-	-	-	7.17	9.63	0.71	3.92	6.65	0.85	7.58	8.58	-0.14
PhysLLM [51]	ICLR'26	0.21	0.57	0.99	0.17	0.35	0.99	6.48	8.48	0.63	4.36	10.76	0.65	4.24	6.81	-	-	-	-
PHASE-Net [58]	CVPR'26	**0.15**	0.53	0.99	**0.14**	0.35	0.99	5.89	7.89	0.48	4.78	8.22	0.71	-	-	-	-	-	-
FreqPhys (Ours)	-	0.17	**0.41**	**1.00**	0.17	**0.25**	**1.00**	**1.13**	**1.31**	**0.99**	4.20	**6.78**	**0.86**	**3.79**	**6.34**	**0.86**	**5.75**	**6.08**	**0.54**

Table 2: HRV and RF estimation results. Models are trained and tested on the UBFC-rPPG dataset. LF, HF, and RF represent low frequency, high frequency, and respiration frequency, respectively. "n.u." denotes normalized units.

Method	Venue	LF (n.u.)			HF (n.u)			LF/HF			RF (Hz)		
		SD↓	RMSE↓	r ↑	SD↓	RMSE↓	r ↑	SD↓	RMSE↓	r ↑	SD↓	RMSE↓	r ↑
Traditional rPPG Methods													
GREEN [43]	Opt Express'08	0.186	0.186	0.280	0.186	0.186	0.280	0.361	0.365	0.492	0.087	0.086	0.111
ICA [27]	Opt Express'10	0.243	0.240	0.159	0.243	0.240	0.159	0.655	0.645	0.226	0.086	0.089	0.102
POS [49]	TBE'16	0.171	0.169	0.479	0.171	0.169	0.479	0.405	0.399	0.518	0.109	0.107	0.087
Deep Learning-based rPPG Methods													
CVD [23]	ECCV'20	0.053	0.056	0.740	0.053	0.065	0.740	0.169	0.168	0.812	0.017	0.018	0.252
Dual-GAN [20]	CVPR'21	0.034	0.035	0.891	0.034	0.034	0.891	0.131	0.136	0.881	0.010	0.010	0.395
Gideon *et al.* [7]	ICCV'21	0.091	0.139	0.694	0.091	0.139	0.694	0.525	0.691	0.684	0.061	0.098	0.103
Contras-Phys [39]	ECCV'22	0.050	0.098	0.798	0.050	0.098	0.798	0.205	0.395	0.782	0.055	0.083	0.347
Contrast-Phys+ [40]	TPAMI'24	0.025	0.025	0.947	0.025	0.025	0.947	**0.064**	**0.066**	0.963	0.029	0.029	0.803
PhysDiff [32]	AAAI'25	0.029	0.022	0.978	0.016	0.022	0.978	0.079	**0.066**	0.979	**0.006**	0.007	0.811
FreqPhys (Ours)	-	**0.016**	**0.014**	**0.988**	**0.013**	**0.014**	**0.988**	0.079	**0.066**	**0.989**	**0.006**	**0.005**	**0.845**

4.2 Intra-dataset Evaluation.

All dataset splits strictly follow the widely used benchmark protocols [32, 51, 58–60], and all evaluations are conducted in a subject-independent manner to avoid identity leakage. Specifically, for UBFC-rPPG, subjects 38 to 49 are used as the test set, while the remaining subjects are used for training. For PURE and BUAA, the datasets are sequentially divided by subject identity into training and testing sets with ratios of 6:4 and 7:3, respectively. For VIPL-HR, we adopt a subject-exclusive 5-fold cross-validation protocol. For MMPD and MR-NIRP-Car, the datasets are sequentially divided by subject identity into training, validation, and testing sets with a ratio of 7:1:2.

Tab. 1 reports the quantitative comparison with state-of-the-art methods. Notably, the evaluated datasets exhibit progressively increasing levels of difficulty. UBFC-rPPG and PURE are relatively controlled environments, whereas

Table 3: Cross-dataset HR estimation results. Models are trained on PURE/UBFC-rPPG, and tested on UBFC-rPPG/PURE/MMPD.

Method	Venue	PURE → UBFC-rPPG			UBFC-rPPG → PURE			PURE → MMPD			UBFC-rPPG → MMPD		
		MAE↓	RMSE↓	r↑	MAE↓	RMSE↓	r↑	MAE↓	RMSE↓	r↑	MAE↓	RMSE↓	r↑
Traditional rPPG Methods													
GREEN [43]	Opt Express'08	19.73	21.00	0.37	10.09	23.85	0.34	21.68	27.69	-0.01	21.68	27.69	-0.01
ICA [27]	Opt Express'10	16.00	25.65	0.44	4.77	16.07	0.72	18.06	24.30	0.01	18.06	24.30	0.01
CHROM [4]	TBE'13	4.06	8.83	0.89	5.77	14.93	0.81	13.66	18.76	0.08	13.66	18.76	0.08
Deep Learning-based rPPG Methods													
DeepPhys [3]	ECCV'18	1.21	2.90	0.99	5.54	18.51	0.66	16.92	24.61	0.05	17.50	25.00	0.05
PhysNet [53]	BMVC'19	1.63	3.79	0.98	9.36	20.63	0.62	13.22	19.61	0.23	10.24	16.54	0.29
TS-CAN [18]	NeurIPS'20	1.30	2.87	0.99	3.69	13.80	0.82	13.94	21.61	0.20	14.01	21.04	0.24
PhysFormer [55]	CVPR'22	1.44	3.77	0.98	12.92	24.36	0.47	14.57	20.71	0.15	12.10	17.79	0.17
EfficientPhys [19]	WACV'23	2.13	3.00	0.99	5.47	17.04	0.71	14.03	21.62	0.17	13.78	22.25	0.09
RhythmMamba [60]	AAAI'25	0.95	1.83	0.99	1.98	6.51	0.96	10.44	16.70	0.36	10.63	17.14	0.34
FreqPhys **(Ours)**	-	**0.43**	**0.79**	**1.00**	**0.95**	**3.15**	**0.99**	**10.11**	**14.34**	**0.48**	**8.91**	**12.86**	**0.57**

BUAA introduces extreme illumination variations. MMPD combines illumination changes with head motion and skin-tone diversity. The VIPL-HR dataset further increases the diversity of camera equipment and is currently the most complex dataset collected in laboratory settings. Among them, MR-NIRP-Car represents the most challenging real-world scenario, as it is collected in a natural driving environment with substantial motion, illumination fluctuation, and environmental interference. Despite these challenges, our method consistently achieves superior performance, with particularly large improvements on the more challenging complex datasets. Specifically, in terms of the most reliable metric, RMSE, our approach reduces the error from 7.89 (PHASE-Net [58]) to **1.31** on BUAA, from 7.27 (RhythmMamba [60]) to **6.78** on MMPD, from 6.65 (PhysDiff [32]) to **6.34** on VIPL-HR, and from 8.58 (PhysDiff [32]) to **8.03** on MR-NIRP-Car. These results demonstrate that incorporating physiological frequency priors enables our model to effectively suppress noise induced by illumination changes, motion artifacts, and cross-subject appearance variations.

In addition to HR estimation, we follow [20,23,32,39,40] and further evaluate our method on two other critical physiological indicators, *i.e.*, HRV and RF, both of which require high-quality rPPG signals for reliable peak detection and temporal analysis. As shown in Tab. 2, our method significantly outperforms existing methods across most metrics. These results indicate that our model not only improves HR estimation accuracy but also reconstructs high-fidelity rPPG signals in the time domain and frequency domain, enabling more reliable downstream physiological analysis.

4.3 Cross-dataset Evaluation.

In addition to intra-dataset evaluation, we further conduct cross-dataset experiments to simulate deployment in unseen environments and rigorously assess domain invariance. As shown in Tab. 3, we evaluate four cross-dataset transfer settings. These datasets exhibit significant domain gaps due to differences in illumination conditions, motion patterns, camera devices, and subject appearance, making cross-dataset generalization particularly challenging. Consequently, most existing methods experience notable performance degradation when transferring

Table 4: Impact of physiological frequency denoiser components.

PBF	PSM	ASS	VIPL-HR MAE↓	VIPL-HR RMSE↓	VIPL-HR ρ↑	MR-NIRP-Car MAE↓	MR-NIRP-Car RMSE↓	MR-NIRP-Car ρ↑
✓	-	-	4.03	6.77	0.82	6.59	8.46	0.29
-	✓	-	3.98	6.55	0.84	6.24	7.62	0.32
-	-	✓	4.10	6.95	0.82	6.81	8.03	0.26
✓	✓	-	3.91	6.42	0.85	6.23	7.11	0.38
✓	-	✓	4.11	6.71	0.84	6.67	7.93	0.31
✓	✓	✓	**3.79**	**6.34**	**0.86**	**5.75**	**6.08**	**0.54**

Table 5: Generality of physiological frequency denoiser (PFD) module.

Method	VIPL-HR MAE↓	VIPL-HR RMSE↓	VIPL-HR ρ↑	MR-NIRP-Car MAE↓	MR-NIRP-Car RMSE↓	MR-NIRP-Car ρ↑
RhythmMamba [60]	4.30	7.49	0.81	8.07	10.03	0.24
RhythmMamba [60]+**PFD (Ours)**	**4.15**	**7.13**	**0.82**	**7.88**	**9.21**	**0.29**
PhysDiff [32]	3.92	6.65	0.85	7.40	8.39	0.31
PhysDiff [32]+**PFD (Ours)**	**3.84**	**6.44**	**0.86**	**7.04**	**8.13**	**0.34**
FreqPhys w/o **PFD (Ours)**	4.22	7.17	0.83	5.93	6.55	0.46
FreqPhys	**3.79**	**6.34**	**0.86**	**5.75**	**6.08**	**0.54**

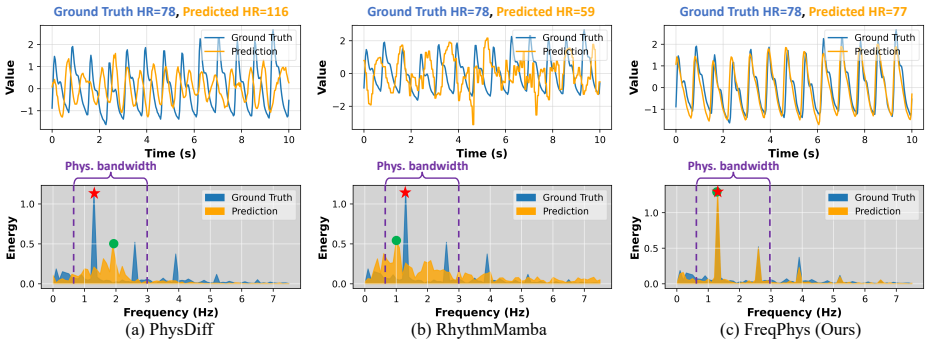
from a relatively simple source domain (*e.g.*, PURE or UBFC-rPPG) to more complex target domains such as MMPD. In contrast, benefiting from physiological spectrum modeling, our method effectively captures domain-invariant physiological priors and consistently achieves the best performance across all transfer settings, demonstrating stronger robustness and generalization capability under substantial cross-dataset domain shifts.

More extended cross-dataset testing results are provided in **Appendix G**.

4.4 Ablation Studies.

How effective are the components of the physiological frequency denoiser? To investigate the contribution of each component in the proposed physiological frequency denoiser, we conduct detailed ablation studies on the three modules: *Physiological Bandpass Filter* (PBF), *Physiological Spectrum Modulation* (PSM), and *Adaptive Spectrum Selection* (ASS). As shown in Tab. 4, each module improves the performance, and their combination yields the best results, indicating that these components work synergistically to suppress spectral noise and enhance physiologically meaningful frequency representations. Furthermore, as reported in Tab. 5, when the physiological frequency denoiser module is removed and only the time-domain MSTmap is used as the conditional input, the performance drops noticeably, further demonstrating the importance of incorporating physiological frequency priors for robust rPPG estimation.

How general is the physiological frequency denoiser? In addition to analyzing the contribution of individual components, we further investigate the generality of the proposed *Physiological Frequency Denoiser* (PFD). It is designed as a modular plug-in that can be seamlessly integrated into existing rPPG architectures. As shown in Tab. 5, we evaluate the complete PFD module (PBF+PSM+ASS) when incorporated into two state-of-the-art models, RhythmMamba [60] and PhysDiff [32]. Specifically, PFD is inserted before each Multi-temporal Constraint Mamba block in RhythmMamba and before each Spatial–Temporal Hybrid Attention block in PhysDiff, while keeping all other settings unchanged. Across all evaluation metrics on the VIPL-HR dataset, both backbones achieve consistent and notable performance gains after integrating PFD. These results demonstrate the strong generality of PFD, indicating that it can be seamlessly integrated into diverse backbone architectures while providing cleaner and more physiologically meaningful spectral clues for rPPG estimation.



(a) PhysDiff (b) RhythmMamba (c) FreqPhys (Ours)

Fig. 5: Time and frequency domain visualizations of rPPG signal predictions on the VIPL dataset under head motion scenario. In the frequency-domain plots, the purple dashed box indicates the physiological signal bandwidth ranging from 0.66 to 3.0 Hz, corresponding to typical human cardiac frequencies. ⋆ and ● denote the spectral peaks of the ground-truth and predicted heart rates, respectively.

4.5 Computational Cost.

To evaluate the deployment efficiency of our method, we conduct a unified 10-second inference benchmark on a single NVIDIA RTX 4090 GPU (24GB). The detailed computational statistics are reported in Tab. 6. Notably, compared with the lightweight RhythmMamba, our method consistently achieves lower computational overhead across all key metrics. Specifically, our **FreqPhys** reduces the parameters from 2.00M to 0.86M (57.0% reduction) and decreases FLOPs from 12.41G to 7.75G (37.6% reduction). In addition, our method significantly improves runtime efficiency, increasing throughput from 27.16 Kps to 85.44 Kps (3.1× faster) and reducing inference latency from 36.82 ms to 11.70 ms. Meanwhile, the peak GPU memory usage is reduced from 2450 MB to 874 MB (64.3% reduction).

Table 6: Computational cost. Best results are marked in bold and the second best in underline.

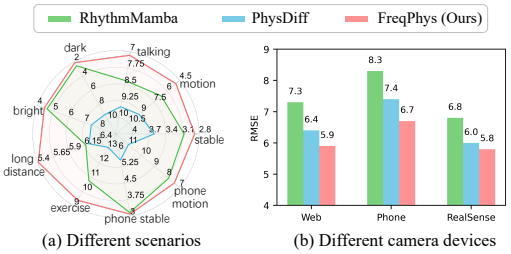
Method	Parameters (M)↓	Flops (G)↓	Throughput (Kps)↑	Inference time (ms)↓	Memory (M)↓
DeepPhys [3]	1.98	111.67	28.89	34.61	10638
PhysNet [53]	**0.77**	65.74	<u>68.73</u>	<u>14.55</u>	3750
TS-CAN [18]	1.98	111.67	26.23	38.13	11834
PhysFormer [55]	7.38	47.44	50.79	19.69	6480
EfficientPhys [19]	1.91	56.06	41.36	24.18	7814
RhythmMamba [60]	2.00	<u>12.41</u>	27.16	36.82	2450
PhysDiff [32]	2.64	22.46	60.23	16.60	<u>1246</u>
FreqPhys (ours)	<u>0.86</u>	**7.75**	**85.44**	**11.70**	**874**

4.6 Qualitative Analysis

Visualization of rPPG Prediction. We present a prediction showcase on the VIPL dataset under the head motion scenario, as shown in Fig. 5. PhysDiff [32], RhythmMamba [60] are selected as the representative SOTA methods. We observe that the rPPG signal predicted by PhysDiff and RhythmMamba exhibits numerous burrs, which reflect the limitations of modeling in the time domain, namely, the difficulty in capturing dominant physiological frequency components and being vulnerable to noise. Our **FreqPhys** addresses this limitation effectively, which not only keeps pace with the label sequence but also

accurately exhibits a smoother appearance with fewer irregularities. In addition, we also observe that the frequency spectra of PhysDiff and RhythmMamba exhibit relatively dispersed energy distributions. In contrast, our method effectively concentrates spectral energy on several key frequency components corresponding to physiological signals, resulting in an obvious sparsity in the frequency domain. This focused energy distribution highlights our method's ability to isolate and enhance vital physiological components while suppressing irrelevant noise.

Qualitative Results for Robustness. To evaluate the robustness of our model across diverse scenarios and camera device conditions, we provide detailed results from the VIPL-HR dataset, encompassing 9 distinct scenarios and 3 types of camera devices. As illustrated in Fig. 6, our proposed method consistently outperforms other methods across these varied conditions, which indicates its robustness in remote physiological signal measurement.



(a) Different scenarios (b) Different camera devices

Fig. 6: HR estimation results on VIPL-HR under different scenarios and camera devices.

5 Conclusion

In this paper, we introduced **FreqPhys**, a diffusion-based framework that leverages implicit physiological frequency priors for robust remote photoplethysmography. Unlike existing approaches that rely primarily on time-domain modeling or incorporate spectral clues only at the loss or representation level, our method integrates physiological spectral priors directly into the iterative diffusion denoising process. This design enables more effective suppression of motion-induced artifacts while preserving meaningful cardiac rhythms. Extensive experiments on six public datasets demonstrate that our method achieves state-of-the-art performance in HR, HRV, and RF estimation, while also exhibiting strong cross-dataset generalization.

Acknowledgment We thank the anonymous reviewers for their constructive feedback. We would also like to thank Dr. Kun Li for his valuable suggestions. This work is supported by Natural Science Foundation of China (72188101, 62272144), National Key R&D Program of China (NO.2024YFB3311602), the Anhui Provincial Natural Science Foundation (2408085J040), and the Major Project of Anhui Provincial Science and Technology Breakthrough Program (202423k09020001), the Fundamental Research Funds for the Central Universities (JZ2024AHST0337), and the New Cornerstone Science Foundation through the XPLORER PRIZE.

References

1. Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., Dubois, J.: Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognition Letters **124**, 82–90 (2019)
2. Chen, S., Wong, K.L., Chin, J.W., Chan, T.T., So, R.H.: Diffphys: Enhancing signal-to-noise ratio in remote photoplethysmography signal using a diffusion model approach. Bioengineering **11**(8), 743 (2024)
3. Chen, W., McDuff, D.: Deepphys: Video-based physiological measurement using convolutional attention networks. In: Proceedings of the European Conference on Computer Vision. pp. 349–365 (2018)
4. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. IEEE Transactions on Biomedical Engineering **60**(10), 2878–2886 (2013)
5. Fons, E., Sztrajman, A., El-Laham, Y., Iosifidis, A., Vyetrenko, S.: Hypertime: Implicit neural representation for time series. arXiv preprint arXiv:2208.05836 (2022)
6. Gedam, S., Paul, S.: 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)
7. Gideon, J., Stent, S.: The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3995–4004 (2021)
8. Guo, D., Li, K., Hu, B., Zhang, Y., Wang, M.: Benchmarking micro-action recognition: Dataset, methods, and applications. IEEE Transactions on Circuits and Systems for Video Technology **34**(7), 6238–6252 (2024)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
10. Huang, B., Hu, S., Liu, Z., Lin, C.L., Su, J., Zhao, C., Wang, L., Wang, W.: Challenges and prospects of visual contactless physiological monitoring in clinical study. NPJ Digital Medicine **6**(1), 231 (2023)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. Li, K., Gu, J., Wang, F., Wu, Z., Fan, H., Guo, D.: Ma-bench: Towards fine-grained micro-action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20118–20128 (2026)
13. Li, K., Liu, P., Guo, D., Wang, F., Wu, Z., Fan, H., Wang, M.: Mmad: Multi-label micro-action detection in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13225–13236 (2025)
14. Li, Q., Guo, D., Qian, W., Tian, X., Sun, X., Zhao, H., Wang, M.: Channel-wise interactive learning for remote heart rate estimation from facial video. IEEE Transactions on Circuits and Systems for Video Technology **34**(6), 4542–4555 (2023)
15. Li, Q., Guo, D., Qian, W., Zhou, Y., Sun, X., Wang, M.: Temporal gated face alignment network for camera-based physiological sensing. IEEE Transactions on Computational Social Systems (2025)
16. Li, Q., Qian, W., Guo, D., Wang, M.: Frepl: Frequency-guided pseudo-labeling for semi-supervised remote physiological measurement. Machine Intelligence Research **23**(2), 467–483 (2026)
17. Li, Z., Yin, L.: Contactless pulse estimation leveraging pseudo labels and self-supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20588–20597 (2023)

18. Liu, X., Fromm, J., Patel, S., McDuff, D.: Multi-task temporal shift attention networks for on-device contactless vitals measurement. In: Advances in Neural Information Processing Systems. vol. 33, pp. 19400–19411 (2020)
19. Liu, X., Hill, B., Jiang, Z., Patel, S., McDuff, D.: Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5008–5017 (2023)
20. Lu, H., Han, H., Zhou, S.K.: Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12404–12413 (2021)
21. Luo, C., Xie, Y., Yu, Z.: Physmamba: Efficient remote physiological measurement with slowfast temporal difference mamba. In: Chinese Conference on Biometric Recognition. pp. 248–259. Springer (2024)
22. 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 (2019)
23. Niu, X., Yu, Z., Han, H., Li, X., Shan, S., Zhao, G.: Video-based remote physiological measurement via cross-verified feature disentangling. In: Proceedings of the European Conference on Computer Vision. pp. 295–310 (2020)
24. Niu, X., Zhao, X., Han, H., Das, A., Dantcheva, A., Shan, S., Chen, X.: Robust remote heart rate estimation from face utilizing spatial-temporal attention. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition. pp. 1–8 (2019)
25. Niu, Z., Tu, X., Zhao, B., Cao, J., Guo, D., Yu, Z.: Intervention-based self-supervised learning: A causal probe paradigm for remote photoplethysmography. arXiv preprint arXiv:2605.00882 (2026)
26. Nowara, E.M., Marks, T.K., Mansour, H., Veeraraghavan, A.: Near-infrared imaging photoplethysmography during driving. IEEE transactions on intelligent transportation systems **23**(4), 3589–3600 (2020)
27. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Optics express **18**(10), 10762–10774 (2010)
28. Qian, W., Guo, D., Li, K., Zhang, X., Tian, X., Yang, X., Wang, M.: Dual-path tokenlearner for remote photoplethysmography-based physiological measurement with facial videos. IEEE Transactions on Computational Social Systems (2024)
29. Qian, W., Li, K., Guo, D., Hu, B., Wang, M.: Cluster-phys: Facial clues clustering towards efficient remote physiological measurement. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 330–339 (2024)
30. Qian, W., Li, Q., Li, K., Wang, X., Sun, X., Wang, M., Guo, D.: Joint spatial-temporal modeling and contrastive learning for self-supervised heart rate measurement. arXiv preprint arXiv:2406.04942 (2024)
31. Qian, W., Li, Z., Ding, Y., Jia, X., Su, G., Wang, F., Li, K., Wang, M.: Rethinking rppg supervision: Distribution and temporal relation alignment for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1338–1348 (2026)
32. Qian, W., Su, G., Guo, D., Zhou, J., Li, X., Hu, B., Tang, S., Wang, M.: Physdiff: Physiology-based dynamicity disentangled diffusion model for remote physiological measurement. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2025)
33. Qian, W., Su, G., Li, K., Ding, Y., Jia, X., Guo, D.: Diffrepss: A diffusion model for remote physiological signal sensing (2025)

34. Shao, H., Luo, L., Qian, J., Yan, M., Chen, S., Yang, J.: Remote photoplethysmography in real-world and extreme lighting scenarios. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10858–10867 (2025)
35. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations. pp. 1–20 (2021)
36. Speth, J., Vance, N., Flynn, P., Czajka, A.: Non-contrastive unsupervised learning of physiological signals from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14464–14474 (2023)
37. Stricker, R., Müller, S., Gross, H.M.: Non-contact video-based pulse rate measurement on a mobile service robot. In: Proc. IISRHIC. pp. 1056–1062 (2014)
38. Su, G., Qian, W., Li, Q., Wu, Y., Feng, W., Guo, D.: Msphys: multiscale fusing-based diffusion model for remote physiological measurement. Machine Vision and Applications **36**(5), 105 (2025)
39. Sun, Z., Li, X.: Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In: Proceedings of the European Conference on Computer Vision. pp. 492–510 (2022)
40. 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 (2024)
41. Tang, J., Chen, K., Wang, Y., Shi, Y., Patel, S., McDuff, D., Liu, X.: Mmpd: multi-domain mobile video physiology dataset. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 1–5. IEEE (2023)
42. Tu, X., Niu, Z., Yin, J., Zhang, Y., Yang, M., Wei, L., Wang, Y., Fan, Z., Zhao, J.: Phys-edigan: A privacy-preserving method for editing physiological signals in facial videos. Pattern Recognition **169**, 111966 (2026)
43. Verkruysse, W.e.a.: Remote plethysmographic imaging using ambient light. Optics Express **16**(26), 21434–21445 (2008)
44. Wang, F., Guo, D., Li, K., Wang, M.: Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 38, pp. 5345–5353 (2024)
45. Wang, F., Guo, D., Li, K., Zhong, Z., Wang, M.: Frequency decoupling for motion magnification via multi-level isomorphic architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18984–18994 (2024)
46. Wang, F., Yang, J., Chen, J., Liu, Y., Li, K., Wei, Y., Guo, D., Wang, M.: Xinsight: Integrative stage-consistent psychological counseling support agents for digital well-being. In: Proceedings of the ACM Web Conference 2026. pp. 9297–9308 (2026)
47. Wang, M., Liu, Z., Li, K., Wang, Y., Wang, Y., Wei, Y., Wang, F.: Task-generalized adaptive cross-domain learning for multimodal image fusion. TMM 2025 (2025)
48. Wang, T., Lin, X., Xu, Y., Ye, Q., Guo, D., Escalera, S., Khoriba, G., Yu, Z.: Micro-gesture recognition: A comprehensive survey of datasets, methods, and challenges. Machine Intelligence Research **23**(2), 308–330 (2026)
49. Wang, W., Den Brinker, A.C., Stuijk, S., De Haan, G.: Algorithmic principles of remote ppg. IEEE Transactions on Biomedical Engineering **64**, 1479–1491 (2016)
50. Xi, L., Chen, W., Zhao, C., Wu, X., Wang, J.: 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)

51. Xie, Y., Zhao, B., Dai, M., Zhou, J.P., Sun, Y., Tan, T., Xie, W., Shen, L., Yu, Z.: Physllm: Harnessing large language models for cross-modal remote physiological sensing. arXiv preprint arXiv:2505.03621 (2025)
52. Yu, Z., Li, X., Wang, P., Zhao, G.: Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection. IEEE Signal Processing Letters **28**, 1290–1294 (2021)
53. Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In: Proceedings of the British Machine Vision Conference. pp. 1–12 (2019)
54. Yu, Z., Shen, Y., Shi, J., Zhao, H., Cui, Y., Zhang, J., Torr, P., Zhao, G.: Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. International Journal of Computer Vision **131**(6), 1307–1330 (2023)
55. Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P., Zhao, G.: 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)
56. 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)
57. Zhang, J., Lin, X., Huang, J., Ye, S., Guo, X., Zhu, D., Hu, R., Guo, D., Liang, Y., Yu, Z., et al.: Multimodal deception detection: A survey. Machine Intelligence Research **23**(2), 284–307 (2026)
58. Zhao, B., Guo, D., Cao, J., Xu, Y., Tan, T., Sun, Y., Zou, B., Zhang, J., Yu, Z.: Phase-net: Physics-grounded harmonic attention system for efficient remote photoplethysmography measurement. arXiv preprint arXiv:2509.24850 (2025)
59. 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)
60. Zou, B., Guo, Z., Hu, X., Ma, H.: Rhythmmamba: Fast remote physiological measurement with arbitrary length videos. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2025)