HRVFusion: Video-based Long-Term Heart Rate Variability Measurement with Conditional Diffusion Models

Xing Yao, Rencheng Song, *Senior Member, IEEE,* Juan Cheng, *Member, IEEE,* Chang Li, *Member, IEEE,* and Xun Chen,*Senior Member,IEEE*

***Abstract*—Remote photoplethysmography (rPPG) has shown great potential for heart rate variability (HRV) analysis due to its non-contact and convenient nature. However, HRV measurement relies on long-term and high-quality blood volume pulse (BVP) signals, which pose substantial challenges for existing rPPG techniques. To address this, we propose HRVFusion, a conditional diffusion model-based framework for HRV estimation from rPPG signals. In the forward diffusion process, raw chrominance (CHROM) signals extracted from a video are incorporated to simulate the complex noise disturbances typically observed in rPPG signals. During the reverse process, to enhance the model's sensitivity to HRVs, the high-precision BVP reconstruction is jointly guided using both the chrominance signal and its time-frequency ridge extracted via the wavelet synchrosqueezed transform (WSST). Furthermore, a convolution and mamba hybrid network is introduced to handle inputs of arbitrary length, meeting the requirements for long-term HRV analysis. A public dataset for long-term video-based HRV extraction, consisting of 15-minute recordings from 18 participants, is presented for the first time. Experimental results demonstrate that the proposed method achieves superior performance across multiple HRV metrics, with a mean absolute error (MAE) of 5–8 ms for SDNN, RMSSD below 18 ms, and Pearson correlation coefficients above 0.85 for frequency-domain indices (LF, HF, LF/HF), outperforming existing approaches. This study provides an accurate and robust non-contact solution for long-term video-based HRV measurement.**

***Index Terms*—Conditional diffusion model, heart rate variability, human-computer interaction, remote photoplethysmography, wavelet synchrosqueezed transform**

I. Introduction

HEART rate variability (HRV) is an important physiological indicator that reflects autonomic nervous system regulation and is widely used in scenarios such as stress assessment [1], emotion recognition [2], and cardiovascular risk monitoring [3]. In particular, long-term and continuous HRV analysis is considered to offer higher physiological relevance and greater clinical value in health management and disease prediction [4]. Traditional HRV measurement relies on contact electrodes, which requires professional operation, and may cause discomfort during long-term monitoring. In recent years, remote photoplethysmography (rPPG) [5] has emerged as a new approach for contactless HRV monitoring by capturing blood volume pulse (BVP) signals from facial videos. HRV analysis relies on high-quality inter-beat interval (IBI) sequences, which places stringent requirements on the integrity and stability of the BVP waveforms. However, rPPG signals are highly susceptible to disturbances such as changes in ambient lighting, facial motion artifacts, and individual skin tone differences. These factors lead to instability in the quality of the extracted BVP waveforms, thereby affecting the accuracy of HRV metrics.

Conventional rPPG algorithms, such as blind source separation methods [6] and model-based approaches [7], [8] have achieved great progress in average heart rate estimation. However, these methods still face limitations in preserving BVP waveform fidelity and struggle to stably extract high-quality IBI sequences. In recent years, deep learning (DL) based rPPG methods [9]–[11] have significantly improved BVP signal quality thanks to its powerful nonlinear modeling capabilities. However, most existing deep learning methods remain focused on average heart rate estimation or HRV estimation within short time windows, lacking the ability to model long-term BVP signal. This limitation is further compounded by the publicly available rPPG datasets. Currently, most public datasets contain videos of no more than five minutes [12]. In related studies, HRV is typically evaluated using short test windows of approximately 30 seconds [13]. However, robust assessment of HRV generally requires continuous inter-beat interval (IBI) sequences of at least five minutes to accurately capture autonomic nervous system regulation [14]. Existing DL methods are mostly trained and evaluated on these short videos, and some studies attempt to approximate long-term HRV by concatenating multiple short segments during inference [15]. Nevertheless, according to HRV evaluation guidelines [16], such discontinuous IBI sequences are prone to rhythm interruptions and loss of inter-beat interval waveform features, which can lead to waveform distortion and compromise the accuracy and physiological interpretability of HRV metrics. As a result, they fall short of meeting the requirements for continuous HRV monitoring in real-world health applications.

Since rPPG signal extraction can be regarded as a reconstruction problem of complex physiological waveforms. In recent years, rPPG methods based on generative models have gradually attracted attention. Among them, generative adversarial networks (GANs) have achieved initial success in various rPPG tasks [17], [18], demonstrating promising generative capacity and adaptability. However, GANs suffer from issues such as training instability and mode collapse. In contrast, diffusion models have emerged as a promising new direction for rPPG signal reconstruction due to their stable training process and powerful modeling of target distributions. These models progressively restore signals through multi-step

X. Yao, R. Song, J. Cheng, and C. Li are with the Department of Biomedical Engineering, and also with the Anhui Province Key Laboratory of Measuring Theory and Precision Instrument, Hefei University of Technology, Hefei 230009, China (e-mail: 2023110087@mail.hfut.edu.cn; rcsong@hfut.edu.cn; chengjuan@hfut.edu.cn; changli@hfut.edu.cn).

X. Chen is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China. E-mail: xunchen@ustc.edu.cn.

iterations, maintaining high fidelity during reconstruction. In particular, their flexible conditional modeling capabilities open up new possibilities for developing rPPG enhancement methods tailored for HRV analysis. However, existing diffusion-based rPPG approaches typically assume the noise in the diffusion process as a simple Gaussian distribution, which fails to capture the non-Gaussian, non-stationary, and structurally complex disturbance patterns commonly present in rPPG signals. This mismatch limits the realism and generalization ability of the generated results.

To address the above challenges, this study proposes HRV-Fusion, a conditional diffusion model, which is able to meet the requirements of long-term video-based HRV analysis under complex disturbances. In the forward diffusion phase, we design a conditional diffusion process that incorporates both the chrominance signal [8] and Gaussian noise into the clean PPG signal, making the noising process more reflective of the interference characteristics commonly observed in real-world rPPG signals. During the reverse denoising stage, the CHROM waveform and its time-frequency ridge extracted by wavelet synchrosqueezing transform (WSST) [19] are both employed as conditions to enforce the model to capture dynamic inter-beat variations. Additionally, we propose a convolution and Mamba hybrid network, where 1-D convolutions are used to enable the model to process rPPG signals of arbitrary duration and supporting long-term HRV analysis tasks. Meanwhile, Mamba blocks are taken to effectively capture long-range temporal dependencies in physiological signals with their state space modeling capability. The main contributions are summarized as follows:

1) By mixing Gaussian noise and chrominance signals during the diffusion process, the HRVFusion can handle non-Gaussian noise such as head movement and uneven lighting in rPPG measurement, thereby improving the model's robustness and generalization ability in complex noise environments.
2) By incorporating CHROM and its time-frequency ridge as guidance during the denoising process, the HRVFusion model aligns closely with interbeat rhythm dynamics, thereby enhancing the consistency of the generated BVP signals with PPG.
3) Benefiting from the proposed hybrid network architecture, the HRVFusion model supports direct long-term waveform prediction after being trained on short fixed-length samples, without the need for retraining. Therefore, it supports continuous and stable BVP signal extraction for long-term HRV analysis.

The rest of this paper is organized as follows. Section II reviews the related work. Section III introduces the proposed method. Sections IV and V describe the experimental setup and result analysis, respectively. Finally, Section VI summarizes the work of this paper.

II. RELATED WORK

HRV analysis highly depends on stable and faithful BVP waveforms to ensure accurate extraction of inter-beat intervals. To improve rPPG signal quality, related research primarily focuses on three main directions, including traditional signal processing methods, network architecture optimization, and generative modeling. This paper will review these aspects accordingly.

A. Signal Processing-driven RPPG Modeling

Traditional rPPG methods based on signal processing can be broadly categorized into three technical directions, statistical approaches, optical reflection models, and time-frequency analysis.

Statistical approaches perform signal separation based on prior assumptions, such as signal independence or correlation. For example, Poh et al. [6] employed independent component analysis (ICA) to effectively decompose RGB channels and extract BVP signals, while Lewandowska et al. [20] applied principal component analysis (PCA) to enhance signal stability. These approaches leverage statistical properties of the input channels to isolate the underlying pulsatile component while suppressing noise. Nevertheless, their performance is highly dependent on the validity of the assumed statistical priors, and they often struggle with severe motion artifacts or heterogeneous illumination.

Optical reflection model–based methods build physical models linking skin optical properties to blood flow variations and design specific signal combination strategies to suppress non-physiological interference and enhance pulsatile components. Typical methods include CHROM [8], plane-orthogonal-to-skin (POS) [7] and the spatial subspace rotation (2SR) [21]. These methods exploit the intrinsic redundancy among RGB channels to construct orthogonal or rotated subspaces that maximize the blood volume pulse while minimizing noise induced by motion and illumination changes. Despite their simplicity and interpretability, their performance heavily depends on carefully designed linear combinations and may degrade under uncontrolled scenarios.

Time-frequency analysis techniques extract BVP signals by utilizing their time-frequency-domain characteristics. These methods typically employ band-pass filters, wavelet transforms, or short-time Fourier transforms to isolate frequency components associated with cardiac activity while attenuating noise from motion or illumination variations. By leveraging the non-stationary nature of rPPG signals, they enable dynamic tracking of instantaneous heart rate [22] and its variability. Chwyl et al. [23] combined pulselet wavelets with continuous wavelet transform to track instantaneous heart rate frame by frame and reconstructed PPG signals using Bayesian methods. However, their effectiveness is often limited by the choice of window size, frequency resolution, and sensitivity to non-periodic disturbances.

Although these approaches perform reasonably well in static or low-interference scenarios, they face challenges in high-quality, long-term HRV measurement. Variations in illumination, facial motion, and individual differences can degrade signal quality, while non-periodic noise and amplitude distortions significantly affect HRV metrics and its physiological interpretability.

B. Network Architecture-driven RPPG Modeling

Network architecture optimization primarily enhances the model's temporal modeling capability to improve the continuity and fidelity of BVP waveforms. Related studies mostly focus on the design of convolutional and attention mechanisms. Early deep learning-based rPPG approaches mostly relied on convolutional architectures to model the non-contact pulse features. For example, DeepPhys, proposed by Chen et al. [24], achieves stable BVP waveform extraction through spatial attention and temporal modeling. EfficientPhys [25] combines self-attention and channel shuffling to accurately reconstruct continuous pulse waveforms without the need for preprocessing. MTTS-CAN [26] further enhances waveform smoothness by introducing temporal shift modules and a multi-task structure.

On another front, transformer architectures have also been applied to rPPG modeling to boost waveform quality. Yu et al. [27] introduced PhysFormer, which leverages temporal difference attention to capture subtle heart rate features and improve the waveform representation. Recently, Zou et al. [10] proposed RhythmFormer to incorporate rhythm priors that guide temporal modeling and further optimize waveform quality.

However, the aforementioned methods mostly focus on short-term signal modeling and have yet to meet the practical demands for long-term, high-quality HRV analysis in clinical settings [21]. Given the limitations of discriminative models in waveform consistency and temporal modeling, recent studies have begun exploring the use of generative models to better support long-term rPPG signal reconstruction.

C. Generative Model-driven RPPG Modeling

Extracting high-quality BVP signals from video can be regarded as a generative task. In recent years, generative model-driven rPPG modeling has demonstrated significant potentials. Early work used GANs to generate BVP waveforms. For example, Song et al. [17] proposed PulseGAN, which leverages a discriminator to guide waveform reconstruction and jointly constrains the time and frequency domains to generate BVP signals with better physiological consistency. Lu et al. [18] designed Dual-GAN, which jointly models the distributions of signals and noise, demonstrating stronger robustness in dynamic environments.

Diffusion models [28], as more advanced generative models, have also been applied to rPPG tasks. For example, Chen et al. [29] proposed DiffPhys, which uses chrominance signals as reconstruction conditions to enhance the signal-to-noise ratio of rPPG signals. Jeong et al. [30] introduced Diffusion-Phys, applying diffusion models to denoise and reconstruct MSTmap, improving waveform quality. Qian et al. [31] developed PhysDiff, which combines diffusion models with decoupled dynamic characteristics to optimize waveform structure and detail. Although the above Diffusion-based rPPG methods have made significant progress in waveform generation quality, most of them [29]–[31] assume isotropic Gaussian noise during the diffusion process [32]. However, motion artifacts and illumination variations in real-world rPPG signals often manifest as non-Gaussian noise, which limits their modeling capability in complex environments.

To address these issues, we propose the HRVFusion framework, which incorporates a mixed noising process combining chrominance signals and Gaussian noise. This approach better reflects the actual interference characteristics in rPPG signals, thereby enhancing the model's robustness and stability in complex noise environments.

III. METHODOLOGY

In this section, we will introduce the details of the rPPG signal reconstruction framework HRVFusion based on the conditional diffusion model. The overall framework of HRVFusion is shown in Fig. 1, and it consists of three main components. (a) BVP signal preprocessing, (b) high-quality BVP signal reconstruction with conditional diffusion model, and (c) convolution and mamba hybrid network structure.

A. BVP Signal Preprocessing

As shown in Fig. 1(a), the HRVFusion first extracts the initial BVP signal from facial videos. The MediaPipe [33] is employed to track 145 facial landmarks to define multiple regions of interest (ROIs) that cover the entire face. The CHROM algorithm is then applied to extract raw BVP signals in each ROI. All BVP signals are then averaged with equal weights to get a fused BVP signal. It should be noted that adaptively weighted fusion may be more effective in scenarios where there are large signal differences in each ROI. A detrending and a sixth-order Butterworth bandpass filter with a frequency range of 0.65-2.5 Hz are further applied to further improve the signal-to-noise ratio (SNR). This processing pipeline ultimately generates the fused chrominance signal y, which is required for the conditional diffusion model.

B. High-quality BVP Signal Reconstruction with Conditional Diffusion Model

The second part of HRVFusion focuses on high-quality BVP signal reconstruction, with its overall framework illustrated in Fig. 1(b). Traditional diffusion models typically assume isotropic Gaussian noise during both the forward diffusion and reverse denoising processes. However, this assumption fails to capture the complex and variable noise characteristics present in rPPG signals. To improve adaptability to real-world environments, HRVFusion incorporates both the chrominance signal extracted from facial video and Gaussian noise into the clean PPG signal during the forward diffusion process, thereby more realistically simulating the types of disturbances that rPPG signals may encounter in practical scenarios.

During the reverse denoising process, we use the waveform information of the chrominance signal and its time-frequency ridge information extracted through the WSST as conditional guidance, jointly guiding the model from both time and frequency domains to generate high-quality rPPG signals. The entire T-step diffusion model consists of two main processes, a conditional diffusion process (with steps $t \in \{0, 1, \ldots, T\}$) and a conditional reverse process (with steps $t \in \{T, T-1, \ldots, 0\}$).

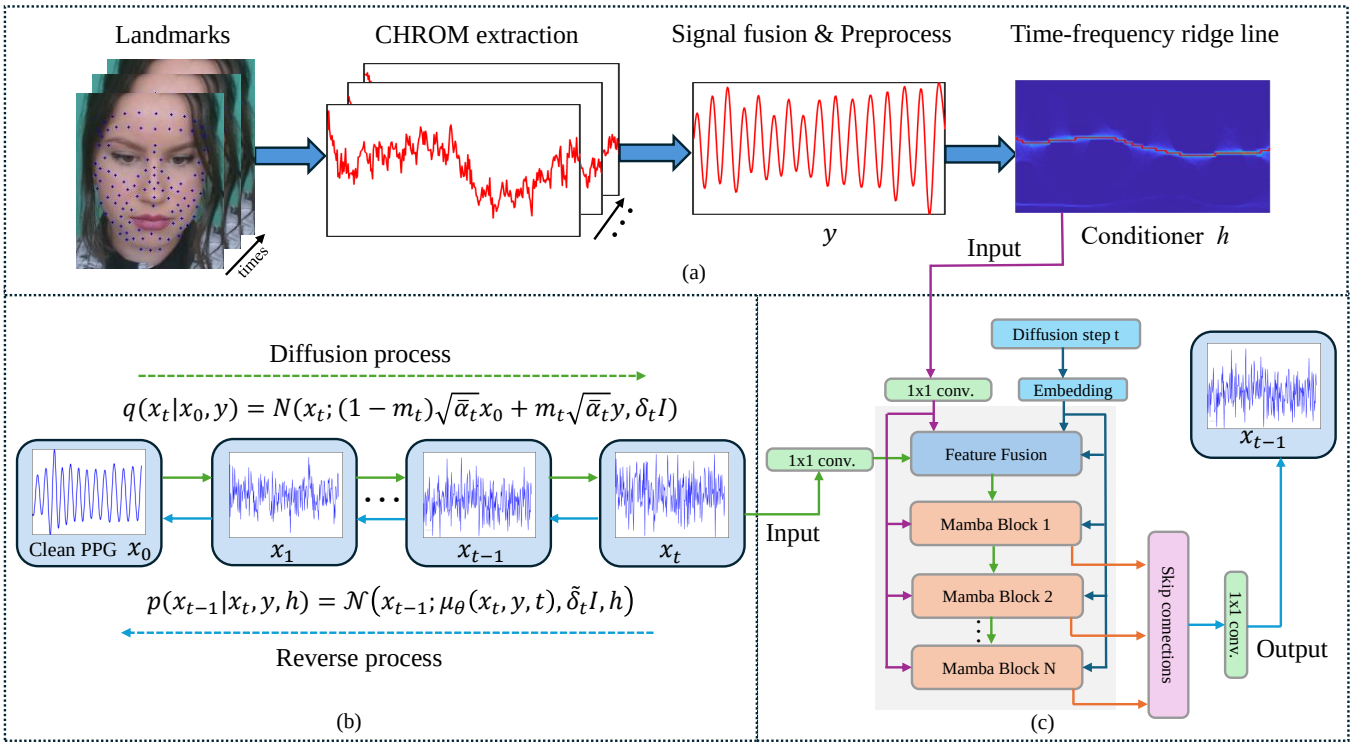

Fig. 1. Flowchart of the proposed HRVFusion method. (a) BVP signal preprocessing, where the chrominance signal y and its time-frequency ridge h are extracted from facial videos and used as input and conditional information for subsequent modules. (b) High-quality BVP signal reconstruction by conditional diffusion model, which includes the diffusion process (incorporating not only Gaussian noise but also non-Gaussian noise from the chrominance signal y), and the reverse process; (c) Convolution and Mamba hybrid network structure, which progressively predicts noise residuals during the reverse process and uses the chrominance signal y and its time-frequency ridge h as conditional guidance for reconstruction.

1) Conditional Diffusion Process: The diffusion process starts from a reference PPG signal and simulates the gradual corruption of the signal by progressively introducing noise. Due to the fact that rPPG signals in practical applications are often disturbed by various complex noises such as illumination changes and head movements, which are diverse and exhibit non-Gaussian characteristics, it is difficult to accurately model them using traditional noise injection methods. Therefore, to more realistically simulate the gradual contamination of signals in real-world environments, we first define the following conditional diffusion process

$$q_{cond}(x_t|x_0,y) = \mathcal{N}\left(x_t;(1-m_t)\sqrt{\bar{\alpha}_t}x_0 + m_t\sqrt{\bar{\alpha}_t}y, \delta_t I\right). \tag{1}$$

Here x_t denotes the noisy signal at time step t, x_0 is the clean reference PPG signal, and y represents the chrominance signal, and m_t is an interpolation function at time step t ($0 \le m_t \le 1$), defined as $m_t = \sqrt{(1-\bar{\alpha}_t)/\sqrt{\bar{\alpha}_t}}$, which dynamically balances the feature contributions of x_0 and y. Here, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{r=1}^{t} \alpha_r$, and β_t follows a linear noise schedule [34], increasing from 1×10^{-4} to 0.035 over $T = 50$ diffusion steps. $\bar{\alpha}_t$ is the cumulative product coefficient in the diffusion process, and $\delta_t = (1-\bar{\alpha}_t) - m_t^2\bar{\alpha}_t$ is the variance parameter at time step t, which ensures that Gaussian noise is added at each step. I denotes the identity matrix. As the time step t increases, m_t gradually approaches 1, causing x_t to transition from a purely Gaussian noise state to a mixed state that incorporates features of the chrominance signal. The state at the final diffusion step T can be expressed as

$$p_{\text{cond}}(x_T \mid y) = \mathcal{N}(x_T, \sqrt{\bar{\alpha}_T}y, \delta_T I). \tag{2}$$

This implies that at timestep T, the signal x_T follows a Gaussian distribution with mean $\sqrt{\bar{\alpha}_T}y$ and variance $\delta_T I$.

2) Conditional Reverse Process: The reverse process starts from x_T and gradually recovers the original signal through a conditional guidance mechanism. We define the following conditional distribution

$$P_{\text{cond}}(x_{t-1} \mid x_t, y, h) = N(x_{t-1}; u_\theta(x_t, y, t), \tilde{\delta}_t I, h), \tag{3}$$

where $\tilde{\delta}_t$ represents the variance term at time step t in the reverse process, and h denotes the time-frequency ridge extracted from the fused chrominance signal y using WSST. The use of y and h at each time step t aims to guide the high-quality BVP signal reconstruction. As shown in Fig. 2, the extracted red ridge line h by WSST demonstrates dominant energy trajectory of the pulse signal in the time-frequency domain. Since instantaneous frequency reflects interbeat dynamics, the ridge h serves as prior guidance to help the model generate signals with improved physiological consistency and accuracy.

The predicted mean of the distribution, denoted as $\mu_\theta(x_t, y, t)$ and parameterized by the learnable network pa-

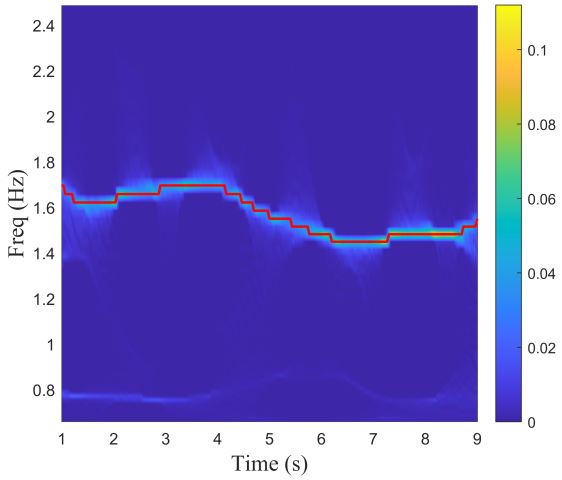

Fig. 2. The time-frequency ridge line (red) of the fused chrominance signal based on WSST.

rameters θ, can be expressed through the following linear combination

$$\mu_\theta(x_t, y, t) = c_{xt}x_t + c_{yt}y - c_{\epsilon t}\epsilon_\theta(x_t, y, t), \tag{4}$$

where c_{xt}, c_{yt}, and $c_{\epsilon t}$ are the weighting coefficients corresponding to x_t, y, and the neural network predicted noise $\epsilon_\theta(x_t, y, t)$, respectively. These coefficients are derived based on the evidence lower bound (ELBO) optimization criterion.

The model is trained based on the ELBO theory by minimizing the following loss function

$$\text{Loss} = c' + \sum_{t=1}^{T} \kappa_t' \mathbb{E}_{x_0,\epsilon,y} \left\| \omega_t - \epsilon_\theta(x_t, y, t) \right\|_2^2, \tag{5}$$

where c' and κ_t' are constant terms, and $\epsilon_\theta(x_t, y, t)$ denotes the noise predicted by the model at time step t. The objective of this loss function is to minimize the discrepancy between the model-estimated noise ϵ_θ and the constructed target noise ω_t, thereby optimizing the model parameters and improving its ability to fit the generative process. The target noise term ω_t is defined as

$$\omega_t = \frac{m_t\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}(y - x_0) + \frac{\sqrt{\delta_t}}{\sqrt{1-\bar{\alpha}_t}}\epsilon. \tag{6}$$

Here, ω_t consists of two parts. The first part $y - x_0$ represents the non-Gaussian noise present in the chrominance signal, while the second part is Gaussian noise ϵ, which preserves the randomness in the generation process.

C. Convolution and Mamba Hybrid Network Structure

The third component of the HRVFusion framework consists of a feature fusion module followed by eight Mamba blocks [35], designed to accurately predict the noise perturbation at each step of the diffusion process. Its overall structure is illustrated in Fig. 1(c). At each diffusion timestep t, the model takes the current signal x_t, the time-frequency ridge h, and the corresponding time embedding as input, and outputs the residual of the predicted noise. After T iterative steps, the model gradually reconstructs a high-quality BVP signal.

The feature fusion module integrates the input signal with conditional information to produce a comprehensive feature representation that encodes both temporal and frequency-domain context, providing rich information for subsequent Mamba blocks. Specifically, before entering the network, the conditional diffusion model generates a time-domain signal x_t aligned with the input length based on the corresponding time–frequency ridge or chrominance signal, ensuring precise matching between the input and conditional features. The input temporal signal x_t is first projected into a 64-dimensional feature space via a one-dimensional (1-D) convolution. The diffusion timestep t is mapped into a temporal embedding vector through a linear transformation. The time–frequency ridge h is processed by a transposed convolution, followed by a 1-D convolution to align its feature dimension with that of the backbone network. The module then applies dilated convolutions to extract multi-scale temporal context and fuses the result with the conditional features, yielding the final fused representation.

The fused representation is then passed through a stack of Mamba blocks, which serve as the backbone for modeling the joint temporal and spectral structure of the signal. Compared to conventional convolutional or sequential models, Mamba exhibits stronger capability in modeling complex cardiovascular dynamics and capturing subtle heart rate fluctuations. Meanwhile, skip connections are added between residual blocks to effectively fuse shallow and deep features, improving both model expressiveness and reconstruction accuracy.

The use of 1-D convolutions allows the model to handle input signals of varying lengths [36], while the selective mechanism of the Mamba module enables efficient processing of variable-length sequences [35]. By training on short videos, the model is capable of processing long videos during testing (the maximum testing length in this study is five minutes).

IV. EXPERIMENTAL SETTINGS

This section introduces the experimental settings including the evaluated datasets, the implementation details, and evaluation metrics of the experiments.

A. Experimental Dataset

Since the lengths of videos in existing public rPPG datasets (such as PURE [37], UBFC-rPPG [38]) are relatively short, they are only suitable for short-term heart rate (HR) and HRV estimation. However, standard HRV analysis typically requires at least five minutes of continuous signals to capture more comprehensive autonomic fluctuations [4]. To address this, we have also collected and released the first rPPG dataset, termed as LTHRV dataset, that specifically designed for long-term HRV analysis.

The three datasets are introduced as below.

1) Public Dataset: The UBFC-rPPG dataset contains 42 real-world scene video segments captured during a mathematical game (Logitech C920 camera, 640×480@30fps). Synchronous reference PPG signals were recorded via a Contec CMS50E pulse oximeter.

TABLE I
SETTINGS FOR CROSS-DATASET VALIDATION

Task No.	Training Dataset	Testing Dataset
1	PURE	UBFC-rPPG
2	UBFC-rPPG	PURE
3	PURE	LTHRV

The PURE dataset includes 60 video segments from 10 subjects under 6 head motion scenarios (ECO274CVGE camera, 640×480@30fps), with reference PPG signals collected synchronously.

2) Self-collected Dataset: This study collects and publicly releases the first rPPG dataset specifically designed for long-term HRV analysis, comprising high-quality RGB videos, PPG signals, and respiratory signals from 18 healthy participants (12 males and 6 females, aged 21–25). A total of 36 recordings are included, each lasting 15 minutes. The data cover two physiological states, spontaneous breathing and short-term breath holding, providing a rich foundation for investigating HRV dynamics under varying respiratory conditions. In contrast to spontaneous breathing, breath-holding typically elicits more pronounced heart rate variability fluctuations [39], facilitating the characterization of dynamic features of autonomic regulation. We confirm that all participants signed written informed consent forms prior to the experiment. The data collection was approved by the Ethics Committee of Hefei University of Technology and was conducted in strict accordance with relevant ethical guidelines. The LTHRV dataset is publicly available by request.

The experimental setup is illustrated in Fig. 3. The camera was positioned 100–105 cm in front of the participant, with its height adjusted according to individual stature. Participants lay in a supine position in a quiet environment and completed two tasks, (1) 15 minutes of natural breathing, and (2) a breath-holding condition involving 20-second apnea periods at the 5th and 10th minutes, with normal breathing maintained during the remaining time. To ensure signal stability, each participant underwent a 10-minute resting period prior to data collection.

The videos were recorded using a Logitech C920 camera using 640×480 resolution at 30 frame-per-second (fps). PPG signals were captured with a Contec CMS50EA pulse oximeter, and respiratory signals were obtained using an HKH-11C respiratory sensor. All signals were synchronously acquired and aligned. Videos were stored in an uncompressed format to preserve maximum image quality.

Three progressive training and testing tasks are defined, each following a cross-dataset validation configuration as summarized in Table I. Considering that deep learning methods often exhibit significant performance drops when applied to data differing from the training set, we adopt cross-dataset validation to provide a more realistic assessment of the model's generalization and robustness. To this end, we construct diverse cross-domain tasks spanning the PURE, UBFC, and long-term LTHRV datasets, allowing us to examine the stability and applicability of the proposed method across different scenarios.

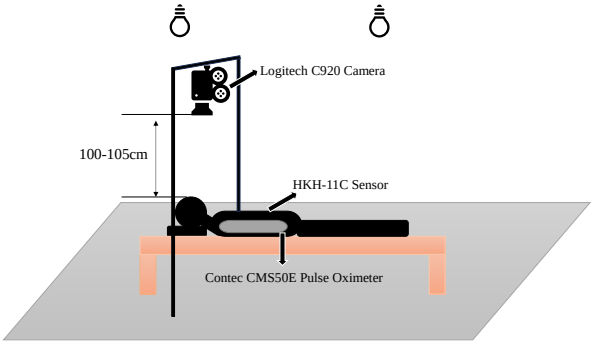

Fig. 3. Illustration of the LTHRV dataset scenarios.

B. Implementation Details

The algorithm in this study is implemented based on the PyTorch framework and trained on an NVIDIA A6000 GPU. The training data comes from PURE and UBFC-rPPG datasets, and sliding window slicing is performed using the pyVHR toolbox [40]. A total of 19,306 samples were generated in PURE dataset under a 10-second processing window with a 5-frame (approximately 0.17 seconds at 30 fps) sliding step. For the UBFC-rPPG dataset, due to the limited number of original videos, the step size was set to 0.1 seconds (3 frames) to maximize the data utilization, resulting in 22,921 samples. During training, the model uses 50-step linear noise schedule with a noise range of [1e-4, 0.35], batch size 64, and learning rate 2e-4, running for a total of 60,000 iterations. In the inference phase, a fast sampling strategy is employed, reducing the number of diffusion time steps from 50 during training to 6 for inference, while the noise schedule [1e-4, 1e-3, 1e-2, 5e-2, 0.2, 0.35] is configured following the recommendation in [34] to improve computational efficiency.

The testing datasets include UBFC-rPPG, PURE, and LTHRV. Considering the signal duration requirements for HRV analysis, the following strategy was adopted during testing. If the video duration is less than five minutes, the maximum duration of each video is used as the window length. if the video exceeds five minutes, a five-minute window with a 30-second stride is applied for sliding window processing. Specifically, sliding window processing on the LTHRV dataset produced a total of 717 samples (360 apnea and 357 normal breathing), with each apnea sample covering a complete event cycle to better capture HRV dynamics. To ensure the stability of chrominance signals under long window conditions, we extract the chrominance signal using a 1.6-second processing window and splice it into a long raw BVP signal. Table II summarizes the key differences between our method and the compared deep learning approaches during the training and testing process. Although our method is also trained with short-window data, it can handle testing signals of arbitrary length. This allows the HRVFusion to generate continuous high-quality rPPG waveforms in one step without the need for subsequent stitching. It significantly ensures the temporal consistency of the signals, and is particularly suitable for long-term HRV analysis.

TABLE II
DIFFERENCES OF HRVFUSION AND COMPARED DEEP LEARNING METHODS IN THE TRAINING AND TESTING PROCESS

Method	Training Length	Testing Length	Stitching Required	Continuity
Compared methods	Short	Same as training	Yes	No
HRVFusion	Short	Arbitrary	No	Yes

TABLE III
TIME- AND FREQUENCY-DOMAIN HRV METRICS EVALUATED

Category	Feature	Definition
Time-domain	AVNN	The average of all NN intervals.
	SDNN	The standard deviation of all NN intervals.
	RMSSD	The root mean square of successive differences between NN intervals.
	pNN50	The proportion of successive NN intervals with differences greater than 50 ms.
Frequency-domain	LF	The power in the low-frequency (0.04–0.15 Hz) range.
	HF	The power in the high-frequency (0.15–0.4 Hz) range.
	LF/HF	The ratio of LF power to HF power.

C. Evaluation Metrics

To comprehensively evaluate the performance of HRVFusion in heart rate (HR) and heart rate variability (HRV) measurement, this study employs a variety of quality metrics for analysis.

1) Heart Rate Evaluation: The heart rates are obtained from the extracted BVP signals using a peak detection method from the open-source toolbox rPPG-Toolbox [41]. The following four metrics are used to evaluate the performance of HR estimation, including mean absolute error (MAE), root mean square error (RMSE), pearson correlation coefficient, and signal-to-noise ratio (SNR).

2) HRV Evaluation: To comprehensively characterize the dynamic properties of HRV, this study used the HRVAS toolbox [42] to evaluate time- and frequency-domain metrics of HRV. The evaluation metrics are shown in Table III. The evaluation metrics for UBFC and PURE datasets include HR and time-domain HRV analysis due to the limitations in signal length. In contrast, the evaluation metrics for LTHRV dataset also include frequency-domain HRV metrics to verify the model's performance for long-term HRV analysis.

V. RESULTS

In this section, the superiority of the proposed HRVFusion method is systematically validated on three datasets. To thoroughly validate the effectiveness of the proposed method, we compared it with several advanced deep learning methods (such as TSCAN [26], PhysFormer [27], DeepPhys [24], EfficientPhys [25], and RhythmFormer [10]). The selection of comparison methods was designed to include representative traditional algorithms, such as CHROM, and mainstream deep learning models, such as TSCAN and RhythmFormer. Specifically, chrominance signals are used as input to provide an intuitive assessment of HRVFusion's ability to enhance signal quality and HRV estimation. TSCAN, as a typical temporal convolutional network, directly models video sequences for heart rate prediction. RhythmFormer leverages self-attention to capture long-term physiological dynamics based on a Transformer architecture, enabling more refined rhythm modeling. This combination of methods allows for a comprehensive benchmark of HRVFusion. All comparison results were generated using the rPPG Toolbox [41] to ensure consistency in experimental settings and evaluation metrics.

TABLE IV
HR EVALUATION RESULTS ON THE UBFC-RPPG DATASET (BEST PERFORMANCE IN BOLD AND SECOND BEST IN UNDERLINE)

Method	MAE (bpm)	RMSE (bpm)	r	SNR (dB)
CHROM [8]	1.86	4.26	0.96	2.38
DeepPhys [24]	1.25	3.23	<u>0.98</u>	2.49
TSCAN [26]	1.32	3.00	<u>0.98</u>	2.35
PhysFormer [27]	2.80	5.07	0.95	1.20
EfficientPhys [25]	1.80	3.43	<u>0.98</u>	1.56
RhythmFormer [10]	<u>1.05</u>	<u>1.69</u>	**0.99**	<u>4.24</u>
HRVFusion	**0.53**	**1.07**	**0.99**	**5.88**

TABLE V
HRV EVALUATION RESULTS ON THE UBFC-RPPG DATASET (BEST PERFORMANCE IN BOLD AND SECOND BEST IN UNDERLINE)

Methods	AVNN (ms)	SDNN (ms)	RMSSD (ms)	pNN50 (%)
	MAE	MAE	MAE	MAE
CHROM [8]	18.07	39.06	58.78	15.73
DeepPhys [24]	11.61	26.88	37.10	11.55
TSCAN [26]	7.96	26.79	43.05	<u>10.43</u>
PhysFormer [27]	20.33	54.99	78.92	20.29
EfficientPhys [25]	13.92	40.34	62.82	13.06
RhythmFormer [10]	<u>6.00</u>	<u>15.04</u>	**21.98**	**8.04**
HRVFusion	**3.29**	**14.69**	<u>25.46</u>	14.74

A. UBFC-rPPG Dataset

In this subsection, we train the model on the PURE dataset and test it on the UBFC-rPPG dataset to evaluate its short-term responsiveness under a normal human-computer interaction scenario. The heart rate (HR) evaluation results are shown in Table IV. HRVFusion achieves a MAE of only 0.53 bpm, which is 49% lower than the second-best method (1.05 bpm from RhythmFormer), ranking first among all compared models. As shown in Table V, HRVFusion achieves the best results in AVNN and SDNN, with values of 3.29 ms and 14.69 ms respectively, representing improvements of 45% and

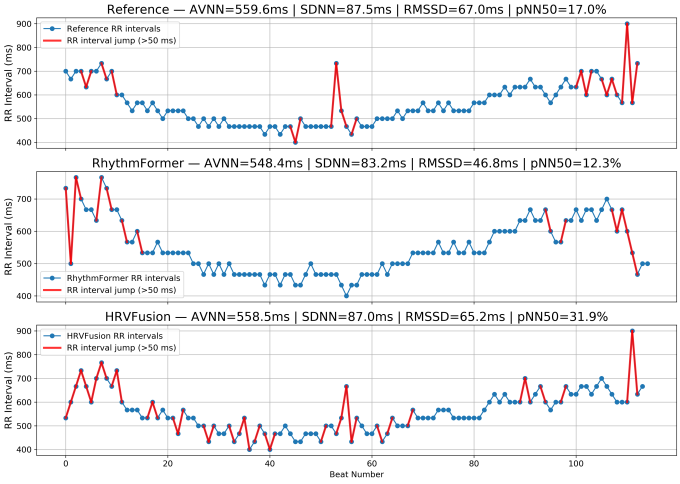

Fig. 4. HRVFusion vs RhythmFormer: RR interval comparison with pNN50 highlights (subject 24 of UBFC-rPPG dataset).

2.3% over the second-best results. These results highlight the unique advantages of HRVFusion in autonomic activity analysis, particularly in the accuracy and stability of short-term HRV measurement.

We observe that HRVFusion achieves excellent performance on the AVNN, SDNN, and RMSSD metrics, but shows relatively larger errors for pNN50. This is mainly because pNN50 reflects the proportion of adjacent RR interval differences exceeding 50 ms, emphasizing short-term abrupt heart rate changes. Its discrete and highly sensitive nature amplifies even slight IBI prediction deviations, resulting in greater fluctuations in this metric.

To provide a more intuitive illustration, Fig. 4 depicts the RR interval curves estimated by HRVFusion and RhythmFormer for subject 24 in the UBFC-rPPG dataset. The red segments visually indicate pNN50, and the corresponding time-domain HRV metrics of the reference, HRVFusion, and RhythmFormer methods are also listed in the title of figures. The MAE (|predicted value − reference value|) calculated on this sample shows that errors on AVNN, SDNN, and RMSSD for HRVFusion are 1.1ms, 0.5ms, and 1.8ms, respectively, significantly better than those (11.2ms, 4.3ms, and 20.2ms) of RhythmFormer. However, the error of pNN50 for HRVFusion's is 14.9%, which is higher than 4.7% of RhythmFormer.

The full results on the UBFC dataset are shown in Table V. Similar to the above example for subject 24, HRVFusion demonstrates overall superior IBI extraction accuracy, but exhibits slightly higher errors on the highly sensitive pNN50 metric compared to the baseline methods.

We also provide the corresponding Bland-Altman plot in Fig. 5, which shows that the differences between the predicted values and the ground truth using the diffusion method are mostly concentrated around the mean, indicating high prediction accuracy and stability.

B. PURE Dataset

To evaluate the generalization capability of HRVFusion under cross-dataset conditions, the model was trained on the UBFC-rPPG dataset and tested on the PURE dataset. The HR and HRV results are summarized in Tables VI and VII, respectively.

TABLE VI
HR EVALUATION RESULTS ON THE PURE DATASET (BEST PERFORMANCE IN BOLD AND SECOND BEST IN UNDERLINE)

Method	MAE (bpm)	RMSE (bpm)	r	SNR (dB)
CHROM [8]	<u>2.14</u>	4.90	<u>0.98</u>	<u>8.66</u>
DeepPhys [24]	9.65	15.84	0.77	2.71
TSCAN [26]	6.29	12.26	0.87	3.14
PhysFormer [27]	22.56	28.58	0.50	1.69
EfficientPhys [25]	9.26	15.63	0.78	2.76
RhythmFormer [10]	2.19	<u>4.17</u>	**0.99**	7.95
HRVFusion	**0.72**	**1.76**	**0.99**	**12.39**

TABLE VII
HRV EVALUATION RESULTS ON THE PURE DATASET (BEST PERFORMANCE IN BOLD AND SECOND BEST IN UNDERLINE)

Methods	AVNN (ms)	SDNN (ms)	RMSSD (ms)	pNN50 (%)
	MAE	MAE	MAE	MAE
CHROM [8]	<u>19.17</u>	<u>47.70</u>	78.99	29.15
DeepPhys [24]	123.39	210.85	291.89	31.73
TSCAN [26]	94.31	203.79	288.03	31.67
PhysFormer [27]	39.04	73.15	123.38	33.98
EfficientPhys [25]	112.11	225.54	322.51	35.16
RhythmFormer [10]	20.82	55.87	<u>77.70</u>	<u>23.32</u>
HRVFusion	**6.94**	**18.15**	**25.13**	**11.77**

In the heart rate estimation task, the traditional CHROM method maintained a low error under domain shift conditions, achieving a mean absolute error (MAE) of 2.14 bpm and a Pearson correlation coefficient of 0.98. In contrast, several deep learning methods showed significant performance degradation: the MAEs of TSCAN and DeepPhys increased to 6.29 bpm and 9.65 bpm, respectively, while PhysFormer's error further expanded to 22.56 bpm, indicating poor adaptability to changes in data distribution. By comparison, HRVFusion achieved the best results across all metrics, with the MAE and RMSE reduced to 0.72 bpm and 1.76 bpm, respectively, a Pearson correlation coefficient of 0.99, and an SNR of 12.39 dB, demonstrating stronger robustness and cross-domain generalization capability.

Regarding HRV estimation, HRVFusion also exhibited consistent performance advantages. It achieved the lowest average errors across four key metrics: AVNN, SDNN, RMSSD, and pNN50. Specifically, the MAEs for SDNN and RMSSD were 18.15 ms and 25.13 ms, respectively, outperforming all compared methods. These results indicate that the proposed method provides higher temporal resolution and greater accuracy in capturing fluctuations in autonomic nervous system rhythms.

To provide a more intuitive demonstration of the effectiveness of the proposed method, we present waveform compar-

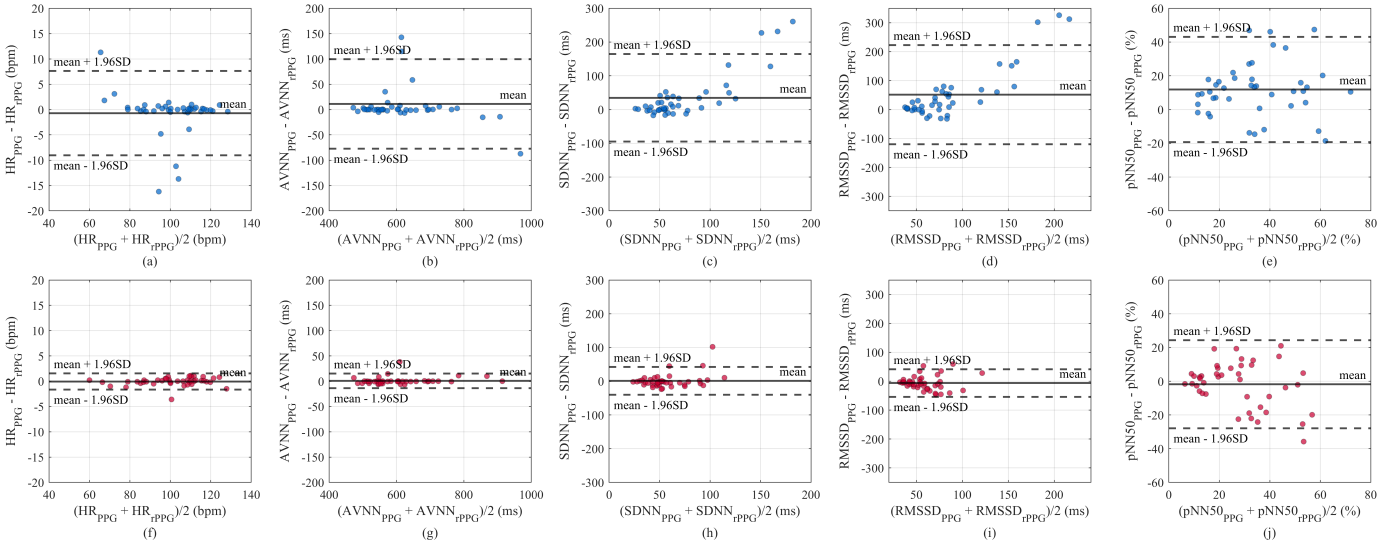

Fig. 5. Bland-Altman plots of HR and HRV metrics on the UBFC-rPPG dataset: CHROM (blue) vs. HRVFusion (red).

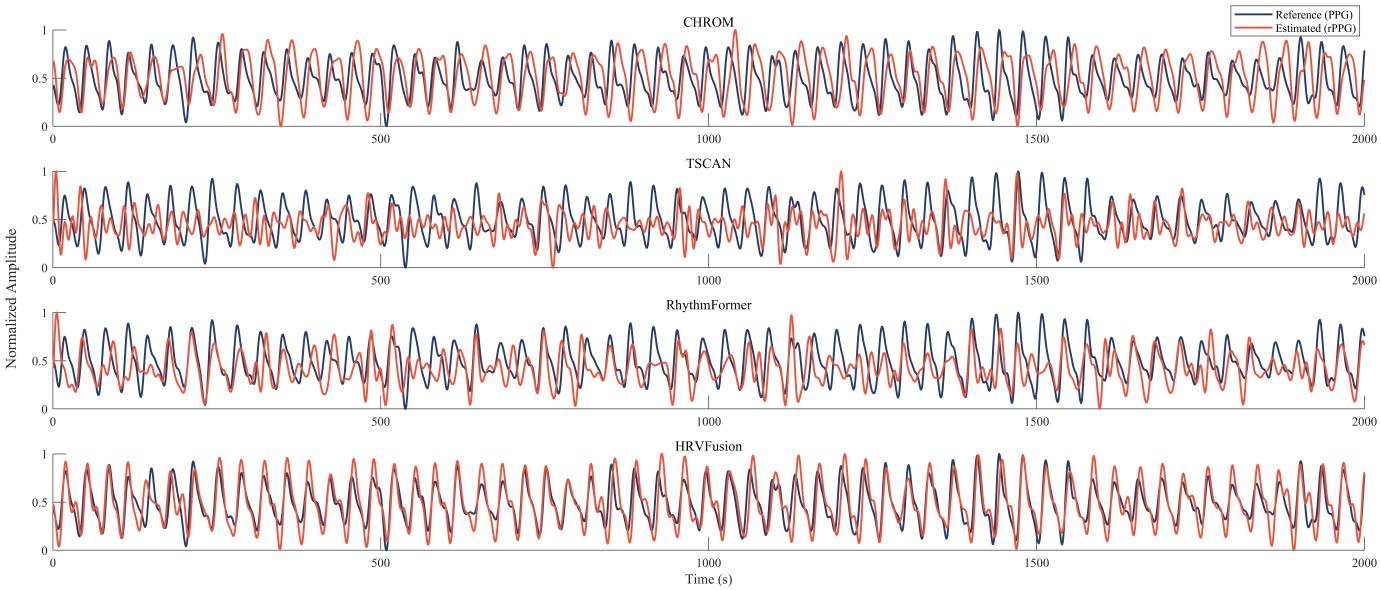

Fig. 6. Comparison of rPPG waveforms estimated by different methods with reference PPG on the PURE dataset (Video ID 05-04).

isons between HRVFusion and baseline methods (CHROM, TSCAN, RhythmFormer). As shown in Fig. 6, the BVP signals extracted by each method for the fast translation video of Subject 5 (ID: 05-04) in the PURE dataset are compared with the reference PPG waveform.

From this figure, it can be observed that the waveform reconstructed by HRVFusion more closely matches the reference PPG in both morphology and rhythm. In contrast, the waveforms generated by CHROM, TSCAN, and RhythmFormer exhibit varying degrees of drift and distortion. HRVFusion shows better alignment of peak–valley positions and greater stability in cycle duration, indicating higher accuracy in modeling rhythm variations. This comparison further demonstrates the method's robustness under complex disturbances and its consistency in signal reconstruction for practical applications. We also show the Bland-Altman plots in Fig. 7. Compared to the CHROM method, the proposed signal reconstruction method demonstrates significant improvement in HRV metrics, showcasing stronger detail retention and noise resistance.

C. LTHRV Dataset

To validate the effectiveness of the proposed HRVFusion method in long-term HRV measurement tasks, we test the model trained on the PURE dataset using the LTHRV dataset.

The evaluation results of time- and frequency-domain HRV metrics are shown in Table VIII and Table IX, respectively. As observed, the HRVFusion method consistently outperforms existing approaches across all HRV metrics. Specifically, for key metrics like SDNN and RMSSD, which reflect heart rate variability, the error remains within the range of 6 to 18 milliseconds. In metrics that are more sensitive to frequency components, such as the LF/HF ratio, LF, and HF, HRVFu-

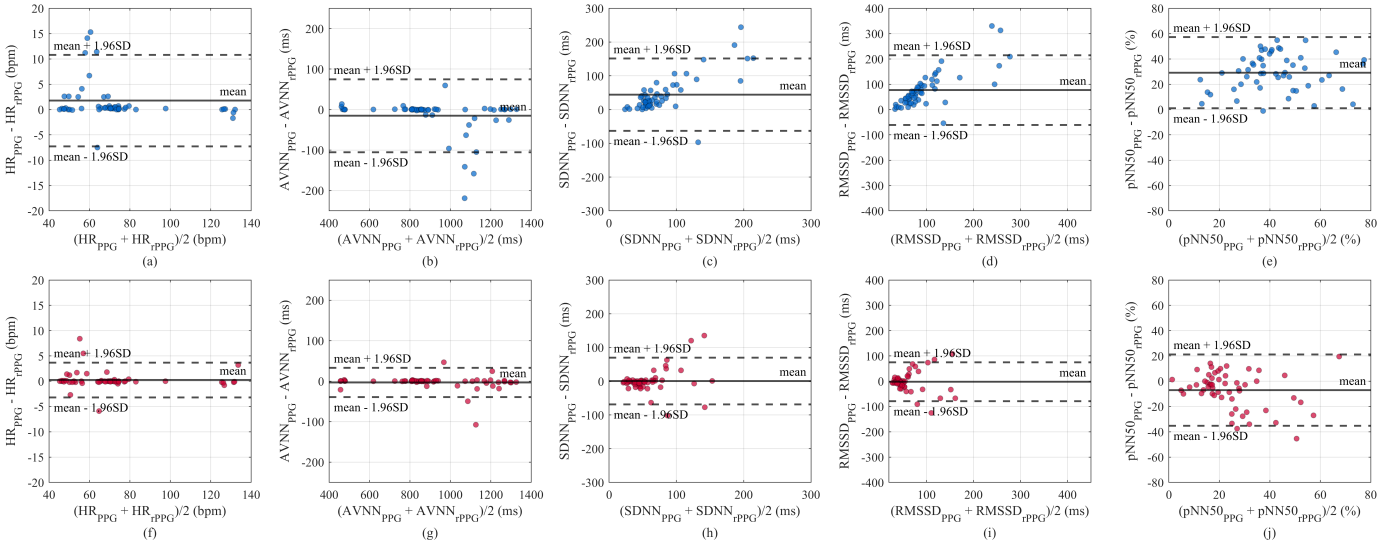

Fig. 7. Bland-Altman plots of HR and HRV metrics on the PURE dataset: CHROM (blue) vs. HRVFusion (red).

TABLE VIII
HRV EVALUATION RESULTS OF NORMAL BREATHING SCENARIOS ON THE LTHRV DATASET (BEST PERFORMANCE IN BOLD AND SECOND BEST IN UNDERLINE)

Methods	SDNN (ms)	RMSSD (ms)	pNN50 (%)	LF(HF) (u.n)		LF/HF	
	MAE	MAE	MAE	MAE	r	MAE	r
CHROM [8]	<u>31.83</u>	<u>71.67</u>	30.32	0.17	<u>0.55</u>	<u>0.47</u>	<u>0.39</u>
DeepPhys [24]	73.84	128.11	13.08	0.19	0.25	0.54	0.22
TSCAN [26]	59.26	116.77	21.09	0.20	0.32	0.56	0.35
PhysFormer [27]	100.82	185.27	22.67	<u>0.15</u>	0.38	0.50	0.23
EfficientPhys [25]	101.62	183.84	25.92	<u>0.15</u>	0.28	0.48	0.16
RhythmFormer [10]	113.28	109.16	<u>11.94</u>	0.21	0.21	0.77	0.13
HRVFusion	**7.90**	**12.47**	**10.30**	**0.08**	**0.84**	**0.34**	**0.85**

TABLE IX
HRV EVALUATION RESULTS OF APNEA BREATHING SCENARIOS ON THE LTHRV DATASET (BEST PERFORMANCE IN BOLD AND SECOND BEST IN UNDERLINE)

Methods	SDNN (ms)	RMSSD (ms)	pNN50 (%)	LF(HF) (u.n)		LF/HF	
	MAE	MAE	MAE	MAE	r	MAE	r
CHROM [8]	<u>29.53</u>	<u>67.22</u>	30.33	0.17	<u>0.48</u>	0.52	<u>0.54</u>
DeepPhys [24]	68.75	97.52	<u>11.8</u>	0.17	0.17	0.50	0.18
TSCAN [26]	71.90	105.88	21.14	0.21	0.25	0.56	0.23
PhysFormer [27]	115.24	182.17	22.93	0.16	0.24	0.47	0.15
EfficientPhys [25]	139.67	219.19	28.89	<u>0.14</u>	0.39	<u>0.44</u>	0.24
RhythmFormer [10]	88.06	81.95	12.58	0.23	0.14	0.84	0.14
HRVFusion	**5.63**	**9.08**	**8.62**	**0.08**	**0.86**	**0.30**	**0.86**

sion also maintains high stability and robustness, with errors significantly lower than all comparison methods.

Although existing deep learning methods may work well in short-term tasks, they typically rely on fixed-length video segments as input and perform concatenation, which limits their capacity to capture long-term temporal dependencies and physiological state variations. As a result, these methods often suffer from information loss and lack of temporal coherence in long-term HRV tasks, leading to suboptimal outcomes on the LTHRV dataset. In contrast, HRVFusion better handles long-range dependencies, suppresses noise, and accurately recovers physiologically meaningful temporal features, thereby achieving superior performance in long-term HRV measurement.

In addition, to verify the reliability of the reconstructed signals, we present the Bland-Altman plots of CHROM and HRVFusion across various HRV metrics in Fig. 8 and Fig. 9 for the normal and apnea scenarios, respectively. The results show that HRVFusion results closely match the reference ones, further demonstrating the effectiveness of the method in long-term HRV measurements.

It is worth noting that we further present visual comparisons of 5-minute IBI sequences in Fig. 10 under normal and apnea breathing conditions. These comparisons clearly illustrate that the proposed method effectively recovers the IBI fluctuation patterns of the reference PPG, demonstrating strong consistency in heart rate variation details. In particular, the model maintains stable reconstruction performance under the apnea condition with pronounced heart rate variability.

D. Ablation Study

We conducted a series of ablation studies on the UBFC-rPPG and LTHRV datasets to systematically evaluate the contributions of key components in our framework. Specifically, we investigated the following factors. (1) the effect of hybrid diffusion compared to pure Gaussian diffusion; (2) the role of incorporating time–frequency ridge guidance during the reverse denoising process; (3) the impact of the number of Mamba blocks; (4) the advantage of the Mamba architecture over standard 1-D CNNs; (5) the influence of the conditional signal selection (CHROM vs. Green) on hybrid noise modeling. The results are summarized in Table X, Table XI, and Table XII, respectively.

Pure Gaussian diffusion vs. hybrid diffusion: To examine the role of hybrid diffusion (combining chrominance signals with Gaussian noise), we simplified the hybrid diffusion process into a pure Gaussian diffusion setting (i.e., setting $m_t = 0$

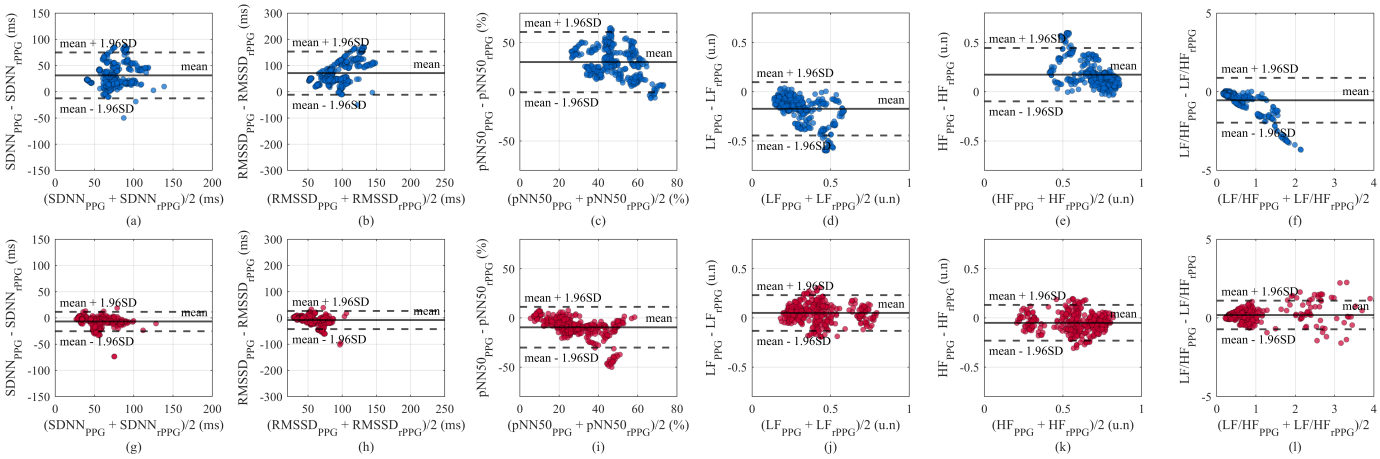

Fig. 8. Bland–Altman plots of HRV metrics on LTHRV dataset with normal breathing: CHROM (blue) vs. HRVFusion (red).

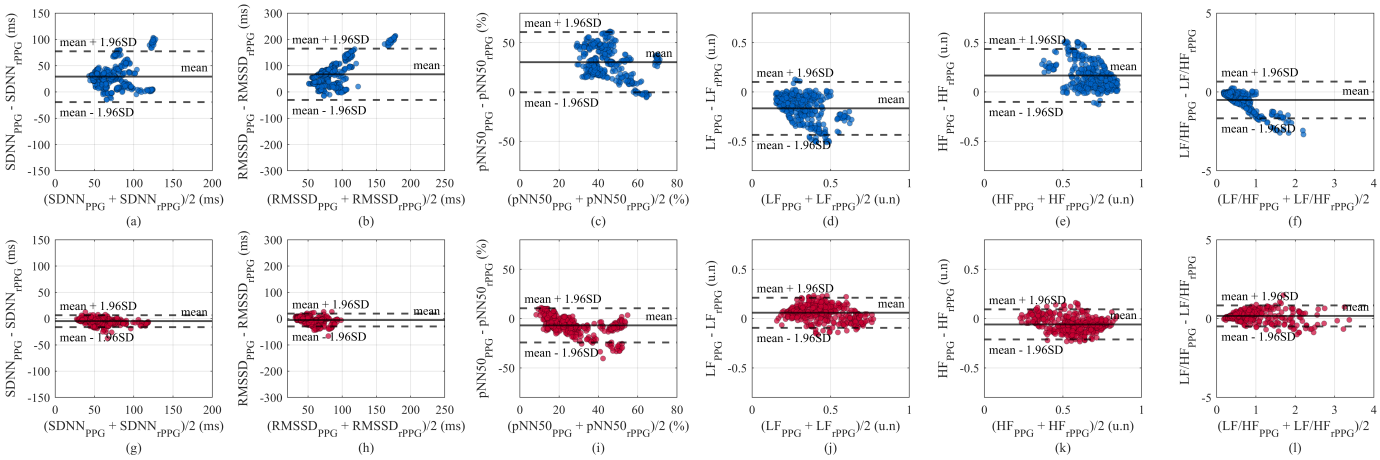

Fig. 9. Bland–Altman plots of HRV metrics on LTHRV dataset with apnea breathing: CHROM (blue) vs. HRVFusion (red).

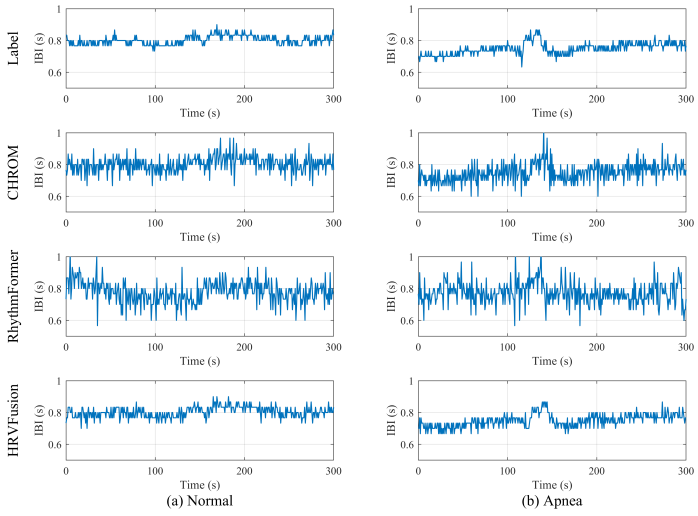

Fig. 10. Comparisons of IBI sequences on the LTHRV dataset. (a) Normal, (b) Apnea.

in Eq. (1)). The results show that hybrid diffusion outperforms pure Gaussian diffusion in both heart rate recovery and HRV estimation. Compared to pure Gaussian noise, hybrid diffusion introduces non-Gaussian perturbation patterns that more accurately reflect real-world conditions, thereby enhancing the model's robustness in complex environments. Moreover, the chrominance signal inherently encodes physiologically relevant cardiac pulsation information, providing additional guidance during reconstruction and facilitating the generation of signals that are more consistent with underlying physiological dynamics. On the UBFC-rPPG dataset, it substantially reduced errors in HR, AVNN, SDNN, and RMSSD. On the LTHRV dataset, hybrid diffusion consistently improved the correlation of LF/HF under both normal and apnea conditions, with the correlation coefficient increasing from 0.56 to 0.86 during apnea. These findings indicate that hybrid diffusion better captures non-Gaussian disturbances, thereby enhancing the model's ability to recover rhythm and long-term HRV characteristics.

With vs. without (w/o) time-frequency ridge : We compared the reconstruction performance with and without ridge guidance during the reverse denoising process. Incorporating the time-frequency ridge led to improvements in both time- and frequency-domain metrics. For instance, on the UBFC-rPPG dataset, the HR error decreased from 0.89 bpm to 0.53 bpm, while SDNN and RMSSD improved by 3-6 ms. On the

TABLE X
ABLATION STUDY RESULTS ON THE UBFC-RPPG DATASET (BEST PERFORMANCE IN BOLD)

Methods	HR				HRV			
	MAE bpm	RMSE bpm	Pearson	SNR dB	AVNN MAE (ms)	SDNN MAE (ms)	RMSSD MAE (ms)	pNN50 MAE (%)
HRVFusion w/o ridge	0.89	1.73	**0.99**	6.32	5.95	17.79	31.70	14.34
6 Mamba blocks	0.81	1.41	**0.99**	5.62	5.55	17.38	29.08	**14.30**
10 Mamba blocks	0.86	1.43	**0.99**	5.22	5.93	19.01	30.25	13.94
Pure Gaussian diffusion	0.84	1.11	**0.99**	**8.49**	4.74	17.20	26.41	17.62
8 1-D CNN blocks	1.02	2.04	**0.99**	2.98	8.87	26.37	40.23	16.31
HRVFusion (w ridge + 8 Mamba blocks + hybrid diffusion)	**0.53**	**1.07**	**0.99**	5.88	**3.28**	**13.97**	**25.59**	14.90

TABLE XI
ABLATION STUDY RESULTS ON THE LTHRV DATASET (BEST PERFORMANCE IN BOLD)

Scenario	Method	SDNN MAE (ms)	RMSSD MAE (ms)	pNN50 MAE (%)	LF		HF	
					MAE (%)	r	MAE (%)	r
Normal	HRVFusion w/o ridge	11.24	16.70	13.61	0.09	0.77	0.35	0.75
	6 Mamba blocks	9.78	17.09	9.22	0.09	0.84	0.38	0.82
	10 Mamba blocks	8.30	12.91	10.78	0.09	0.82	0.35	0.85
	Pure Gaussian diffusion	15.71	18.51	21.54	0.12	0.68	0.47	0.71
	8 1-D CNN blocks	9.79	24.89	19.17	0.09	0.84	0.31	0.80
	HRVFusion (w ridge + 8 Mamba blocks + hybrid diffusion)	**7.90**	**12.47**	**10.30**	**0.08**	**0.84**	**0.34**	**0.85**
Apnea	HRVFusion w/o ridge	11.33	20.06	11.12	0.10	0.66	0.37	0.61
	6 Mamba blocks	8.61	17.27	7.80	0.09	0.77	0.36	0.73
	10 Mamba blocks	5.86	9.14	9.00	0.09	0.86	0.31	0.85
	Pure Gaussian diffusion	12.67	12.73	18.24	0.12	0.59	0.44	0.56
	8 1-D CNN blocks	10.45	26.17	20.89	0.09	0.85	0.30	0.86
	HRVFusion (w ridge + 8 Mamba blocks + hybrid diffusion)	**5.63**	**9.08**	**8.62**	**0.08**	**0.86**	**0.30**	**0.86**

TABLE XII
HR AND HRV METRICS OF CHROM AND GREEN SIGNALS ON THE UBFC-RPPG DATASET (BEST PERFORMANCE IN BOLD).

Methods	HR				HRV			
	MAE bpm	RMSE bpm	Pearson	SNR dB	AVNN MAE (ms)	SDNN MAE (ms)	RMSSD MAE (ms)	pNN50 MAE (%)
Green	5.17	13.54	0.66	-0.74	99.12	193.79	141.97	26.46
CHROM	1.86	4.26	0.96	2.38	18.07	39.06	58.78	15.73
HRVFusion (Green)	3.40	6.22	0.93	**7.69**	29.72	45.95	61.72	18.00
HRVFusion (CHROM)	**0.53**	**1.07**	**0.99**	5.88	**3.28**	**13.97**	**25.59**	**14.90**

LTHRV dataset, the inclusion of ridge guidance significantly reduced the errors of SDNN and RMSSD under both normal and apnea conditions, and also improved the correlation of LF/HF. These results demonstrate that ridge guidance enhances the physiological consistency of the generated signals, helping the model to more accurately recover cardiac rhythms.

Number of Mamba blocks: We evaluated network architectures comprising 6, 8, and 10 Mamba blocks. The results indicate that the configuration with 8 blocks achieved the optimal performance on both datasets. Relative to the 6-block configuration, 8 blocks provided enhanced modeling capacity. Compared to the 10-block configuration, it avoided the performance degradation and computational overhead associated with excessive depth. Therefore, the 8-block configuration offers a favorable trade-off between accuracy and computational efficiency.

Mamba vs. CNN: To further validate the effectiveness of the Mamba architecture, we replaced the 8 Mamba blocks with 8 1-D CNN blocks. The CNN-based model yielded larger errors in time-domain metrics such as HR, SDNN, and RMSSD, and showed notable declines in frequency-domain indices including LF/HF and pNN50. In contrast, Mamba demonstrated clear advantages in modeling long-range dependencies, handling non-stationary signals, and preserving physiological rhythms.

CHROM vs. Green: To investigate the impact of conditional signal selection on signal recovery, we conducted a comparative experiment on the UBFC-rPPG dataset by replacing CHROM with the green channel signal (Green), which is more susceptible to noise as the conditional input. As shown in Table XII, all performance metrics dropped significantly when using Green. This is mainly because, during the reverse denoising process, the model relies on the observed signal and its time–frequency ridge as conditions to guide the generation of high-quality signals. When the conditional signal is heavily contaminated by noise, the model struggles to distinguish true cardiac information from artifacts, thereby reducing prediction accuracy. In contrast, the chrominance signal provides a more reliable capture of heart rate–related features, and its non-Gaussian noise can be effectively modeled by the framework, thus demonstrating greater robustness in signal recovery. Based on these findings, CHROM was selected as the conditional input for hybrid noise modeling, as it balances

computational simplicity, accessibility, the ability to model non-Gaussian noise, and practical feasibility. Nevertheless, as the experiments indicate, other suitable signals could also serve as alternative conditional inputs for non-Gaussian hybrid noise modeling.

E. Computational Efficiency

To further validate the feasibility of the proposed method in real-world application scenarios, we analyzed the model's runtime efficiency and computational cost.

All the experiments were conducted on an NVIDIA A6000 GPU. The network model contains approximately 0.93 million parameters. For inference, we introduced a fast sampling strategy that reduces the diffusion process steps from 50 during training to 6, effectively lowering the number of iterations and significantly improving the running speed. It takes on average 249.12 milliseconds to process a 10-second sample with a memory usage of about 585 MB. Accordingly, processing a 5-minute sample requires only approximately 490.20 milliseconds, with memory usage increasing slightly to about 680 MB. The actual runtime does not scale strictly linearly with input length, as GPU parallelism and amortized overhead allow longer sequences to be processed more efficiently.

In terms of computational load, processing a 10-second sample requires approximately 0.18 GFLOPs, while processing a 5-minute sample requires about 5.72 GFLOPs. The computational complexity scales linearly with the input length, i.e., $O(n)$. This efficiency is attributed to the Mamba architecture used in the core network of HRVFusion, which leverages a hardware-aware parallel scan algorithm and selective state update mechanism, enabling the network to maintain high scalability when processing long sequences.

These results demonstrate that HRVFusion achieves high-precision signal reconstruction while maintaining strong computational efficiency, meeting the practical demands of long-term contactless HRV measurement.

VI. CONCLUSION

This paper presented HRVFusion, a conditional diffusion model-based framework for rPPG signal reconstruction, which can reconstruct long-term, high-quality BVP signals, thereby improve the accuracy of video-based HRV estimation. In the forward diffusion process, chrominance signals extracted from videos and Gaussian noise are introduced to simulate complex real-world disturbances in rPPG signals. During the reverse process, CHROM waveforms and their time-frequency ridges are jointly used as conditional guidance to constrain signal generation from temporal and frequency domains, effectively enhancing signal quality and stability. Compared to existing deep learning methods, HRVFusion achieved significantly lower mean absolute errors across key HRV metrics such as SDNN, RMSSD, pNN50, and LF/HF. In future work, HRVFusion will further validate its practicality and generalizability in clinical applications such as arrhythmia detection, stress assessment, and sleep quality analysis.

REFERENCES

[1] Y. Haque, R. S. Zawad, C. S. A. Rony, H. Al Banna, T. Ghosh, M. S. Kaiser, and M. Mahmud, "State-of-the-art of stress prediction from heart rate variability using artificial intelligence," *Cogn Comput*, vol. 16, no. 2, pp. 455–481, 2024.

[2] B. M. Appelhans and L. J. Luecken, "Heart rate variability as an index of regulated emotional responding," *Rev. Gen. Psychol.*, vol. 10, no. 3, pp. 229–240, 2006.

[3] R. E. Kleiger, P. K. Stein, and J. T. Bigger Jr, "Heart rate variability: measurement and clinical utility," *Ann. Noninvas. Electro.*, vol. 10, no. 1, pp. 88–101, 2005.

[4] D. Kotecha, G. New, M. Flather, D. Eccleston, J. Pepper, and H. Krum, "Five-minute heart rate variability can predict obstructive angiographic coronary disease," *Heart*, vol. 98, no. 5, pp. 395–401, 2012.

[5] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." *Opt. Express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[6] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Opt. Express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.

[7] W. Wang, A. C. Den Brinker, S. Stuijk, and G. De Haan, "Algorithmic principles of remote ppg," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1479–1491, 2016.

[8] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, 2013.

[9] X. Zhang, W. Sun, H. Lu, Y. Chen, Y. Ge, X. Huang, J. Yuan, and Y. Chen, "Self-similarity prior distillation for unsupervised remote physiological measurement," *IEEE Trans. on Multimedia*, vol. 26, pp. 10 290–10 305, 2024.

[10] B. Zou, Z. Guo, J. Chen, J. Zhuo, W. Huang, and H. Ma, "Rhythmformer: Extracting patterned rppg signals based on periodic sparse attention," *Pattern Recogn.*, vol. 164, p. 111511, 2025.

[11] X. Liu, Y. Zhang, Z. Yu, H. Lu, H. Yue, and J. Yang, "rppg-mae: Self-supervised pretraining with masked autoencoders for remote physiological measurements," *IEEE Trans. on Multimedia*, vol. 26, pp. 7278–7293, 2024.

[12] C. Y. Zhou, X. L. Zhan, Y. W. Wei, X. C. Zhang, Y. G. Li, C. C. Wang, and X. L. Ye, "Review of heart rate variability parameter estimation methods in facial video," *J. Image Graph.*, vol. 30, no. 4, pp. 953–976, 2025.

[13] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," *arXiv preprint arXiv:1905.02419*, 2019.

[14] J. Kranjec, S. Beguš, G. Geršak, and J. Drnovšek, "Non-contact heart rate and heart rate variability measurements: A review," *Biomed. Signal Process. Control*, vol. 13, pp. 102–112, 2014.

[15] H. Kuang, F. Lv, X. Ma, and X. Liu, "Efficient spatiotemporal attention network for remote heart rate variability analysis," *Sensors*, vol. 22, no. 3, p. 1010, 2022.

[16] T. F. o. t. E. S. o. C. t. N. A. S. o. P. Electrophysiology, "Heart rate variability: standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[17] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, and X. Chen, "Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography," *IEEE J. Biomed. Health. Inf.*, vol. 25, no. 5, pp. 1373–1384, 2021.

[18] H. Lu, H. Han, and S. K. Zhou, "Dual-GAN: Joint bvp and noise modeling for remote physiological measurement," in *Proc. IEEE Int. Conf. Comput. Vision*, 2021, pp. 12 404–12 413.

[19] I. Daubechies, J. Lu, and H.-T. Wu, "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool," *Appl. Comput. Harmon. A.*, vol. 30, no. 2, pp. 243–261, 2011.

[20] M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity," in *2011 federated conference on computer science and information systems (FedCSIS)*. IEEE, 2011, pp. 405–410.

[21] W. Wang, S. Stuijk, and G. De Haan, "A novel algorithm for remote photoplethysmography: Spatial subspace rotation," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 9, pp. 1974–1984, 2015.

[22] R. Song, Z. Du, J. Cheng, C. Li, and X. Yang, "Video-based heart rate estimation with spectrogram signal quality ranking and fusion," *Biomed. Signal Process. Control*, vol. 100, p. 107094, 2025.

[23] B. Chwyl, A. G. Chung, R. Amelard, J. Deglint, D. A. Clausi, and A. Wong, "Time-frequency domain analysis via pulselets for non-contact heart rate estimation from remotely acquired photoplethysmograms," in *Proc. Conf. Comput. Robot Vis.*, 2016, pp. 201–207.

[24] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 349–365.

[25] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff, "Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 5008–5017.

[26] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," *Adv. Neural Inform. Process. Syst.*, vol. 33, pp. 19 400–19 411, 2020.

[27] Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, and G. Zhao, "Physformer: Facial video-based physiological measurement with temporal difference transformer," in *Proc. IEEE Int. Conf. Comput. Vision*, 2022, pp. 4186–4196.

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.

[29] S. Chen, K.-L. Wong, J.-W. Chin, T.-T. Chan, and R. H. So, "Diffphys: Enhancing signal-to-noise ratio in remote photoplethysmography signal using a diffusion model approach," *Bioengineering*, vol. 11, no. 8, p. 743, 2024.

[30] Y.-H. Jeong and Y.-S. Choi, "Diffusion-phys: noise-robust heart rate estimation from facial videos via diffusion models," *Biomed. Eng. Lett.*, pp. 1–11, 2025.

[31] W. Qian, G. Su, D. Guo, J. Zhou, X. Li, B. Hu, S. Tang, and M. Wang, "Physdiff: Physiology-based dynamicity disentangled diffusion model for remote physiological measurement," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 6, 2025, pp. 6568–6576.

[32] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *IEEE Int. Conf. Acoust. Speech Signal Process.* Ieee, 2022, pp. 7402–7406.

[33] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee *et al.*, "Mediapipe: A framework for perceiving and processing reality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2019, 2019.

[34] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," *arXiv preprint arXiv:2009.00713*, 2020.

[35] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[36] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[37] R. Stricker, S. Müller, and H.-M. Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," in *IEEE Ro-man 2014*. IEEE, 2014, pp. 1056–1062.

[38] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recogn. Lett.*, vol. 124, pp. 82–90, 2019.

[39] X. Chen, T. Chen, F. Yun, Y. Huang, and J. Li, "Effect of repetitive end-inspiration breath holding on very short-term heart rate variability in healthy humans," *Physiol Meas.*, vol. 35, no. 12, p. 2429, 2014.

[40] G. Boccignone, D. Conte, V. Cuculo, A. D'Amelio, G. Grossi, R. Lanzarotti, and E. Mortara, "pyvhr: a python framework for remote photoplethysmography," *PeerJ Comput. Sci.*, vol. 8, p. e929, 2022.

[41] X. Liu, G. Narayanswamy, A. Paruchuri, X. Zhang, J. Tang, Y. Zhang, R. Sengupta, S. Patel, Y. Wang, and D. McDuff, "rppg-toolbox: Deep remote ppg toolbox," *Adv. Neural Inform. Process. Syst.*, vol. 36, pp. 68 485–68 510, 2023.

[42] J. T. Ramshur Jr, "Design, evaluation, and application of heart rate variability analysis software (HRVAS)," MSc thesis, Univ. of Memphis, 2010.