M^{3}PD Dataset: Enabling Dual-view Photoplethysmography on Smartphones in Lab and Clinical Settings

JIANKAI TANG[*†], Tsinghua University, China
TAO ZHANG[*], Tsinghua University, China
JIA LI[*], Beijing Anzhen Hospital, China
YUNTAO WANG, Tsinghua University and Qinghai University, China
YIRU ZHANG, Tsinghua University, China
MINGYU ZHANG, Tsinghua University, China
KEGANG WANG, Tsinghua University, China
YUMING HAO, Beijing Anzhen Hospital, China
BOLIN WANG, Beijing Anzhen Hospital, China
HAIYANG LI, Beijing Anzhen Hospital, China
XINGYAO WANG, Agency for Science, Technology and Research, Singapore
YUANCHUN SHI, Tsinghua University and Qinghai University, China
SICHONG QIAN, Beijing Anzhen Hospital, China

Portable physiological monitoring is essential for early detection and management of cardiovascular disease, but current methods often require specialized equipment that limits accessibility or imposes impractical postures that patients cannot maintain. Video-based photoplethysmography on smartphones offers a convenient non-invasive alternative, yet it still faces reliability challenges caused by motion artifacts, lighting variations, and single-view constraints. Few studies have demonstrated that this technology can be reliably applied to physiological monitoring of cardiovascular patients, and no widely used open datasets exist for researchers to examine its cross-device accuracy. To address these limitations, we introduce the M^{3}PD dataset—the first publicly available dual-view mobile photoplethysmography dataset—comprising synchronized facial and fingertip videos captured simultaneously via front and rear smartphone cameras from 60 participants (including 47 cardiovascular patients). Building on this dual-view setting, we further propose the F^{3}Mamba, which fuses the facial and fingertip views through Mamba-based temporal modeling. The model reduces heart-rate error by 21.9–30.2% over existing single-view baselines while showing enhanced robustness across challenging real-world scenarios, *and*

[*]Co-first authors.
[†]Corresponding author.

Authors' Contact Information: Jiankai Tang (corresponding author), Tsinghua University, China, tjk24@mails.tsinghua.edu.cn; Tao Zhang, Tsinghua University, China, zt19375356@gmail.com; Jia Li, Beijing Anzhen Hospital, China, 79333529@qq.com; Yuntao Wang, Tsinghua University and Qinghai University, China, yuntaowang@tsinghua.edu.cn; Yiru Zhang, Tsinghua University, China, yr-zhang24@mails.tsinghua.edu.cn; Mingyu Zhang, Tsinghua University, China, zhang-my22@mails.tsinghua.edu.cn; Kegang Wang, Tsinghua University, China, kegang.wang@foxmail.com; Yuming Hao, Beijing Anzhen Hospital, China, mingheguyu@163.com; Bolin Wang, Beijing Anzhen Hospital, China, bjm79125@icloud.com; Haiyang Li, Beijing Anzhen Hospital, China, ocean0203@163.com; Xingyao Wang, Agency for Science, Technology and Research, Singapore, wang_ xingyao@a-star.edu.sg; Yuanchun Shi, Tsinghua University and Qinghai University, China, shiyc@tsinghua.edu.cn; Sichong Qian, Beijing Anzhen Hospital, China, drqsc1990a@163.com. For correspondence, please contact Jiankai Tang, Yuntao Wang or Sichong Qian.



This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2474-9567/2026/6-ART61
https://doi.org/10.1145/3810228

further achieves a 4.4–16.4% MAE reduction over the best standard dual-view fusion baselines. Data and code are released at https://github.com/Health-HCI-Group/F3Mamba.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Dataset, remote photoplethysmography (rPPG), smartphone physiological sensing, dual-view fusion, mobile health, cardiovascular monitoring, deep Learning

ACM Reference Format:
Jiankai Tang, Tao Zhang, Jia Li, Yuntao Wang, Yiru Zhang, Mingyu Zhang, Kegang Wang, Yuming Hao, Bolin Wang, Haiyang Li, Xingyao Wang, Yuanchun Shi, and Sichong Qian. 2026. M^{3}PD Dataset: Enabling Dual-view Photoplethysmography on Smartphones in Lab and Clinical Settings. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 2, Article 61 (June 2026), 32 pages. https://doi.org/10.1145/3810228

1 INTRODUCTION

Portable health monitoring is important for everyday well-being, with heart rate (HR) serving as a key physiological indicator for cardiovascular health assessment [2, 22]. Traditional HR measurement methods rely on specialized medical devices such as electrocardiography (ECG) or photoplethysmography (PPG) sensors that are often inconvenient to carry and impractical for continuous monitoring [33]. This creates a critical gap in surveillance particularly for patients with arrhythmias, hypertension, and coronary artery disease, where cardiovascular events can occur suddenly and unpredictably [20]. By the time patients reach medical facilities, transient cardiac abnormalities may have already normalized, making timely diagnosis and appropriate treatment decisions more challenging [8].

Recent studies in ubiquitous and mobile computing have shown that smartphones themselves can be turned into physiological sensors: inertial-sensor–based methods can reconstruct ECG-like signals or estimate HR and heart rate variability (HRV) from accelerometers and gyroscopes [23, 41], and acoustic sensing can even pick up heartbeat-induced chest motion using commodity smart speakers [48]. These efforts clearly validate the idea of phone-centric health monitoring. However, to reach their reported accuracy, most of these systems still assume static or semi-static postures, stable phone–body placement, and low ambient noise. Once the phone is truly handheld, the user is elderly, or the scene contains natural head/hand movements, the inertial or acoustic channels are easily flooded by motion artifacts. This suggests that we need a more stable sensing modality that works *with* natural smartphone use rather than against it.

Video-based physiological sensing has emerged as a promising solution to address these challenges, enabling non-invasive extraction of vital signs from facial [34] or fingertip regions [12] without requiring specialized equipment. With recent advancements in smartphone camera technology and computational capabilities, an increasing number of studies have explored mobile phone-based remote physiological sensing applications that can be integrated into everyday life [18, 39]. Unlike traditional fixed-camera approaches used in clinical settings [20], smartphone cameras offer superior portability and accessibility [19], making physiological monitoring feasible across diverse environments and populations.

However, extracting physiological signals from videos using smartphone cameras still faces reliability challenges in healthcare settings. Smartphones are usually operated in a handheld manner, which introduces motion artifacts and jitter that can compromise physiological signal quality [17]. These challenges are even more pronounced for elderly users and cardiovascular disease (CVD) patients, who may find it difficult to maintain a stable posture or fixed device position during measurement.

Existing methods therefore often resort to stationary or constrained setups—such as tripod-mounted smartphones for facial monitoring [24, 29] or requiring the hand to be placed on a stable surface [1, 2]—but such assumptions do not reflect real-world mobile usage. Variations in camera-to-subject distance, viewing angle, and ambient illumination further complicate video-based physiological signal extraction [39]. These limitations are

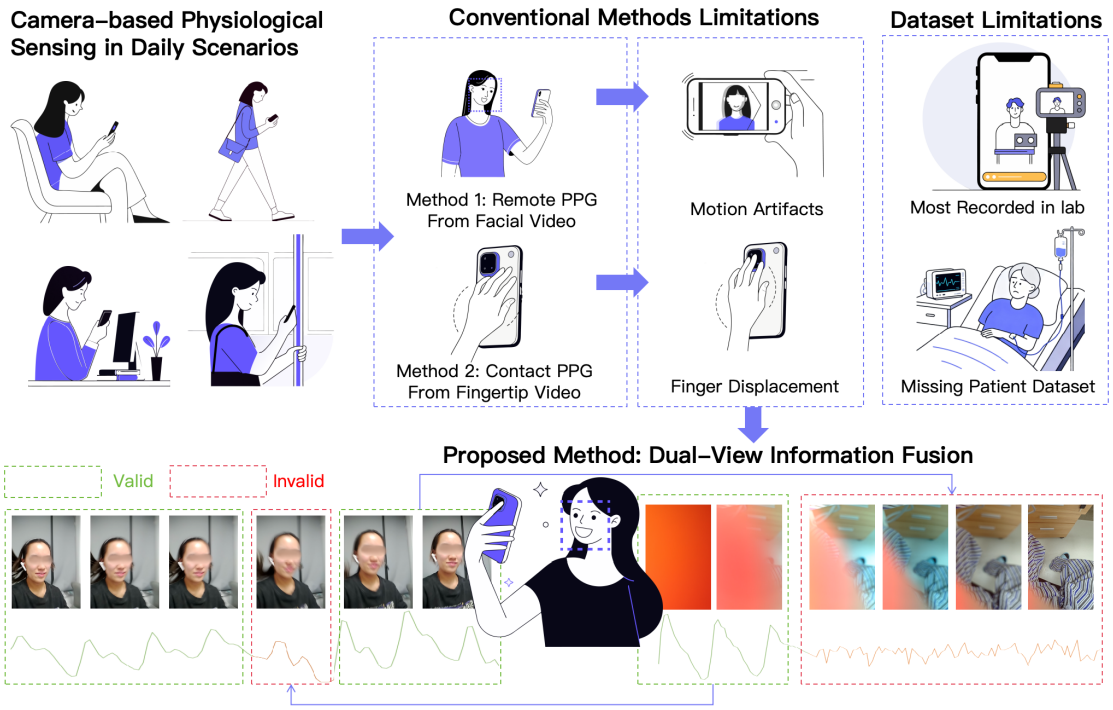


Fig. 1. **Fusion of video-based physiological sensing.** Video-based physiological sensing faces challenges from motion artifacts, lighting variations, and position instability. Traditional approaches rely on single views (facial or fingertip), limiting robustness. Our dual-position fusion method integrates signals from both front camera (facial) and rear camera (fingertip) videos. The F^{3}Mamba framework leverages this dual-view approach to enhance algorithm robustness and accuracy in heart rate estimation across real-world scenarios.

particularly problematic in clinical or pre-clinical screening scenarios, where measurement accuracy directly affects diagnostic value. Recent cross-dataset evaluations [18] on the MMPD smartphone video dataset [33] have shown that even advanced algorithms can yield heart-rate estimation errors exceeding 10 beats per minute (BPM), which falls short of the precision required for many cardiovascular assessments.

A key reason for this performance gap is that current smartphone-based physiological sensing pipelines typically rely on a single view (for example, facial [29] or fingertip [1]) and thus fail to exploit the complementary information available from multiple sensing sites. This is a critical limitation for CVD patients, whose peripheral perfusion and physiological waveforms may fluctuate across time and body locations, making one view (e.g., face) unreliable while another view (e.g., fingertip) still contains usable pulsatile components. Yet, to date, few studies have systematically explored the potential of simultaneous front–rear smartphone camera recording to improve robustness. Early work such as MobilePhys [19] showed that rear-camera signals can enhance performance but required subject-specific retraining, which restricts deployment. The most relevant study used two USB cameras on only 10 subjects [16], suggesting that multi-position video input can improve heart-rate estimation, but it relied on external cameras rather than on-board dual cameras, limiting its practicality for everyday mobile health monitoring.

To address these limitations in smartphone-based physiological monitoring—especially the lack of dual-view mobile video data for CVD patients—we introduce the **M**ulti-view **M**ulti-scenario **M**obile **P**hysiology **D**ataset

(M^3**PD**). As illustrated in Figure 1, our dual-view fusion approach integrates signals from both front camera (facial) and rear camera (fingertip) videos to enhance robustness against motion artifacts, lighting variations, and position instability. The M^3PD dataset is, to the best of our knowledge, the first publicly available smartphone dual-view physiological sensing dataset that explicitly targets handheld and clinically relevant scenarios. It contains synchronized facial and fingertip videos recorded simultaneously by front and rear smartphone cameras in both lab and clinical settings, together with clinical-grade physiological measurements including PPG waveforms, blood oxygen (SpO2), and blood pressure (BP). The dataset comprises recordings from 60 subjects, among which 47 are CVD patients, enabling validation in both laboratory (n=13) and clinical (n=47) environments. Building on this dual-view setting, we further propose the **F**acial-**F**ingertip **F**usion **Mamba** (F^3**Mamba**) framework to integrate complementary physiological information from dual-view streams. F^3Mamba dynamically updates state representations across views and performs temporal fusion through a Fusion Mamba (F-Mamba) architecture, yielding more reliable heart-rate estimates even when one view is corrupted by motion, low perfusion, or lighting artifacts.

Table 1. **Datasets comparison.** Details of wide-use video physiological sensing datasets.

Dataset	Scenarios	Subjects	Camera	Position	Vitals
PURE [31]	Lab	10	eco274CVGE	Face	PPG/SpO_2
UBFC-rPPG [3]	Lab	42	Logitech C920	Face	PPG
Oximetry [12]	Lab	6	Google Nexus 6P	Finger	SpO_2
MMPD [33]	Lab	33	Galaxy S22 Ultra	Face	PPG
RLAP [40]	Lab	58	Logitech C930c	Face	PPG
SUMS [16]	Lab	10	Logitech C922	Face+Finger	PPG/SpO_2/RR
LADH [22]	Lab	21	Logitech C922	Face(RGB+IR)	PPG/SpO_2/RR
M^3PD(**Ours**)	Lab	13	OPPO A52	Face+Finger	PPG/SpO_2/RR/BP
	Clinic	47	XiaoMi 14	Face+Finger	PPG/SpO_2/RR/BP

The main contributions of this paper are:

- We present M^3PD, the first dual-view smartphone dataset that records front-camera (face) and rear-camera (fingertip) videos from *CVD patients* (n=47) and healthy subjects (n=13). This resource addresses realistic handheld challenges faced in point-of-care cardiovascular monitoring, including motion artifacts and unstable handling, particularly among elderly CVD patients.
- We conduct an extensive benchmark of existing methods on M^3PD, identifying key reliability gaps in single-view sensing under handheld clinical conditions. Based on these insights, we propose F^3Mamba, a dual-view framework that dynamically fuses facial and fingertip information to overcome individual modality failures.
- We demonstrate the effectiveness of the dual-view paradigm through extensive validation in both lab and clinic environments. The improvements over single-view baselines underscore the clinical relevance of our approach for robust mobile health screening.

2 RELATED WORKS

2.1 Video Physiology Dataset

The development of multi-view datasets has been essential for advancing remote photoplethysmography (rPPG) in patient care applications. These datasets typically include synchronized recordings of facial videos alongside

cardiovascular measurements such as PPG waveforms and heart rate. The PURE dataset [31] represents one of the earliest contributions, featuring facial videos captured under laboratory conditions with corresponding PPG and SpO_2 measurements from pulse oximeters. The UBFC-rPPG dataset [3] expanded this approach with a larger participant pool using USB webcams. The Oximetry dataset [12] shifted focus to fingertip videos from rear smartphone cameras to predict SpO_2 levels during controlled oxygen desaturation protocols. The MMPD dataset [33] addressed variety by including multiple skin tones, lighting conditions, and movement patterns using fixed-position smartphones. The RLAP dataset [40] further improved data quality with standardized recording protocols across various scenarios.

Recent datasets have widened the scope of physiological monitoring beyond basic heart rate detection, reflecting the growing potential of non-contact sensing for thorough cardiovascular assessment. The SUMS dataset [16] introduced dual-view collection (face and fingertip) specifically designed for monitoring hypoxic conditions in highland, with oxygen saturation levels as low as 90%—medically relevant for patients with respiratory disorders. The LADH dataset [22] advanced monitoring capabilities by incorporating infrared facial recordings that maintain accuracy despite face coverings, enabling continuous physiological tracking over extended periods. These developments represent a natural progression toward integrated monitoring systems capable of assessing multiple cardiovascular parameters simultaneously, supporting more complete patient evaluation in both clinical and home settings.

However, existing datasets fail to address a key challenge in applying rPPG technology to everyday medical monitoring—the natural movement artifacts introduced during handheld smartphone use. Most current datasets rely on stationary or tripod-mounted cameras in controlled environments, creating a notable gap between laboratory performance and real-world clinical utility. While some datasets like VIPL [24] and MMPD [33] have incorporated limited handheld scenarios, they mainly feature brief, stable recordings that do not reflect typical patient usage patterns. This limitation is especially important for cardiovascular monitoring in elderly patients and those with limited dexterity, who often struggle to maintain stable device positioning during measurement.

To address this critical gap in clinical usefulness, we developed the **M**ulti-view **M**ulti-scenario **M**obile **P**hysiology **D**ataset (M^{3}PD), which captures both facial and fingertip videos using handheld smartphones in both laboratory and clinic environments. By intentionally including the natural movement patterns observed in everyday clinical practice, this dataset enables the development of more robust algorithms that can maintain accurate cardiovascular measurements despite the variable recording conditions encountered in real-world patient monitoring.

2.2 rPPG Algorithms

rPPG algorithms have evolved significantly, leveraging advancements in computer vision and signal processing to extract physiological signals from video data. These algorithms can be broadly categorized into two main approaches: traditional unsupervised methods and supervised deep-learning methods.

Traditional rPPG methods primarily rely on signal processing and color space analysis to extract physiological signals from facial videos. Verkruysse et al. [37] first demonstrated that ambient light and the green channel of RGB video can be used for remote plethysmographic imaging, laying the foundation for subsequent research. Poh et al. [28] introduced Independent Component Analysis (ICA) to separate the pulse signal from noise in webcam videos, improving robustness to environmental variations. De Haan et al. [9] proposed the CHROM method, which leverages chrominance-based signal processing to enhance pulse rate estimation accuracy under varying lighting conditions. To address motion artifacts, Wang et al. [42] introduced the Plane-Orthogonal-to-Skin (POS) algorithm, which formulates rPPG extraction as a projection problem in color space, significantly enhancing signal quality. Álvarez et al. [5] proposed Face2PPG, an unsupervised pipeline for extracting blood volume pulse signals from facial videos, further advancing the field toward practical, real-world applications.

However, these traditional methods often rely on restrictive assumptions about stable illumination and minimal motion, limiting their effectiveness in unconstrained real-world environments. To address these limitations, recent research has shifted toward deep learning approaches for enhanced robustness and generalization in variable conditions.

Early deep learning contributions include DeepPhys [6] and PhysNet [45], which pioneered end-to-end neural networks for video-based vital sign measurement with improved spatio-temporal feature learning. Subsequently, Transformer architectures were adapted to rPPG research, leveraging their ability to model long-range dependencies in temporal data. PhysFormer [46] and RhythmFormer [51] demonstrated significant accuracy improvements through effective capture of complex temporal and rhythmic patterns.

More recent advances include PhysMamba [21], which combines TD-Mamba blocks with a dual-stream SlowFast architecture to enhance local dynamics while maintaining long-range context for robust heart rate estimation. Similarly, MaKAN-Mixer [49] integrates Eulerian Video Magnification with Temporal Shift Module Amplification to enhance subtle physiological signals, while employing a Mamba-KAN Fusion Module for efficient temporal modeling and channel mixing.

Despite these advances, most existing rPPG algorithms are still designed and evaluated in a *single-view* setting: they operate either on facial videos (remote, non-contact) or on fingertip videos (contact, rear camera + flash), and their feature modeling is mainly based on (i) color-space redundancy and (ii) periodic temporal patterns. Because these principles hold for both facial and fingertip recordings, we include representative unsupervised and deep-learning baselines in our evaluation on M^3PD. Yet, to the best of our knowledge, no prior work explicitly exploits *both* synchronized views from the same smartphone to perform complementary physiological estimation, especially for low-perfusion or arrhythmic cardiovascular patients. To fill this gap, we propose F^3Mamba, which performs cross-view fusion over temporally aligned facial and fingertip streams, and we benchmark it against these widely used single-view baselines on our dual-view dataset to quantify the benefit of multi-view modeling.

2.3 Mamba Fusion

Recent advances in multimodal fusion have driven significant progress across diverse domains including medical imaging[13], autonomous driving [27], remote sensing[25], and human-computer interaction[11]. Traditional approaches to fusion include early/late integration strategies [30, 36], hybrid feature combination[7], and attention-based cross-modal interaction[15, 44]. While these methods can combine information from different modalities, they often struggle with computational efficiency and comprehensive fusion when processing high-dimensional data.

The emergence of Mamba, a state-space model (SSM) based architecture, offers a promising solution by maintaining linear time complexity while achieving better scalability than Transformers. Vision Mamba [50] first applied SSMs to visual tasks, inspiring subsequent multimodal fusion frameworks. Several recent studies have tailored Mamba for multimodal applications with notable success. Xie et al. [43] developed a cross-modal fusion Mamba specifically designed for detailed interaction between modalities. Dong et al. [10] introduced HFMamba, a lightweight network that uses dual Mamba branches to extract and hierarchically fuse complementary features from different perspectives.

In the remote sensing domain, researchers have adapted Mamba for specialized fusion needs. Peng et al. [26] created a dual-input Mamba block that dynamically combines spatial and spectral features through an interactive SSM update mechanism. Similarly, Cao et al. [4] designed a cross-attention module (Cross-SS2D) that efficiently exchanges information between multimodal data by using complementary inputs to refine SSM parameters.

While these approaches show Mamba's effectiveness in fusing homogeneous views (such as spatial and spectral features from the same region), they typically rely on strong inherent correlations between the data sources. Our task presents a different challenge: we must integrate physiological signals from two distinct locations (face and

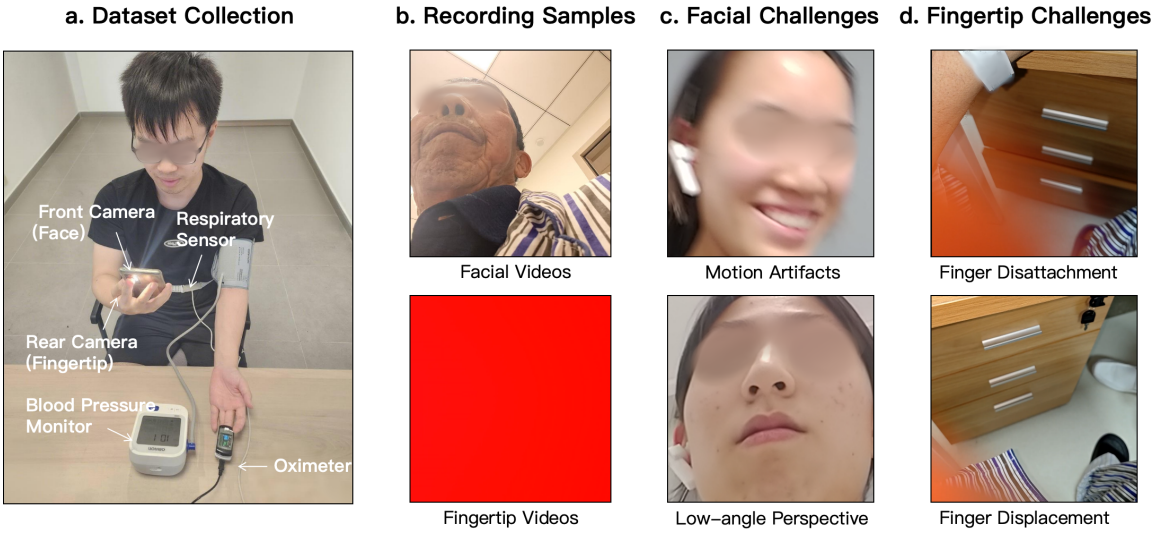


Fig. 2. **Data collection setup and real-world challenges in dual-view mobile rPPG.** (a) Synchronized data acquisition system capturing facial and fingertip videos simultaneously via front and rear smartphone cameras, with concurrent physiological measurements including respiratory sensor, blood pressure monitor, and pulse oximeter. (b) Representative recording samples showing facial videos from elderly cardiovascular patients and fingertip videos with characteristic red appearance from rear camera flash. (c) Facial video challenges during handheld recording: motion artifacts from natural head movements and low-angle perspective distortions common in patient self-monitoring. (d) Fingertip video challenges: finger disattachment from camera surface and lateral finger displacement, particularly prevalent among elderly users with limited dexterity.

fingertip) that lack direct spatial correspondence. This separation introduces unique difficulties in developing efficient state-space model interactions capable of accurately estimating heart rate from these complementary yet weakly aligned views. Unlike previous approaches designed for closely related inputs, our method must bridge the physiological gap between these different vascular regions.

3 DATASET

In this section, we introduce the M^{3}PD dataset, which is the first publicly available dual-view physiological sensing dataset captured using handheld smartphones. We start by describing the data collection system in subsection 3.1, followed by details of the lab dataset in subsection 3.2 and the Clinic dataset in subsection 3.3. We summarize the dataset characteristics in Table 1.

3.1 Collection System

This section describes the synchronized multi-modal data acquisition system used to collect the M^{3}PD dataset, including hardware components, software applications, and data synchronization methods.

3.1.1 Hardware. The hardware setup comprises a central Windows computer, an Android smartphone, and several medical-grade sensors. Two different smartphone models were used to capture data across the two study environments: an OPPO A52 in the lab and a Xiaomi 14 in the clinic. These devices feature distinct camera sensors and image signal processors (ISPs), resulting in inherent differences in color reproduction, noise characteristics, and video processing pipelines. To characterize these variations, we recorded a standard color chart with both

phones, confirming that the captured data reflects the diversity of consumer devices, as illustrated in Figure 3. This hardware variability is crucial for developing and validating rPPG algorithms that can generalize across different devices in real-world settings.

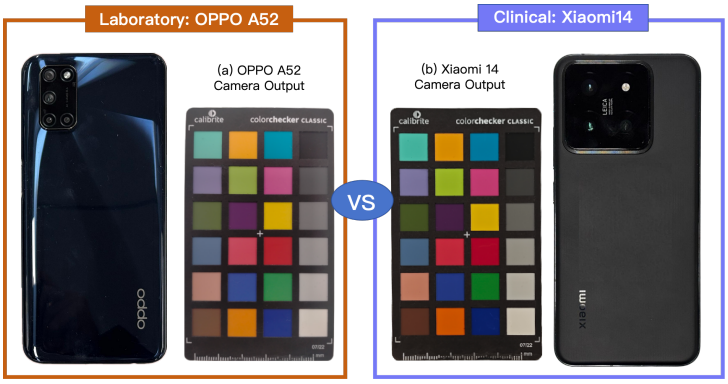


Fig. 3. **Camera color reproduction variability across devices and environments.** Comparison of ColorChecker Classic captured by (a) Xiaomi 14 in clinical settings and (b) OPPO A52 in laboratory settings. The distinct color reproductions reflect differences in camera sensors and image signal processors (ISPs) between devices, demonstrating the hardware variability that algorithms must handle for robust cross-device generalization in real-world mobile health monitoring applications.

For ground-truth physiological measurements, we used a CMS50E pulse oximeter to record PPG waveforms at 20 Hz and SpO_2 at 1 Hz, an HKH11C respiratory belt for breathing waveforms at 50 Hz, and an OMRON U726J automated cuff for blood pressure readings. Ground-truth devices are demonstrated in Figure 2(a).

3.1.2 Software. As illustrated in Figure 4, our system includes two custom software applications: a data acquisition platform on the Windows computer and a video recording application on the Android smartphone.

Inspired by PhysRecorder [40], the Windows platform provides a centralized interface for initiating and monitoring data streams from the connected medical sensors. The Android application simultaneously records video from the front (face) and rear (fingertip) cameras at a resolution of 1280x720 and a frame rate of 30 fps, embedding precise timestamps for each frame.

3.1.3 Data Synchronization. Ensuring precise temporal alignment between video streams and physiological signals is critical for rPPG research. Our system achieves this through a multi-level synchronization strategy. The smartphone and the Windows computer are synchronized to the same Network Time Protocol (NTP) server, establishing a common time reference. The medical sensors are connected to the computer via serial ports, which allows the data acquisition software to record incoming physiological data with millisecond-precision timestamps relative to the system clock. During data processing, all data streams—facial video, fingertip video, PPG, and respiration—are aligned using their respective timestamps, guaranteeing accurate correspondence between video frames and ground-truth physiological events.

3.2 Controlled Lab Dataset

3.2.1 Participants. We recruited 13 healthy adults (6 male, 7 female; age 18–30 years, mean 21.38 ± 3.78) to participate in the laboratory study approved by the institutional review board (IRB) of Tsinghua University. All participants provided written informed consent before the experiment. This subset is intended to provide a well-controlled source domain that can be contrasted with the more challenging patient recordings in subsection 3.3.

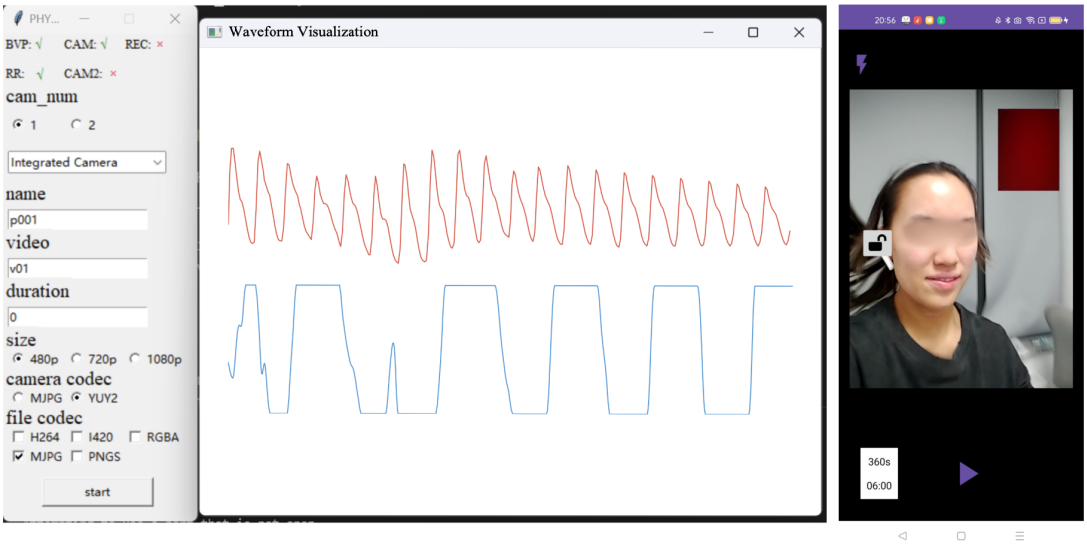


Fig. 4. **Synchronized multi-modal data acquisition system.** The system interface displays real-time physiological waveforms including blood volume pulse (BVP, top) and respiratory rate (RR, bottom) signals synchronized with simultaneous dual-view smartphone recording. The right panel shows the mobile application interface capturing both facial (front camera) and fingertip (rear camera) videos with real-time preview and recording controls.

3.2.2 Data Collection Procedure. Data were recorded using an OPPO A52 smartphone configured to capture *simultaneous* dual-view videos: the front camera recorded the participant's face, while the rear camera (with LED flash) recorded a fingertip video under contact illumination. The phone was time-synchronized with the Windows-based acquisition computer described in subsection 3.1, which received physiological signals from a CMS50E pulse oximeter (PPG waveform at 20 Hz, SpO_2 and HR at 1 Hz), an HKH11C respiratory belt (50 Hz), and an OMRON U726J automated cuff (pre/post blood-pressure readings). All videos were stored at 1280×720 resolution and 30 fps, with per-frame timestamps embedded to support alignment with the physiological streams.

The laboratory protocol was designed to emulate typical mobile cardiovascular self-monitoring behaviors and consisted of five phases (see Figure 5):

(i) 5 min seated rest (baseline), (ii) 1 min breath-holding to elicit an autonomic response, (iii) 2 min seated recovery, (iv) 1 min high leg lifts to induce exertional heart-rate changes, and (v) 4 min final seated rest for post-exertion monitoring. Each session lasted about 15 min in total, yielding roughly 13 min of effective dual-view recording per participant. During the high leg-lift phase, strong body motion caused noticeable corruption in the contact oximeter reference; therefore, this phase is *not* used in our quantitative benchmarks, but the corresponding facial and fingertip videos are kept in the released dataset to support research on motion-robust rPPG, view completion, and quality assessment.

3.2.3 Dataset Characteristics and Challenges. Although collected in a laboratory, the recordings still exhibit common real-world artifacts (see Figure 2): (1) video jitter introduced by handheld or slightly moving smartphones; (2) variations in facial pose, distance, and partial face visibility when participants adjust their sitting posture; and (3) fingertip displacement or partial detachment from the rear camera surface, especially during transitions between phases. These effects reduce the effective pulsatile component in both facial and fingertip videos and

make the lab subset more representative of actual mobile health usage than fully constrained datasets in Table 1. By releasing both the clean resting segments and the more challenging motion/transitional segments, M^{3}PD allows researchers to evaluate best-case rPPG accuracy, to test robustness to short motion bursts, and—most importantly for our work—to study whether dual-view fusion can compensate for temporary signal degradation in either view.

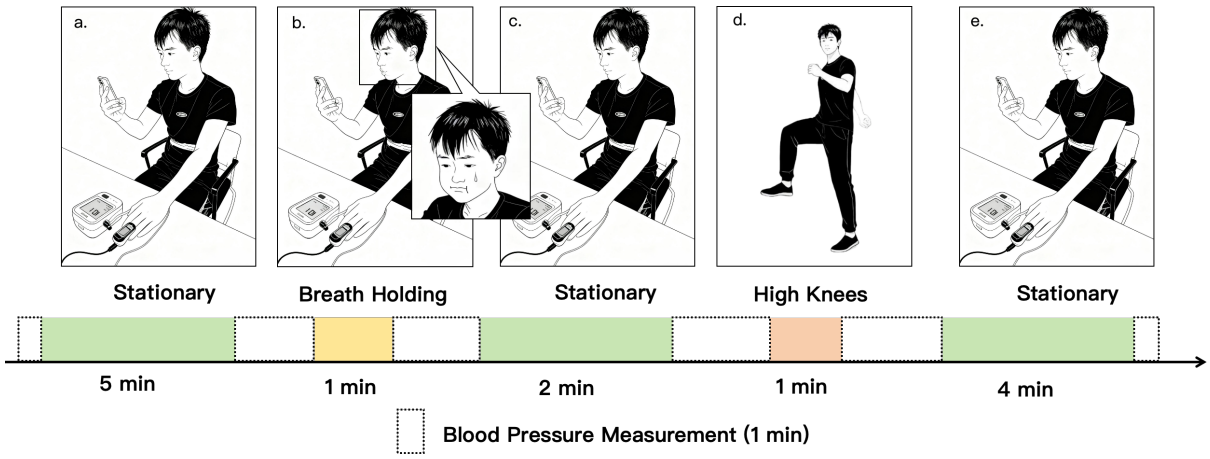


Fig. 5. **Experimental protocol for data collection.** The protocol consists of five phases designed to simulate real-world cardiovascular monitoring scenarios: baseline resting state (5 min), breath-holding for autonomic response testing (1 min), recovery period (2 min), high leg lifts for exertional heart rate changes (1 min), and final recovery phase (4 min). Blood pressure measurements were taken during the breath-holding phase to capture comprehensive cardiovascular parameters.

3.3 Clinic Dataset

3.3.1 Participants. We collected the clinical subset from 47 outpatients with documented cardiovascular conditions (30 male, 17 female; age 24–78 years, mean 60.3 ± 10.6). The cohort covered common diagnoses such as coronary artery disease, chronic heart failure, and atrial fibrillation. These conditions are characterized by unstable hemodynamics, rhythm irregularity, or reduced peripheral perfusion, all of which are known to make camera-based rPPG less reliable. All participants signed informed consent, and the study was approved by the institutional review board of the collaborating clinical site.

3.3.2 Dataset Collection Procedure. Recordings were conducted in a seated position in a real clinical environment. Each participant was asked to hold a Xiaomi 14 smartphone in front of them and look toward the screen while the device *simultaneously* captured (i) a facial video from the front camera and (ii) a fingertip video from the rear camera with flash. Importantly, we did *not* constrain how participants gripped the phone or how stably they maintained the device and head pose; small hand tremors, low-angle views, and micro-adjustments of the phone were allowed because they frequently occur in outpatient self-check scenarios. Each recording lasted about 30 s, which fits into the clinical workflow and is sufficient for heart-rate estimation from both views. The smartphone was time-aligned with the Windows-based acquisition system described in subsection 3.1, so that dual-view videos and physiological references (CMS50E PPG/HR/SpO_2, and spot BP when available) share a common timestamp. All videos were stored at 1280×720 resolution and 30 fps, identical to the lab subset to enable joint training and cross-subset evaluation.

Table 2. Sample counts for Lab and Clinic subsets.

Scenario	Facial Frames	Fingertip Frames	BVP	HR	RESP	SpO2	BP
Lab	366,501	366,718	376,942	14,369	959,573	14,369	52
Clinic	49,242	49,014	28,585	1,412	–	1,412	47

3.3.3 Dataset Characteristics and Challenges. As shown in Figure 2, the clinical recordings expose two challenges:

(1) **Physiological variability.** Many of the enrolled patients presented arrhythmias (e.g., AF), lower pulsatile amplitude, or disease-related changes in peripheral circulation. In patients with cardiovascular disease, hemodynamic alterations and impaired peripheral perfusion often lead to weak or irregular pulsations at distal sites (e.g., fingertip and face). Because PPG relies on detecting small blood-volume changes in the microvascular bed, such low-perfusion or low-pulsatility conditions reduce the signal-to-noise ratio and make beat detection less reliable, especially for camera-based rPPG. This mechanism has been reported to cause pulse underestimation and increased error in the presence of arrhythmias or peripheral vascular dysfunction, which is exactly the population we capture here. As a result, even when the medical reference reports a stable or elevated heart rate, the facial or fingertip optical signals in our dataset may show intermittently missing pulses.

(2) **Handling variability.** Because participants were not forced to fix their posture or grip, natural hand tremors, brief fingertip detachment from the rear camera, and changes in facial angle are commonly observed. These factors introduce frame-to-frame motion and view changes that are rarely seen in controlled datasets.

Unlike the lab subset, *all* clinical recordings, including those with these real-world artifacts, are included in our benchmark experiments. This design choice is intentional: mobile health systems for cardiovascular patients must operate under exactly these conditions, so we preserve them to let researchers evaluate robustness, view-level failure handling, and cross-view fusion on a realistic patient population.

3.4 Dataset Structure

The M^{3}PD dataset is organized into two main subsets: the Lab dataset and the Clinic dataset. Each subset contains synchronized facial and fingertip videos along with corresponding physiological measurements. The data sample points statistics are shown in Table 2. Demographic distribution of the dataset is illustrated in Figure 6. The dataset structure is as follows:

```
Dataset/
|-- 001/                        # subject ID
|   |-- dual_camera_session_20250115_103012/
|   |   |-- front_camera_20250115_103012.mp4     # facial view (front camera)
|   |   |-- back_camera_20250115_103012.mp4      # fingertip view (rear camera)
|   |   |-- front_camera_meta_20250115_103012.txt
|   |   `-- back_camera_meta_20250115_103012.txt
|   `-- v01/              # synchronized physiological labels
|       |-- BVP.csv
|       |-- HR.csv
|       |-- RR.csv
|       |-- SpO2.csv
|       `-- frames_timestamp.csv  # mapping video frame -> signal time
|-- 002/...
```

3.5 Data Release and Access Policy

The M^{3}PD dataset contains two types of sensitive information: (i) full facial videos that can reveal the participant's identity, and (ii) clinical physiological measurements from cardiovascular patients. For the clinical subset, all participants signed an informed-consent form that explicitly allows the use of their video and physiological data for non-commercial, academic research *under access control*, but does not permit fully public, unrestricted distribution. This means the raw data cannot be posted on open file-sharing platforms without a data-use agreement.

To balance reproducibility and privacy, we will adopt an *on-request* release model similar to widely used camera-based physiological datasets such as PURE [32] and MMPD [33]. Authorized researchers from recognized academic or medical institutions can request access by signing a joint usage and non-redistribution agreement that (1) restricts the data to non-commercial research, (2) prohibits re-identification or face recognition, and (3) forbids secondary sharing with third parties. This model has been successfully practiced in multiple international dataset projects and has been accepted by ethics boards in many institutions [14].

We will release the raw facial and fingertip video files (i.e., without blurring) together with synchronized physiological CSV files so that rPPG algorithms can be trained and evaluated fairly on the original signals. At the same time, we **require** downstream users to apply minimal privacy protection when presenting qualitative examples in papers, talks, or online materials. In particular, for patient recordings, the eye region (or the whole face when necessary) should be blurred or masked in figures to prevent casual identification. This policy preserves the scientific utility of the dataset while honoring the restrictions in the patients' consent forms.

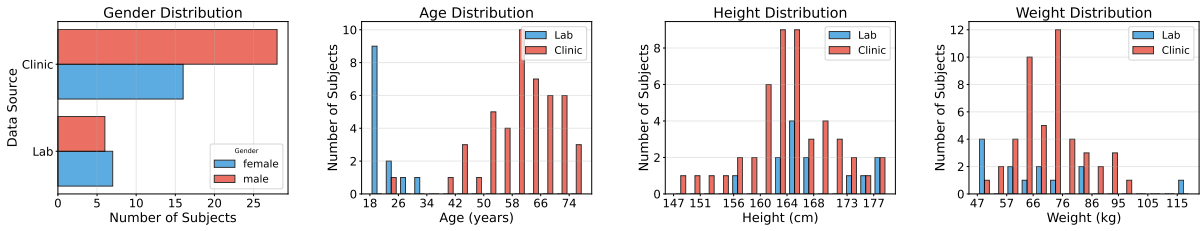


Fig. 6. **Demographic distribution of participants in the Lab and Clinic datasets.** From left to right: (a) Gender distribution, (b) Age distribution, (c) Height distribution, and (d) Weight distribution. The Clinic dataset includes a higher proportion of older participants, reflecting the target population for cardiovascular monitoring.

3.6 Motivation: Dual-view Mobile Sensing for Cardiovascular Patients

Handheld smartphone videos suffer from view-specific failure modes: facial recordings are easily corrupted by head motion or suboptimal illumination, while fingertip recordings require stable contact that many elderly or cardiovascular patients cannot consistently maintain. These failures, however, are often *complementary*—when the face view degrades, the fingertip may still retain a clean pulsatile signal, and vice versa. This observation motivates recording *simultaneous* front- and rear-camera videos so that algorithms can dynamically rely on the better view instead of a single, potentially unreliable source.

At the same time, our clinical target population is physiologically more variable than the young, cooperative subjects commonly used in prior smartphone rPPG datasets. As summarized in Table 2, the Lab subset provides long, well-instrumented dual-view sequences (over 360k facial and fingertip frames each) suitable for method development, whereas the Clinic subset contributes a large number of patient recordings with synchronized videos and medical references collected under real outpatient conditions. The demographic distribution in Fig. 6 further shows that the Lab participants are mostly young adults, while the Clinic cohort is dominated by older

patients with a much wider range of height and weight, reflecting the actual diversity of cardiovascular users. This shift in age and body habitus is clinically relevant because skin properties, peripheral perfusion, and motion control all tend to deteriorate with age.

4 METHODOLOGY

In this section, we first describe the data preprocessing steps applied to both facial and fingertip videos (Section 4.1). We then provide a detailed explanation of the network architecture (Section 4.2). Next, we discuss the view-specific branch (Section 4.3) and the Fusion Mamba block (Section 4.4) in depth. Finally, we describe the loss function (Section 4.5).

4.1 Data Processing

To prepare the dual-view videos for physiological signal extraction, we applied several preprocessing steps to standardize the input data. Both facial and fingertip videos were resized to 128 × 128 pixels to maintain computational efficiency while preserving sufficient spatial information for capturing subtle color variations related to blood volume changes. For facial videos, we additionally applied face detection and cropping to extract stable facial regions, reducing the impact of background clutter and head movements during handheld recording. The temporal length of the input video segments was set to 160 frames, corresponding to approximately 5.3 seconds at 30 fps, which provides sufficient temporal context for capturing multiple cardiac cycles while maintaining computational tractability.

In the post-processing stage, we applied a bandpass filter (0.5-3 Hz, corresponding to 30-180 BPM) to the predicted PPG waveform to remove high-frequency noise and low-frequency baseline drift. Heart rate was then estimated from the filtered signal using the Welch method for power spectral density estimation [18], identifying the dominant frequency component within the physiological range.

4.2 Facial-fingertip Fusion Mamba (F^{3}Mamba) Framework

To effectively integrate physiological signals from facial videos and fingertip videos for reliable cardiovascular monitoring, we propose a novel framework named F^{3}Mamba, as illustrated in Figure 7. This framework addresses common clinical challenges in mobile health monitoring by combining two complementary vascular beds - the facial region with rich microvasculature and the fingertip with dense capillary networks. Our design consists of parallel view-specific branches, a view fusion branch, and an rPPG predictor head. The view-specific branches process each input source using three Temporal Difference Mamba (TD-Mamba) blocks [21], while the view fusion branch uses Fusion Mamba (F-Mamba) blocks to combine their physiological information. This approach maintains measurement continuity when one signal source is temporarily compromised by patient movement.

For processing, stable facial regions are first extracted from facial videos through cropping. These facial frames and the raw finger frames are then fed into the F^{3}Mamba framework, where the DiffNormalized technique [6] extracts frame difference features to highlight subtle blood volume changes. These processed frames are sent to separate view-specific branches where a basic stem network extracts initial spatial features. Three TD-Mamba blocks then model the temporal patterns essential for capturing cardiovascular pulsations across video frames.

The multi-stage fusion branch is a key innovation that enables robust physiological monitoring in challenging daily healthcare clinical scenarios. This branch contains three F-Mamba blocks that align with the view-specific branches. Each F-Mamba block takes both facial and finger feature maps as input and creates a fused representation that captures the most reliable physiological signals from each source. These fused features connect back to their respective view branches through residual connections, allowing mutual enhancement of signal quality. After the final F-Mamba block, the combined features are processed by the rPPG predictor to generate the pulse waveform needed for heart rate calculation.

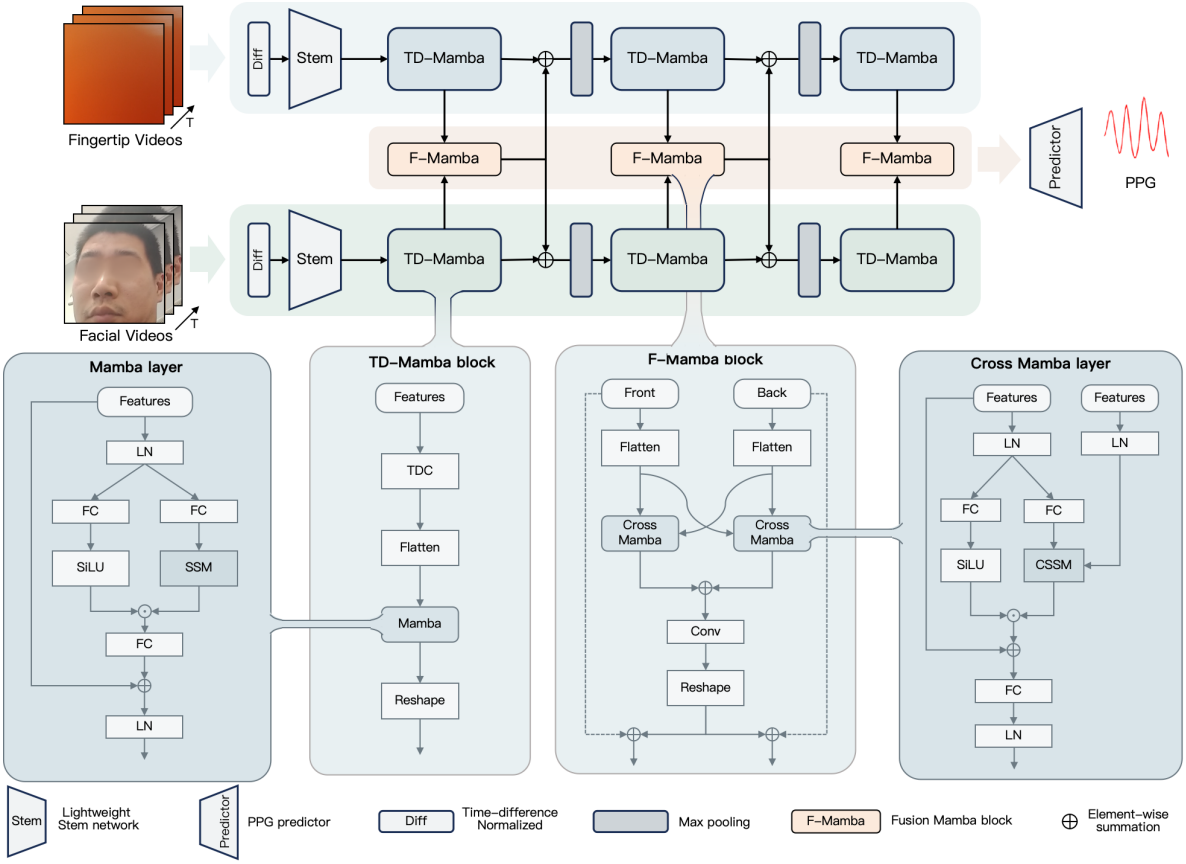


Fig. 7. **The proposed F³Mamba framework.** The architecture processes facial and fingertip videos from front and rear cameras. After differentiation-normalization preprocessing, inputs flow through view-specific branches with TD-Mamba blocks for temporal modeling of each view separately. F-Mamba blocks facilitate cross-view fusion by integrating complementary information and dynamically updating state parameters. This fusion approach maintains robustness when one view contains artifacts, ultimately outputting a PPG waveform for heart rate estimation.

4.3 Dual View (Front and Rear) Feature Extraction

The preprocessed video frames F_{pre}^{d} are processed through a stem network consisting of three simple convolutional layers to achieve spatial downsampling:

$$F_{\text{stem}}^{d} = \text{stem}(F_{\text{pre}}^{d}) \tag{1}$$

Where $F_{\text{stem}}^{d} \in \mathbb{R}^{C\times T\times H\times W}$ are the downsampled feature maps, C denotes the channel dimension, T is the temporal length, and H/W represent the downsampled height and width, respectively. The subscript d indicates the input view (face or fingertip). F_{stem}^{d} is then fed into a view-specific branch for further feature extraction.

Each branch consists of three TD-Mamba blocks, where each block integrates a temporal difference convolution (TDC) layer followed by a Mamba layer. The following provides a detailed explanation of these layers. The TDC

block efficiently captures fine-grained, local, spatio-temporal dynamics, which are crucial for tracking subtle color changes [47]:

$$TDC(F_i^d) = \underbrace{\sum_{p_n \in C} w(p_n) \cdot F_i^d(p_0 + p_n)}_{\text{vanilla 3D convolution}} + \theta \cdot \underbrace{\left(-F_i^d(p_0) \cdot \sum_{p_n \in R''} w(p_n) \right)}_{\text{temporal CD term}} \tag{2}$$

Where, $i \in \{0, 1, 2\}$ denotes the feature refinement stage and F_i^d denotes the input feature of the TD-mamba block in stage i. w is a learnable parameter, p_0, C and R'' represent the current spatio-temporal location, the sampled local neighborhood, and the sampled adjacent neighborhood, respectively. Unlike vanilla convolution, TDC explicitly models temporal correction, enhancing its ability to extract time-dependent features. The TDC-processed features $F_i^d \in \mathbb{R}^{C \times T \times H_i \times W_i}$ are flattened to $F_i^d \in \mathbb{R}^{C \times (TH_iW_i)}$ and passed through a Mamba layer:

$$x_i^d, z_i^d = \text{Linear}_x \left(\text{LN} \left(F_i^d \right) \right), \text{Linear}_z \left(\text{LN} \left(F_i^d \right) \right) \tag{3}$$

Here, LN denotes layer normalization, Linear_x and Linear_z represent two distinct fully connected layers. The transformed sequence x_i^d is then processed through a Selective State Space (SSM) block for feature extraction, yielding the output sequence:

$$y_i^d = SSM \left(x_i^d \right) \tag{4}$$

After gating with z_i^d, the sequence is processed through a fully connected layer and added to the original F_i^d. We then use layer normalization on the output, resulting in the final output of Mamba layer:

$$F_i^{Out,d} = \text{LN}(\text{Linear}_{out} \left(y_i^d \cdot \text{SiLU} \left(z_i^d \right) \right) + F_i^d) \tag{5}$$

Here, $\text{Linear}_{\text{out}}$ denotes the fully connected layer, and SiLU is the activation function. The output feature map $F_i^{\text{Out,d}} \in \mathbb{R}^{C \times (TH_iW_i)}$ from the Mamba layer is reshaped to $\mathbb{R}^{C \times T \times H_i \times W_i}$. Following each TD-Mamba block, a 2×2 max-pooling layer performs spatial downsampling, after three such TD-Mamba blocks, the final output feature map from each view branch is $F_3^{\text{Out,d}} \in \mathbb{R}^{C \times T \times \frac{H}{8} \times \frac{W}{8}}$.

4.4 Aligned Facial and Fingertip Feature Fusion

To combine the dual-view features $F_i^{Out,d}$ extracted from TD-Mamba block, we introduce the F-Mamba block. This block processes paired facial and fingertip features as inputs and generates the fused representations $F_i^f \in \mathbb{R}^{C \times T \times H_i \times W_i}$.

The F-Mamba block employs a parallel, dual-branch architecture to seamlessly integrate multi-view features. Each branch processes view-specific features $F_i^d \in \mathbb{R}^{C \times T \times H_i \times W_i}$, which are flattened along the temporal dimensions to form a new representation $\widetilde{F}_i^d \in \mathbb{R}^{C \times (TH_iW_i)}$. Cross-view interaction is facilitated through a symmetric cross Mamba layer, designed to operate on these flattened sequence representations. Unlike the standard Mamba layer, the cross Mamba layer incorporates a Cross State Space Model (CSSM), enabling bidirectional information exchange between complementary views. This mechanism dynamically updates each branch's state space, ensuring robust feature fusion, as detailed in Algorithm 1.

Variable Explanations:

Algorithm 1 CSSM Block

Input: x^a, x^b: (B, L, D) ▷ Input features
Output: y^a: (B, L, D) ▷ Fused features
$A : (B, D, N) \leftarrow \text{Parameter}_A$ ▷ Learnable state matrix
$B_a : (B, L, N) \leftarrow \text{Linear}_B(x^a)$
$B_b : (B, D, N) \leftarrow \text{Linear}_B(x^b)$
$B : (B, L, N) \leftarrow (1 - \lambda_b)B_a + \lambda_b B_b$ ▷ Fuse views
$C_a : (B, L, N) \leftarrow \text{Linear}_C(x^a)$
$C_b : (B, L, N) \leftarrow \text{Linear}_C(x^b)$
$C : (B, L, N) \leftarrow (1 - \lambda_c)C_a + \lambda_c C_b$ ▷ Fuse views
$\Delta : (B, L, D) \leftarrow \log(1 + \exp(\text{Linear}_\Delta(x^a) + \text{Parameter}_\Delta)$ ▷ Softplus discretization
$\overline{A} : (B, L, D, N) \leftarrow \exp(\Delta \otimes A)$ ▷ Discretize state transition
$\overline{B} : (B, L, D, N) \leftarrow \Delta \otimes B_a$ ▷ Discretize input matrix
$y^a \leftarrow \text{SSM}(\overline{A}, \overline{B}, C)(x^a)$ ▷ State space computation
return y^a

- x^a: Primary view input features
- x^b: Complementary view input features
- y^a: Enhanced output features of the primary view
- A: Learnable state transition matrix
- B_a: Input projection of primary view x^a
- B_b: Input projection of complementary view x^b
- C: Output projection matrix derived from x^a
- Δ: Discretization step size
- $\overline{A}$: Discretized state transition matrix
- $\overline{B}$: Discretized input matrix
- λ_b: Fusion weight for input projection
- λ_c: Fusion weight for output projection

The fusion feature $F_i^f \in \mathbb{R}^{C\times(TH_iW_i)}$ is generated by the convolutional integration of these cross-view features:

$$\widetilde{F}_i^f = \text{Conv}\left(\left[\widetilde{F}_i^{d_1\to d_2} + \widetilde{F}_i^{d_2\to d_1}\right]\right) \tag{6}$$

Here, $\widetilde{F}_i^{d_1\to d_2}$ and $\widetilde{F}_i^{d_2\to d_1}$ are outputs from the Cross Mamba layer. Each represents features where one view's state space (d_1 or d_2) is updated with information from the other view.

To maintain dimensional consistency, we reshape $\widetilde{F}_i^f$ to match F_i^d. The reshaped features are then added to the original features through residual connections:

$$F_i^d = F_i^d + \text{Reshape}\left(\widetilde{F}_i^f\right) \tag{7}$$

With $i \in 0, 1$ denoting the fusion stages. In the final stage, the F-Mamba block output F_3^f is processed through adaptive average pooling to reduce dimensions, followed by convolution to generate PPG predictions:

$$\text{PPG} = \text{Conv}(\text{AdaptiveAvgPool}(F_3^f)) \tag{8}$$

This dual-view approach combines advantages from both input sources. When facial videos contain motion artifacts or poor lighting, fingertip data with its dense capillary network provides more stable signals. When fingertip contact is unstable, facial vasculature data maintains measurement quality. This complementary design improves heart rate estimation across various real-world conditions.

4.5 Time-frequency Combined Loss Function

To ensure robust learning, the loss function combines time-domain loss and frequency-domain loss, inspired by prior work [51]. Specifically, we employ the negative Pearson (NP) loss as the time-domain loss, defined as:

$$\mathcal{L}_{time} = -\frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{9}$$

where y_i and $\hat{y}_i$ represent the ground truth and predicted signals, respectively, $\bar{y}$ and $\bar{\hat{y}}$ are their means, and N is the number of samples.

The frequency-domain loss $\mathcal{L}_{freq}$ is calculated using the cross-entropy (CE) between the aligned power spectrum of the predicted signal and the ground truth signal. To align the power spectra, the index of the maximum value in the ground truth spectrum is used as a reference. It is expressed as:

$$\mathcal{L}_{freq} = \mathrm{CE}(\mathrm{MaxIndex}(y_{PSD}, \hat{y}_{PSD})) \tag{10}$$

where MaxIndex indicates the index of the maximum value, and y_{PSD}, $\hat{y}_{PSD}$ represent the power spectrum decomposition (PSD) of the ground true signals y and the predicted signals $\hat{y}$. Additionally, we introduce a PSD distribution loss $\mathcal{L}_{PSD}$ to constrain the predicted PSD distribution to closely match the ground truth PSD distribution. This loss is defined using the Kullback-Leibler (KL) divergence as:

$$\mathcal{L}_{PSD} = \mathrm{KL}\left(P(y_{PSD}) \| P(\hat{y}_{PSD})\right) \tag{11}$$

In this equation, $P(y_{PSD})$ and $P(\hat{y}_{PSD})$ represent the probability distributions of the ground truth and predicted PSD values, respectively. The KL divergence measures the difference between these two distributions, ensuring that the predicted HR aligns closely with the ground truth. Finally, the overall loss function is formulated as a weighted sum of the three components:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{time}} + \lambda_2 \mathcal{L}_{\text{freq}} + \lambda_3 \mathcal{L}_{\text{PSD}} \tag{12}$$

Where λ_1, λ_2, and λ_3 are hyperparameters that control the relative importance of each loss component. Following the parameter selection strategy in prior work [51], we set $\lambda_1 = 0.2$, $\lambda_2 = 1$, and $\lambda_3 = 1$ in our experiments.

5 RESULTS

5.1 Experimental Settings and Evaluation Metrics

The programs were developed in Python 3.8 using PyTorch and executed on NVIDIA GeForce RTX 3090 GPUs. The TD-Mamba block followed the original settings from [21]. In the Fusion Mamba block, the CSSM dimension was set to 64 with a dropout rate of 0.1, the hyperparameters λ_b and λ_c were set to 0.5, indicating equal importance for both views in the fusion process.

We trained the model for 15 epochs with a batch size of 4, using the Adam optimizer with an initial learning rate of 5×10^{-5} and employing a OneCycleLR scheduler to adjust the learning rate dynamically. The subjects were randomly divided into three groups, and a three-fold cross-validation protocol was performed to obtain the average results.

To assess the accuracy of HR measurement, we used the mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE), and Pearson correlation coefficient (ρ) as the evaluation metrics. All reported values are rounded to three decimal places for readability. Best results are presented in **bold**, and sub-optimal results are underlined.

5.2 Evaluation of Single-view Video Physiology with Existing Methods

To establish a clear baseline for our dual-view fusion approach, we analyzed the performance of traditional unsupervised rPPG or contact PPG (cPPG) methods on single-view inputs. These methods, including GREEN [37], ICA [28], POS [42], and OMIT [5], serve as common baselines to highlight the inherent limitations of relying on a single camera view in real-world handheld scenarios. The unsupervised results are in the supplementary materials (see Appendix A).

While traditional methods provide a foundational baseline, recent advancements in deep learning have led to state-of-the-art (SOTA) performance in rPPG/cPPG. We evaluate several deep learning models, including PhysNet [45], PhysFormer [46], RhythmFormer [51], and PhysMamba [21]. For a fair comparison, we chose the common PURE [31] dataset for cross-dataset deep learning experiments.

We also incorporated supplementary standard fusion baselines, as follows: **EarlyFuse-Concat**: Concatenates dual-view features along channels at the input for unified encoding; **LateFuse-Concat**: Extracts features independently, then concatenates high-level features for inference;**LateFuse-Avg**: Generates separate predictions for each view, with the final result as their average; **SummitVital**: A dual-view two-stream method from the [16], used as a cross-method benchmark. All baselines share F^{3}Mamba's backbone (from [21]), differing only in dual-view fusion strategies:

5.2.1 Challenges with Face rPPG in Handheld Scenarios. As shown in Table 10 and Table 3, traditional methods perform poorly on facial videos in our handheld setting, with MAEs ranging from 16.5 to 32.9 BPM and near-zero correlation (ρ), reflecting high sensitivity to motion and illumination. Deep learning baselines trained on PURE also show a marked degradation on M^{3}PD; on the Lab dataset, the best model achieves an MAE of 15.0 BPM, while on the Clinic dataset, MAEs are between 23 and 29 BPM, evidencing a strong domain shift.

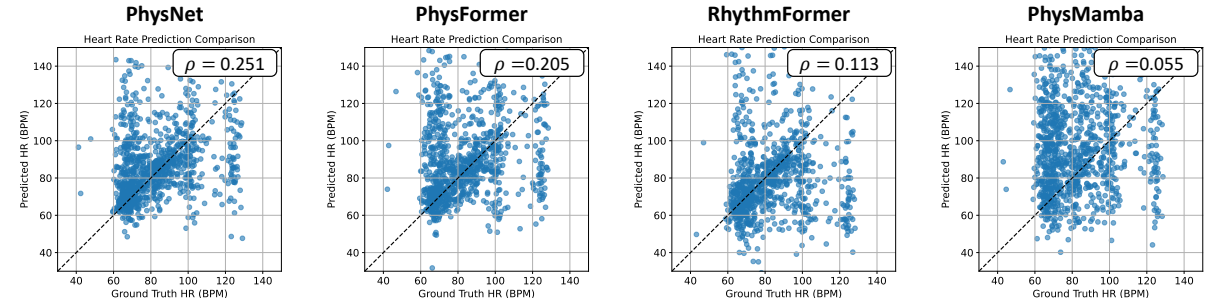
To further illustrate these limitations, we visualize the heart rate estimation results using scatter plots in Figure 12. Ideally, the estimated heart rate should be highly correlated with the reference heart rate, with data points closely distributed along the diagonal. However, the scatter plots for both datasets reveal significant deviation and dispersion, indicating poor correlation and large estimation errors. This demonstrates that single-view facial rPPG under handheld motion and lighting variations is unstable for accurate heart rate measurement.

Table 3. Performance of deep learning methods on the M^{3}PD dataset using face inputs.

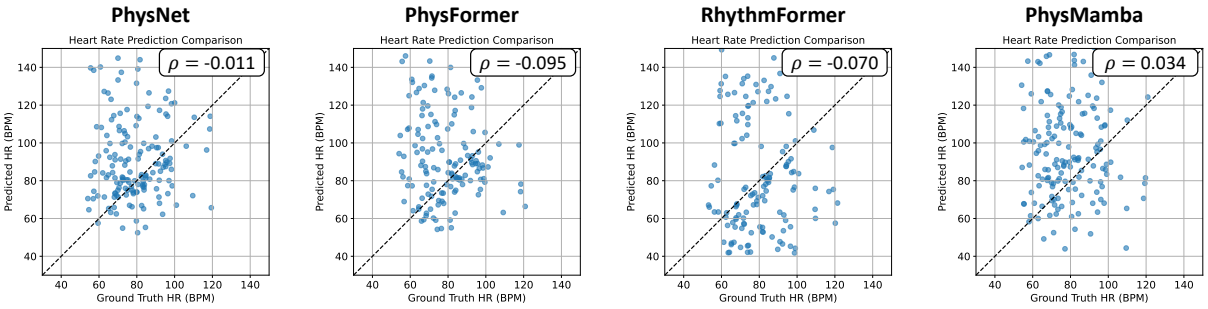
Method	PURE→Lab				PURE→Clinic			
	MAE ↓	MAPE ↓	RMSE ↓	ρ ↑	MAE ↓	MAPE ↓	RMSE ↓	ρ ↑
PhysNet	**15.041**	**18.633**	**22.501**	**0.251**	23.497	33.361	**32.426**	-0.011
PhysFormer	17.186	21.457	24.652	0.205	**23.473**	**31.296**	32.450	-0.095
RhythmFormer	17.239	20.184	25.974	0.113	29.337	38.062	38.230	-0.070
PhysMamba	24.998	31.988	32.101	0.055	26.590	33.682	35.622	**0.034**

MAE = mean absolute error in HR estimation (beats/min), RMSE = root mean square error in HR estimation (beats/min), MAPE = mean percentage error (%), ρ = Pearson correlation in HR estimation.

5.2.2 Challenges with Fingertip cPPG in Handheld Scenarios. The performance on fingertip inputs, detailed in Table 11 and Table 4, reveals that while certain methods like ICA show improved accuracy on the Lab dataset compared to facial inputs (with an MAE of 7.749 BPM), the overall results remain inconsistent across different algorithms and datasets and fail to achieve the desired high precision. This variability underscores the limitations



(a) Lab Dataset (Face inputs).



(b) Clinic Dataset (Face inputs).

Fig. 8. **Scatter plots of estimated HR vs. reference HR for face inputs.** Top: Lab; Bottom: Clinic. The diagonal indicates ideal correlation.

of relying on a single view, as factors like finger placement and pressure significantly impact signal quality, highlighting the reliability challenges that persist in practical handheld scenarios.

To further illustrate these limitations, we visualize the heart rate estimation results in Figure 13. The scatter plots reveal considerable spread and deviation, indicating that even fingertip cPPG is susceptible to significant estimation errors and poor correlation under handheld conditions. This instability is often caused by incomplete finger coverage, variable pressure, or movement during measurement, factors that are common in real-world use, especially among elderly patients. These results demonstrate that single-view fingertip cPPG, while sometimes more robust than facial rPPG, still faces significant reliability challenges in practical handheld scenarios.

5.3 F^{3}Mamba Performs best with Dual-view Fusion

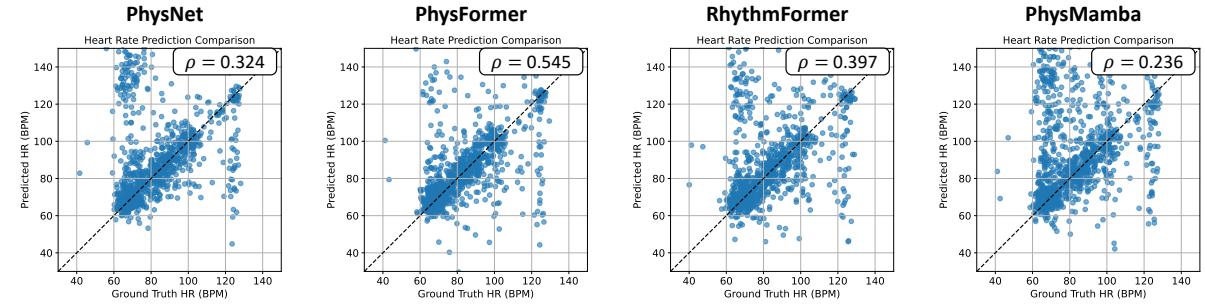
Having established the limitations of traditional single-view methods, we now evaluate the performance of modern deep learning approaches, culminating in our proposed dual-view fusion model, F^{3}Mamba. This section details the results of intra-dataset and cross-dataset testing, demonstrating the superior accuracy and robustness achieved by integrating complementary physiological signals from both facial and fingertip videos.

5.3.1 Intra-dataset Testing. We conducted intra-dataset tests on the two subsets of M^{3}PD dataset (**Lab** and **Clinic**) and compared our results with deep learning methods[21, 45, 46, 51]. All method parameters followed the original settings. We employed a three-fold cross-subject validation to prevent data leakage.

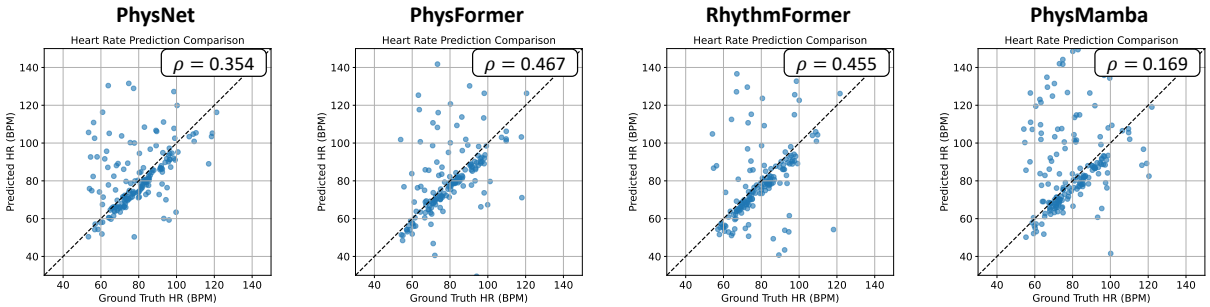
Table 4. Performance of deep learning methods on the M^{3}PD dataset using fingertip inputs.

Method	PURE→Lab				PURE→Clinic			
	MAE ↓	MAPE ↓	RMSE ↓	ρ ↑	MAE ↓	MAPE ↓	RMSE ↓	ρ ↑
PhysNet	13.021	17.017	23.689	0.324	11.312	18.859	15.581	0.354
PhysFormer	**9.637**	**11.761**	**17.333**	**0.545**	**10.532**	**17.661**	**13.694**	**0.467**
RhythmFormer	12.966	16.302	22.880	0.397	12.845	22.302	16.182	0.455
PhysMamba	17.749	22.907	28.425	0.236	16.715	26.248	22.784	0.169

MAE = mean absolute error in HR estimation (beats/min), RMSE = root mean square error in HR estimation (beats/min), MAPE = mean percentage error (%), ρ = Pearson correlation in HR estimation.



(a) Lab Dataset (Fingertip inputs).



(b) Clinic Dataset (Fingertip inputs).

Fig. 9. **Scatter plots of estimated HR vs. reference HR for fingertip inputs.** Top: Lab; Bottom: Clinic. The diagonal indicates ideal correlation.

As shown in Table 5, our model consistently outperforms existing approaches across both datasets. In the Lab dataset, our dual-view fusion approach reduces MAE by 30.2% (from 9.542 to 6.664 BPM) compared to the best single-camera baseline. In the Clinic dataset with cardiovascular patients, we observe 21.9% MAE reduction (from 9.480 to 7.405 BPM). This level of accuracy is particularly important in clinical cardiovascular monitoring, where heart rate measurement precision directly impacts arrhythmia detection and treatment decisions.

To further validate the effectiveness of our dual-view fusion strategy, we compare F^{3}Mamba against several standard fusion baselines. As shown in Table 5, our method consistently surpasses these baselines across all metrics on both datasets. On the Lab subset, F^{3}Mamba achieves an MAE of 6.664 BPM, outperforming the best baseline, LateFuse-Avg, which records 7.121 BPM. Similarly, on the Clinic subset, our model attains an MAE of 7.405 BPM, compared to 8.853 BPM for LateFuse-Avg. These results demonstrate the superior ability of our fusion approach to effectively integrate dual-view information for more accurate heart rate estimation.

Table 5. Intra-dataset testing results on subsets of M^{3}PD

Method	Input	Lab				Clinic			
		MAE ↓	MAPE ↓	RMSE ↓	ρ ↑	MAE ↓	MAPE ↓	RMSE ↓	ρ ↑
PhysNet	Face	31.651	37.350	39.238	-0.057	25.159	32.158	30.951	0.047
	Finger	10.325	10.464	19.563	**0.640**	16.476	20.971	22.738	0.385
PhysFormer	Face	23.691	27.268	28.923	0.031	19.570	26.432	23.933	0.094
	Finger	16.054	17.242	24.834	0.363	13.885	17.384	17.447	0.350
RhythmFormer	Face	26.633	30.341	34.772	0.014	28.157	37.103	34.190	-0.241
	Finger	21.790	23.571	29.379	0.025	24.107	31.836	31.081	-0.341
PhysMamba	Face	14.041	13.341	22.759	0.428	15.481	20.269	20.032	0.032
	Finger	9.542	9.247	18.088	0.630	9.480	11.411	15.524	0.460
EarlyFuse-Concat	Face+Finger	10.891	9.894	22.195	0.350	9.235	11.709	14.484	0.533
LateFuse-Concat	Face+Finger	15.930	16.820	20.965	0.045	14.066	17.443	17.463	0.152
LateFuse-Avg	Face+Finger	7.121	7.208	13.622	0.598	8.853	10.940	14.179	0.539
SummitVital	Face+Finger	6.971	7.168	13.636	0.612	9.954	12.645	15.492	0.460
F^{3}Mamba **(Ours)**	Face+Finger	**6.664**	**6.859**	**12.796**	0.636	**7.405**	**9.308**	**10.669**	**0.753**

MAE = mean absolute error in HR estimation (beats/min), RMSE = root mean square error in HR estimation (beats/min), MAPE = mean percentage error (%), ρ = Pearson correlation in HR estimation.

We visualize the predicted and ground-truth waveforms in Figure 10. The first row shows results using facial video input, while the second row shows fingertip video input results. The waveforms of our model closely match the ground truth, demonstrating accurate heart rate estimation through view fusion. Additionally, the Bland-Altman plot in Figure 11 shows that 91.01% of measurements fall within the 95% confidence interval, with color intensity indicating data density. The balanced distribution around zero suggests no systematic bias, confirming measurement reliability for clinical applications.

Existing methods performed poorly on facial video inputs across both datasets due to two key clinical challenges. First, handheld smartphone recordings introduce motion artifacts that distort facial blood volume signals, a common issue in patient self-monitoring. Second, non-frontal recording angles (often from below) caused image distortion, mimicking real-world scenarios where patients struggle to maintain ideal device positioning.

When comparing input views, fingertip videos produced better results than facial videos across both datasets. This is consistent with physiological principles, as fingertips contain denser capillary networks with more direct hemodynamic signals and fewer confounding factors than facial regions. However, fingertip-based models still faced practical limitations in clinical use: patients often could not maintain steady contact with the rear camera, causing light interference and signal loss. Additionally, most algorithms were originally designed for facial video processing and could not fully capture the unique vascular properties of fingertip signals. Our fusion

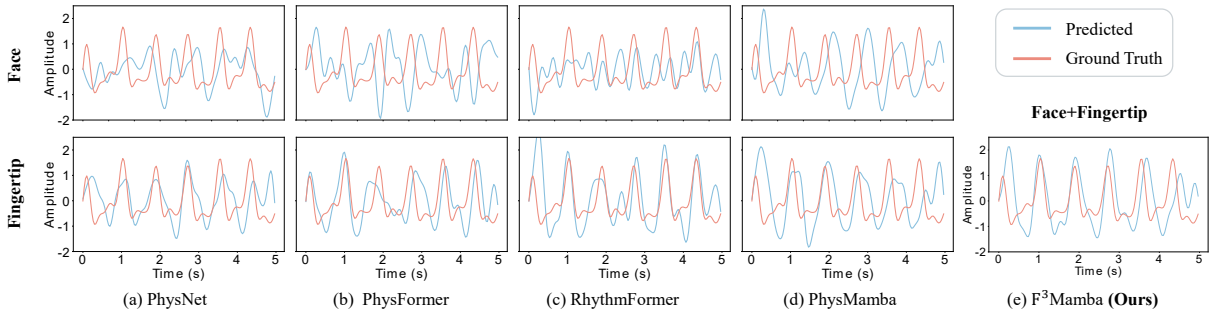


Fig. 10. **Comparison between the predicted and ground-truth waveforms in the Lab dataset.** First row: Use facial view as input, second row: Use fingertip view as input. The predicted PPG waveforms generated by our F^{3}Mamba framework show better consistency and accuracy.

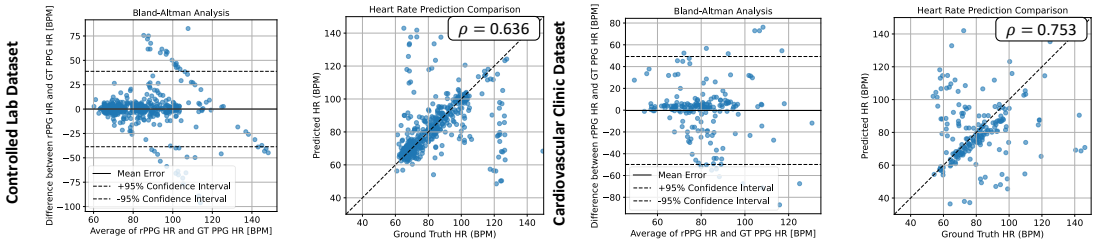


Fig. 11. **Bland-Altman analysis of F^{3}Mamba on intra-dataset M^{3}PD**: (a) Lab dataset and (b) Clinic dataset. Mean error and 95% limits of agreement are shown. Lighter colors indicate denser distributions; small random jitter is added to reduce overlap.

approach addresses these complementary limitations by maintaining measurement continuity when either source is temporarily compromised.

5.3.2 Cross-dataset Testing. To evaluate the reliability of our model in different clinical environments, we conducted two cross-dataset experiments: (1) training on the Lab dataset and testing on the Clinic dataset (Lab→Clinic), and (2) training on the Clinic dataset and testing on the Lab dataset (Clinic→Lab). These experiments simulate real-world healthcare scenarios where monitoring systems must work across varied patient populations and recording conditions.

As shown in Table 6, in the Lab→Clinic setting, our model achieves the best performance across all metrics, with an MAE of 8.204 BPM and a ρ of 0.644. For comparison, the best-performing single-view model, the fingertip-based PhysMamba, achieves an MAE of 8.629 BPM and a ρ of 0.599. The relatively small performance decrease from intra-dataset testing (with an MAE of 7.405 BPM) to cross-dataset testing (with an MAE of 8.204 BPM) highlights the adaptability of our F^{3}Mamba model to new clinical environments—a crucial feature for practical cardiac monitoring systems.

Among the fusion baselines, EarlyFuse-Concat performs second best, with an MAE of 8.250 BPM and a ρ of 0.623. In contrast, LateFuse-Concat and LateFuse-Avg show larger errors and lower correlations, indicating that

Table 6. Cross-dataset testing results on subsets of M^{3}PD.

Method	Input	Lab → Clinic				Clinic → Lab			
		MAE ↓	MAPE ↓	RMSE ↓	ρ ↑	MAE ↓	MAPE ↓	RMSE ↓	ρ ↑
PhysNet	Face	21.177	29.079	28.409	0.322	27.234	34.283	34.287	-0.143
	Finger	24.383	31.654	38.786	0.102	18.537	21.579	27.728	0.143
PhysFormer	Face	15.926	21.773	19.831	0.106	18.771	23.137	23.594	0.008
	Finger	14.673	19.173	19.693	0.099	15.789	19.238	22.590	0.120
RhythmFormer	Face	18.431	23.582	26.705	0.263	21.250	25.244	27.542	-0.041
	Finger	15.489	19.822	19.413	-0.090	19.160	22.908	25.616	-0.085
PhysMamba	Face	12.352	16.917	16.776	0.274	14.053	16.740	19.352	0.218
	Finger	8.629	10.840	12.850	0.599	**8.522**	**9.302**	15.640	0.523
EarlyFuse-Concat	Face+Finger	8.250	10.841	12.634	0.623	10.091	11.085	17.347	0.439
LateFuse-Concat	Face+Finger	14.788	16.857	19.404	-0.184	16.357	19.126	21.108	-0.046
LateFuse-Avg	Face+Finger	8.974	10.757	16.494	0.358	9.624	10.011	18.527	0.337
SummitVital	Face+Finger	10.445	11.627	18.439	0.409	8.515	9.451	16.589	**0.575**
F^{3}Mamba **(Ours)**	Face+Finger	**8.204**	**10.115**	**12.383**	**0.644**	9.360	10.938	**15.059**	0.546

MAE = mean absolute error in HR estimation (beats/min), RMSE = root mean square error in HR estimation (beats/min), MAPE = mean percentage error (%), ρ = Pearson correlation in HR estimation.

simple concatenation or averaging of high-level features may not effectively handle the domain shift between controlled lab and clinical environments. The specialized model SummitVital also underperforms in this setting.

In the Clinic→Lab setting, our model performs well for stability and correlation metrics (best RMSE and ρ), while slightly trailing PhysMamba in average error metrics. This pattern reflects several clinical monitoring challenges: the Clinic dataset contains less total monitoring time (24 minutes vs. 195 minutes), limiting training data; elderly cardiovascular patients exhibit more hand tremors affecting video quality; and these patients have difficulty maintaining stable camera contact. While our fusion approach mitigates these issues, the dataset differences still impact performance. These results highlight both the challenges and potential of smartphone-based vital sign monitoring across diverse patient populations.

While standard fusion methods struggle to handle these domain shifts effectively, our F^{3}Mamba still demonstrates competitive generalization, outperforming all fusion baselines in overall correlation and error stability. Notably, SummitVital achieves the best ρ (0.575) among baselines but shows higher RMSE, suggesting a trade-off between correlation and error magnitude that our method better balances.

5.3.3 Ablation Study on F^{3}Mamba Components. This section evaluates the importance of various components within the F^{3}Mamba model through a series of ablation studies. We constructed distinct model variants (denoted V0-V5) by adjusting or removing key modules, and subsequently performed a detailed comparative analysis. Table 7 presents the results of the ablation experiments. All ablation studies were conducted on the Lab dataset using a three-fold cross-subject-validation protocol to obtain the average results.

Impact of View Quantity: Dual-view Outperforms Single-view, with Fingertip View Being More Informative: Comparisons between single-view (V0 and V1) and dual-view (V2-V5) variants emphasise the importance of the quantity of views. The fingertip view (V1) consistently outperforms the face view (V0) among single-view variants, and dual-view variants outperform any single-view counterpart. The basic dual-view variant

Table 7. Ablation study on various components inside F^{3}Mamba.

Type	Face Stream	Finger Stream	Fusion Stream	CSSM	MAE ↓	MAPE ↓	RMSE ↓	ρ ↑
V0	✓	×	×	×	14.510	13.810	24.078	0.353
V1	×	✓	×	×	11.132	10.764	21.304	0.538
V2	✓	✓	×	×	9.256	8.994	17.614	0.568
V3	✓	✓	×	✓	7.184	7.229	13.951	0.578
V4	✓	✓	✓	×	7.537	7.556	14.338	0.541
V5	✓	✓	✓	✓	**6.664**	**6.859**	**12.796**	**0.636**

MAE = mean absolute error in HR estimation (beats/min), RMSE = root mean square error in HR estimation (beats/min), MAPE = mean percentage error (%), ρ = Pearson correlation in HR estimation.

Table 8. Performance comparison of standard fusion baselines and F^{3}Mamba on SpO_2 estimation.

Method	Lab				Clinic				SUMS			
	MAE ↓	MAPE ↓	RMSE ↓	ρ ↑	MAE ↓	MAPE ↓	RMSE ↓	ρ ↑	MAE ↓	MAPE ↓	RMSE ↓	ρ ↑
SummitVital	1.078	1.111	1.263	0.026	1.386	1.450	1.604	-0.051	3.881	2.357	5.707	-0.003
F^{3}Mamba **(Ours)**	1.084	1.117	1.270	0.009	1.357	1.421	1.577	0.005	3.118	2.250	4.735	0.009

(V2) reduces the MAE by 17% compared to the best single-view variant (V1). This confirms that multi-view inputs enhance the robustness of physiological signals by leveraging complementary spatiotemporal information, thereby improving performance.

Effectiveness of Hierarchical Fusion: Progressive Cross-view Interaction Surpasses Single-stage Fusion: Although fusing dual-view features at the final layer (V2) is better than using a single view, it still results in relatively high MAE of 9.26 BPM, indicating that cross-view features are not sufficiently aligned via one-time fusion. Conducting multi-stage fusion after each TD-Mamba layer reduces the MAE to 7.54 BPM (18.5% lower than V2), while retaining the complementary advantages of dual views. This shows that, through 'early interaction + progressive alignment', hierarchical fusion continuously optimises cross-view information at different feature abstraction levels and suppresses view-specific noise more effectively than one-time fusion only at the final layer.

Effectiveness of CSSM: Dynamic cross-stream regulation improves fusion: Adding CSSM to single-stage fusion variants (V3 vs. V2) reduces the MAE by 22.4% (from 9.256 to 7.184 BPM). With hierarchical fusion variants (V5 vs. V4), CSSM lowers MAE by a further 6.2% (from 7.537 to 6.664 BPM), confirming its key role in robust multi-scale fusion. From a mechanistic perspective, CSSM adaptively updates the state and projection matrices of one stream using cues from the paired stream. This enables reliability-aware gating when a view is corrupted by motion or illumination.

5.4 Other Vital Signs

We extended our F^{3}Mamba framework to estimate blood oxygen saturation (SpO_2) alongside HR using dual-view fusion. The model architecture remains unchanged, with the final regression layer modified to output both BVP and SpO_2 values. We employed a multi-task loss function in [16], which combines MSE losses for both BVP and SpO_2:

$$\mathcal{L}_{multi} = MSE_{BVP} \times 100 + MSE_{SpO_2} \times (100 - Mean(SpO_2)) \tag{13}$$

We compared our method against the SummitVital fusion baseline on the Lab, Clinic, and SUMS datasets. As shown in Table 8, our model achieves comparable performance to SummitVital across all datasets, with slight improvements in the Clinic and SUMS datasets. This demonstrates the versatility of our fusion approach for estimating multiple vital signs from dual-view smartphone videos.

5.5 Computational Cost

In this section, we compare the computational complexity of our model with existing SOTA methods. The results are presented in Table 9. All metrics are calculated under a standardized input dimension of $160 \times 3 \times 128 \times 128$. The number of parameters (Param), floating-point operations (FLOPs), and storage requirements are reported for each method.

Table 9. Complexity comparison.

Methods	Param(M) ↓	FLOPs(G) ↓	Storage(MB) ↓	Latency(ms) ↓
PhysNet	8.85	70.32	3.38	16.99
PhysFormer	73.81	38.53	12.69	15.71
RhythmFormer	33.26	49.53	75.8	17.07
PhysMamba	7.59	60.40	2.90	23.74
F^{3}Mamba **(Ours)**	13.87	113.46	5.29	26.70

The latency is calculated using 160 frames as input on single RTX 3090 GPU.

Our model, F^{3}Mamba, has a total of 13.87 million parameters, which is higher than PhysNet and PhysMamba, but significantly lower than PhysFormer and RhythmFormer. The FLOPs of our model are higher than PhysMamba, due to two factors: the introduction of a new view and the addition of cross-view fusion. However, the performance gains justify the increased computational cost. The storage requirement of our model is 5.29 MB, which is also well-suited for deployment on mobile devices, ensuring practicality and efficiency. We further deployed F^{3}Mamba on an iPhone 17 Pro and verified on-device feasibility, achieving an average inference latency of 819.47 ± 11.49 ms per 160-frame chunk (mean±SD, $n = 5$). Since recording 160 frames at 30 fps takes ~5.33 s, the measured latency is well below the acquisition time, indicating practical potential for real-time operation via pipelined execution (processing chunk t while capturing chunk t+1). The real-device test video is provided in the Supplemental Video.

6 DISCUSSION

6.1 Dual-view Video-based Physiology Sensing in Mobile Scenarios

A central design choice in M^{3}PD is to record facial and fingertip videos *at the same time* from a single smartphone in mobile scenarios. This is not a "more data is better" decision; rather, it is motivated by reliability under real mobile failures, where one view can degrade while the other remains usable. In real mobile use, the two views behave very differently: facial rPPG is convenient, contactless, and works well when the user is looking at the screen, but it is sensitive to head pose, distance, and illumination changes, as illustrated by the motion artifacts in Figure 2 (c) and the poor performance of single-view facial methods in Table 3; fingertip cPPG, in contrast, usually has a stronger pulsatile component because of the flash-assisted, contact illumination, but it is more vulnerable to hand instability and partial detachment, which are common in older or clinical users (Figure 2 (d)). By capturing both views in parallel we can (i) study view-specific failure modes on the *same* heartbeat, (ii) train fusion models that fall back to the remaining view when one signal is degraded, and (iii) quantify how much accuracy actually improves when the two optical channels are combined. The superior performance of our dual-view model, shown in Table 5 and Table 6, validates this approach. This is particularly valuable for continuous smartphone usage

(e.g., reading and messaging), where users naturally expose the front camera but can occasionally place their finger on the rear camera, making opportunistic dual-view monitoring feasible without extra hardware.

6.2 Including Cardiovascular Disease Patients in the Video Physiology Measurements

Most existing video-based physiological datasets—e.g., PURE or other webcam-based collections—are dominated by young, healthy volunteers under good perfusion and controlled lighting. As our cross-dataset experiments show (Table 3, Table 4), the models trained solely on such data often fail in real telemedicine scenarios because cardiovascular patients present exactly the opposite characteristics: arrhythmias, reduced stroke volume, peripheral vasoconstriction, as evidenced by the HRV metrics in Table 12. These factors, combined with difficulties in holding the phone steadily (Figure 2), lead to weak or intermittent optical pulses at the face or fingertip and therefore to systematic HR underestimation or beat omission. Moreover, in arrhythmia populations, a pulse-based reference may not perfectly reflect beat-level rhythm due to pulse deficit, which motivates future ECG-enabled validation for finer-grained rhythm analysis. By explicitly adding 47 cardiovascular patients into M^{3}PD and keeping their recordings *in* the benchmark, we expose rPPG algorithms to the population that needs remote monitoring the most. This also makes the dataset more trustworthy for clinicians and nurses, since they can verify algorithm behavior on elderly and symptomatic subjects rather than extrapolating from student volunteers.

6.3 Reliable Video-based Pysiological Sensing with F^{3}Mamba Model

Our experimental results underscore the limitations of single-view approaches and highlight the significant benefits of dual-view fusion for robust heart rate monitoring in real-world settings. As shown in Table 5 and Table 6, single-view models, whether using facial or fingertip videos, struggle to achieve consistent accuracy, particularly in the challenging Clinic dataset. Facial rPPG is highly susceptible to motion artifacts and lighting variations common in handheld use, leading to high error rates. While fingertip cPPG generally performs better due to stronger signal quality, it is still vulnerable to signal loss from unstable finger contact, a common issue with elderly patients. Our ablation study (Table 7) confirms that even a simple fusion of both views outperforms the best single-view model. By integrating complementary information, our full F^{3}Mamba model achieves a 21.9–30.2% reduction in MAE compared to the best single-view baseline, and further reduces MAE by 4.4–16.4% over the best standard dual-view fusion baselines, demonstrating that the gains are not merely from using two synchronized streams but from the designed fusion mechanism.

6.4 Enable Broader Applications with M^{3}PD Datasets

Although we position M^{3}PD mainly as an rPPG benchmarking resource, its synchronized dual-view (face–fingertip) videos together with physiological references make it useful well beyond single-task HR estimation. Each short capture can be leveraged in daily smartphone interactions to accumulate longitudinal cardiovascular records *without* requiring 24/7 wearables, which is attractive for older or cardiac patients who cannot tolerate continuous devices. The fingertip (rear-camera) channel further enables pulse-presence and signal-quality checks [38], so that a system can fall back to the more reliable view or trigger an alert when the facial signal is corrupted. In addition, the dataset contains respiration, and spot BP labels [35], making it possible to train multi-task models that (i) estimate HR while conditioning on posture/activity, and (ii) learn harder or sparser targets (e.g., BP trend) from the dual-view video streams by exploiting the natural inter-site PPG delay between the facial and fingertip vascular beds [29]. In other words, M^{3}PD is not limited to "HR from face," but can serve as a seed dataset for opportunistic, phone-based cardiopulmonary sensing and for building mobile health applications that combine convenience, redundancy, and clinical relevance.

6.5 Limitations

Our study has several limitations that suggest directions for future work. First, while the M^{3}PD dataset includes a valuable clinical cohort of 47 cardiovascular patients, its scale and diversity are still limited. It does not fully represent the wide spectrum of cardiac conditions, and transient events like short atrial fibrillation bursts are sparse [20]. The current cohort also has a narrow skin-tone range, which may limit the generalization of models trained on this dataset, as camera-based rPPG performance is known to be affected by skin pigmentation [33, 34]. While we demonstrated on-device feasibility across three platforms (iPhone 17 Pro, OPPO A52, and Xiaomi 14), hardware dependence remains incompletely characterized. The two data-collection devices exhibited measurable inter-device differences in color reproduction and temporal noise profiles—confirmed via color-chart recordings—yet the degree to which ISP-induced spectral shifts affect rPPG accuracy across a broader device ecosystem has not been systematically quantified. To address this, future work should evaluate against mobile phone benchmarks such as MMPD [33], and explore mitigation strategies including radiometric calibration and ISP-aware data augmentation to improve cross-device generalization. At the same time, the ground-truth modalities, while clinically appropriate for short sessions, have their own limitations. The reliance on pulse oximetry and spot-check blood pressure, without continuous ECG, restricts beat-level arrhythmia analysis and the development of continuous BP estimation models. Future work should aim to expand the dataset with a more diverse patient population (in terms of both clinical conditions and skin tones) and incorporate ECG for more precise validation of cardiac rhythms [41]. Extending the recording duration in longitudinal studies would also be crucial for capturing transient cardiovascular events and assessing long-term trends.

7 CONCLUSION

In this paper, we presented M^{3}PD, the first smartphone *dual-view* mobile physiological sensing dataset that simultaneously covers a controlled lab study (13 healthy adults) and, 47 cardiovascular patients in clinical cohort. By capturing handheld operation, device heterogeneity, and disease-related physiological variability, M^{3}PD fills a missing benchmark for realistic mobile rPPG evaluation. Built on this dataset, we further proposed F^{3}Mamba, a facial–fingertip fusion framework that uses TD-Mamba branches and an F-Mamba fusion module to dynamically rely on the cleaner view, which reduces heart-rate error by **21.9–30.2%** compared to single-view baselines, and further achieves a **4.4–16.4%** MAE reduction over the best standard dual-view fusion baselines. We hope that releasing M^{3}PD together with F^{3}Mamba will encourage the community to move toward multi-view, cross-device, and clinically grounded mobile physiological sensing.

Acknowledgement

This work is supported by the National Key R&D Program of China under Grant No.2024YFB2808800 and No.2024YFB2808803.

LLM Usage Clarification

We used ChatGPT 5.0 for grammar checking and language polishing to improve the clarity and readability of the manuscript. We used ChatGPT 5.0 for generating illustrative simulated character images in Figure 1 and Figure 5, to visually support the manuscript.

References

[1] Md Hasanul Aziz, Md Kamrul Hasan, Arafat Mahmood, Richard R Love, and Sheikh Iqbal Ahamed. 2021. Automated cardiac pulse cycle analysis from photoplethysmogram (PPG) signals generated from fingertip videos captured using a smartphone to measure blood hemoglobin levels. *IEEE Journal of Biomedical and Health Informatics* 25, 5 (2021), 1385–1396.

[2] Haoyu Bian, Bin Guo, Sicong Liu, Yasan Ding, Shanshan Gao, and Zhiwen Yu. 2024. UbiHR: Resource-efficient Long-range Heart Rate Sensing on Ubiquitous Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 163 (Nov. 2024), 26 pages.

doi:10.1145/3699771

[3] Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. 2019. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters* 124 (2019), 82–90.

[4] Mingxiang Cao, Weiying Xie, and Yunsong Li. [n. d.]. M3amba: CLIP-Driven Mamba Model for Multi-Modal Remote Sensing Classification. ([n. d.]).

[5] Constantino Alvarez Casado and Miguel Bordallo López. 2023. Face2PPG: An unsupervised pipeline for blood volume pulse extraction from faces. *IEEE Journal of Biomedical and Health Informatics* 27, 11 (2023), 5530–5541.

[6] Weixuan Chen and Daniel McDuff. 2018. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*. 349–365.

[7] Jae-Ho Choi, Ki-Bong Kang, and Kyung-Tae Kim. 2024. Fusion-vital: video-RF fusion transformer for advanced remote physiological measurement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 1344–1352.

[8] Jo Woon Chong, David D. McManus, and Ki H. Chon. 2013. Arrhythmia discrimination using a smart phone. In *2013 IEEE International Conference on Body Sensor Networks*. 1–4. doi:10.1109/BSN.2013.6575493

[9] Gerard De Haan and Vincent Jeanne. 2013. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering* 60, 10 (2013), 2878–2886.

[10] Hongwei Dong, Shun Na, and Fengye Hu. 2025. Hybrid-Fusion Mamba for Multitask Point Cloud Learning With Visual Perception Sensors. *IEEE Internet of Things Journal* 12, 8 (2025), 10183–10193. doi:10.1109/JIOT.2024.3512598

[11] Baole Fu, Wenhao Chu, Chunrui Gu, and Yinhua Liu. 2024. Cross-Modal Guiding Neural Network for Multimodal Emotion Recognition From EEG and Eye Movement Signals. *IEEE Journal of Biomedical and Health Informatics* (2024).

[12] Jason S Hoffman, Varun K Viswanath, Caiwei Tian, Xinyi Ding, Matthew J Thompson, Eric C Larson, Shwetak N Patel, and Edward J Wang. 2022. Smartphone camera oximetry in an induced hypoxemia study. *NPJ digital medicine* 5, 1 (2022), 146.

[13] Huafeng Li, Dayong Su, Qing Cai, and Yafei Zhang. 2025. Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 4725–4733.

[14] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. 2018. The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 242–249.

[15] Tengfei Liang, Yi Jin, Wu Liu, and Yidong Li. 2023. Cross-modality transformer with modality mining for visible-infrared person re-identification. *IEEE Transactions on Multimedia* 25 (2023), 8432–8444.

[16] Ke Liu, Jiankai Tang, Zhang Jiang, Yuntao Wang, Xiaojing Liu, Dong Li, and Yuanchun Shi. 2024. Summit Vitals: Multi-Camera and Multi-Signal Biosensing at High Altitudes. In *The 21st IEEE International Conference on Ubiquitous Intelligence and Computing (UIC 2024)*.

[17] Mingxuan Liu, Jiankai Tang, Haoxiang Li, Jiahao Qi, Siwei Li, Kegang Wang, Yuntao Wang, and Hong Chen. 2024. Spiking-PhysFormer: Camera-Based Remote Photoplethysmography with Parallel Spike-driven Transformer. *Neural Networks* (2024).

[18] Xin Liu, Girish Narayanswamy, Akshay Paruchuri, Xiaoyu Zhang, Jiankai Tang, Yuzhe Zhang, Roni Sengupta, Shwetak Patel, Yuntao Wang, and Daniel McDuff. 2023. rppg-toolbox: Deep remote ppg toolbox. *Advances in Neural Information Processing Systems* 36 (2023), 68485–68510.

[19] Xin Liu, Yuntao Wang, Sinan Xie, Xiaoyu Zhang, Zixian Ma, Daniel McDuff, and Shwetak Patel. 2022. MobilePhys: Personalized Mobile Camera-Based Contactless Physiological Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 24 (mar 2022), 23 pages. doi:10.1145/3517225

[20] Xuenan Liu, Xuezhi Yang, Dingliang Wang, Alexander Wong, Likun Ma, and Longwei Li. 2022. VidAF: A Motion-Robust Model for Atrial Fibrillation Screening From Facial Videos. *IEEE Journal of Biomedical and Health Informatics* 26, 4 (April 2022), 1672–1683. doi:10.1109/JBHI.2021.3124967

[21] Chaoqi Luo, Yiping Xie, and Zitong Yu. 2024. PhysMamba: Efficient Remote Physiological Measurement with SlowFast Temporal Difference Mamba. In *Chinese Conference on Biometric Recognition*. Springer, 248–259.

[22] Xulin Ma, Jiankai Tang, Zhang Jiang, Songqin Cheng, Yuanchun Shi, Dong LI, Xin Liu, Daniel McDuff, Xiaojing Liu, and Yuntao Wang. 2025. Non-Contact Health Monitoring During Daily Personal Care Routines. arXiv:2506.09718 [cs.CV]

[23] Reham Mohamed and Moustafa Youssef. 2017. HeartSense: Ubiquitous Accurate Multi-Modal Fusion-based Heart Rate Estimation Using Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 97 (Sept. 2017), 18 pages. doi:10.1145/3132028

[24] Xuesong Niu, Hu Han, Shiguang Shan, and Xilin Chen. 2018. VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In *Asian conference on computer vision*. Springer, 562–576.

[25] Siran Peng, Chenhao Guo, Xiao Wu, and Liang-Jian Deng. 2023. U2net: A general framework with spatial-spectral-integrated double u-net for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3219–3227.

[26] Siran Peng, Xiangyu Zhu, Haoyu Deng, Liang-Jian Deng, and Zhen Lei. 2024. FusionMamba: Efficient Remote Sensing Image Fusion with State Space Model. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–16. arXiv:2404.07932 [cs] doi:10.1109/TGRS.2024.3496073

[27] Xiangyuan Peng, Yu Wang, Miao Tang, Bierzynski Kay, Lorenzo Servadei, and Robert Wille. 2025. MoRAL: Motion-aware Multi-Frame 4D Radar and LiDAR Fusion for Robust 3D Object Detection. *arXiv preprint arXiv:2505.09422* (2025).

[28] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. 2010. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering* 58, 1 (2010), 7–11.

[29] Monay Mokhtar Shoushan, Bersain Alexander Reyes, Aldo Mejia Rodriguez, and Jo Woon Chong. 2021. Non-Contact HR Monitoring via Smartphone and Webcam During Different Respiratory Maneuvers and Body Movements. *IEEE Journal of Biomedical and Health Informatics* 25, 2 (Feb. 2021), 602–612. doi:10.1109/JBHI.2020.2998399

[30] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems* 27 (2014).

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

[32] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. 2014. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. 1056–1062. doi:10.1109/ROMAN.2014.6926392

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

[34] Jiankai Tang, Xinyi Li, Jiacheng Liu, Xiyuxing Zhang, Zeyu Wang, and Yuntao Wang. 2024. Camera-Based Remote Physiology Sensing for Hundreds of Subjects Across Skin Tones. In *CHI'24 Workshop PhysioCHI'24*.

[35] Jiankai Tang, Xin Liu, Daniel McDuff, Zhang Jiang, Hongming Hu, Luxi Zhou, Nodoka Nagao, Haruta Suzuki, Yuki Nagahama, Wei Li, Linhong Ji, Yuanchun Shi, Izumi Nishidate, and Yuntao Wang. 2025. Camera Measurement of Blood Oxygen Saturation. (2025). arXiv:2503.01699 [cs.CE] https://arxiv.org/abs/2503.01699

[36] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, Vol. 2019. 6558.

[37] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. 2008. Remote plethysmographic imaging using ambient light. *Optics express* 16, 26 (2008), 21434–21445.

[38] Jiyao Wang, Xiao Yang, Qingyong Hu, Jiankai Tang, Can Liu, Dengbo He, Yuntao Wang, Yingcong Chen, and Kaishun Wu. 2025. PhysDrive: A Multimodal Remote Physiological Measurement Dataset for In-vehicle Driver Monitoring. *arXiv preprint arXiv:2507.19172* (2025).

[39] Kegang Wang, Jiankai Tang, Yantao Wei, Mingxuan Liu, Xin Liu, and Yuntao Wang. 2024. A Plug-and-Play Temporal Normalization Module for Robust Remote Photoplethysmography. *arXiv preprint arXiv:2411.15283* (2024).

[40] Kegang Wang, Yantao Wei, Jiankai Tang, Yuntao Wang, Mingwen Tong, Jie Gao, Yujian Ma, and Zhongjin Zhao. 2024. Camera-Based HRV Prediction for Remote Learning Environments. In *The 21st IEEE International Conference on Ubiquitous Intelligence and Computing (UIC 2024)*.

[41] Lei Wang, Xingwei Wang, Dalin Zhang, Xiaolei Ma, Yong Zhang, Haipeng Dai, Chenren Xu, Zhijun Li, and Tao Gu. 2023. Knowing Your Heart Condition Anytime: User-Independent ECG Measurement Using Commercial Mobile Phones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 131 (Sept. 2023), 28 pages. doi:10.1145/3610871

[42] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. 2016. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering* 64, 7 (2016), 1479–1491.

[43] Xinyu Xie, Yawen Cui, Tao Tan, Xubin Zheng, and Zitong Yu. 2025. FusionMamba: Dynamic Feature Enhancement for Multimodal Image Fusion with Mamba. arXiv:2404.09498 [cs] doi:10.48550/arXiv.2404.09498

[44] Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for computational linguistics* 4 (2016), 259–272.

[45] Zitong Yu, Xiaobai Li, and Guoying Zhao. 2019. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. *arXiv preprint arXiv:1905.02419* (2019).

[46] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. 2022. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4186–4196.

[47] Zitong Yu, Benjia Zhou, Jun Wan, Pichao Wang, Haoyu Chen, Xin Liu, Stan Z Li, and Guoying Zhao. 2021. Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. *IEEE Transactions on Image Processing* 30 (2021), 5626–5640.

[48] Fusang Zhang, Zhi Wang, Beihong Jin, Jie Xiong, and Daqing Zhang. 2020. Your Smart Speaker Can "Hear" Your Heartbeat! *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 161 (Dec. 2020), 24 pages. doi:10.1145/3432237

[49] Hengrui Zhang, Feiyang Liao, Gang Yuan, Haoyang Jin, Biao Xie, Xu Cao, Mingcui Fu, and Jian Zheng. 2025. MaKAN-Mixer: Channel Interaction-Based Mamba Method for rPPG Extraction. *IEEE Journal of Biomedical and Health Informatics* (2025).

[50] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. arXiv:2401.09417 [cs] doi:10.48550/arXiv.2401.09417

[51] Bochao Zou, Zizheng Guo, Jiansheng Chen, Junbao Zhuo, Weiran Huang, and Huimin Ma. 2025. RhythmFormer: Extracting patterned rPPG signals based on periodic sparse attention. *Pattern Recognition* 164 (2025), 111511.

A APPENDIX

In this supplementary material, we report the performance of representative traditional unsupervised rPPG methods (GREEN, ICA, POS, and OMIT) on M^{3}PD for both face (front camera) and fingertip (rear camera with flash) inputs, together with scatter plots of estimated vs. reference HR. We further compare HRV descriptors between the Lab and Clinic subsets to quantify the increased rhythm irregularity in patients and motivate the necessity of dual-view acquisition and fusion for robust mobile HR estimation.

Table 10. Performance of traditional unsupervised methods on the M^{3}PD dataset using face inputs.

Method	**Lab**				**Clinic**			
	MAE ↓	**MAPE ↓**	**RMSE ↓**	ρ ↑	**MAE ↓**	**MAPE ↓**	**RMSE ↓**	ρ ↑
GREEN	28.924	32.966	37.173	-0.023	32.892	40.550	38.202	-0.038
ICA	19.438	21.902	26.897	0.060	18.531	22.627	24.079	**0.134**
POS	**16.549**	**19.851**	**23.755**	**0.115**	**16.475**	**21.190**	**22.210**	0.087
OMIT	25.672	29.145	34.364	-0.006	28.189	34.704	34.473	-0.003

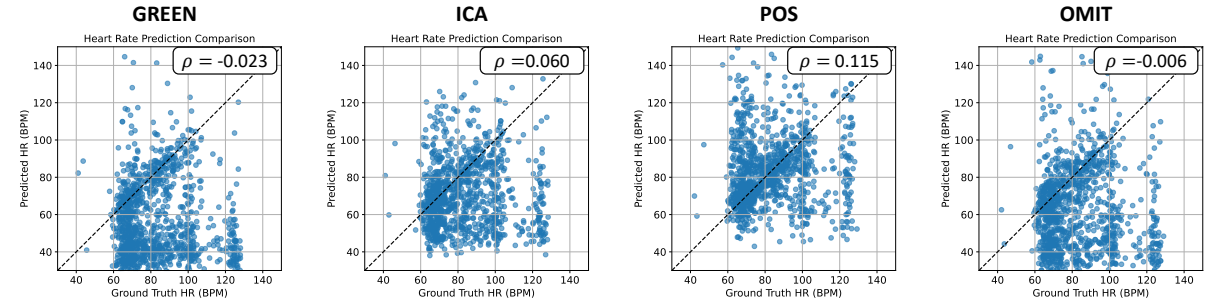
MAE = mean absolute error in HR estimation (beats/min), RMSE = root mean square error in HR estimation (beats/min), MAPE = mean percentage error (%), ρ = pearson correlation in HR estimation.

Table 11. Performance of traditional unsupervised methods on the M^{3}PD dataset using fingertip inputs.

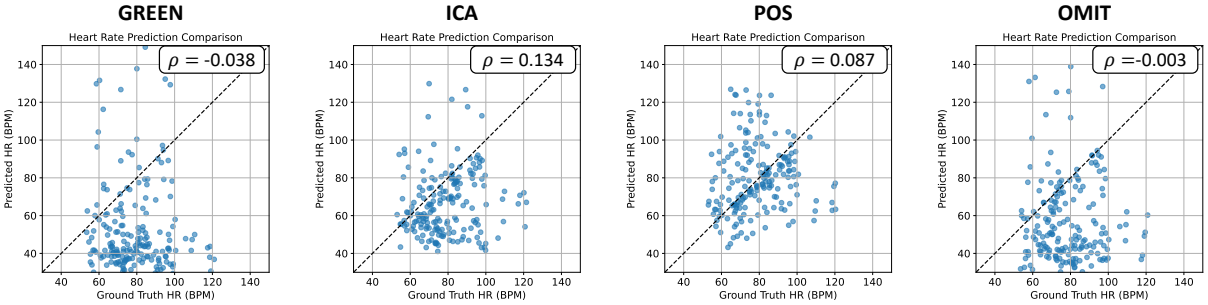
Method	**Lab**				**Clinic**			
	MAE ↓	**MAPE ↓**	**RMSE ↓**	ρ ↑	**MAE ↓**	**MAPE ↓**	**RMSE ↓**	ρ ↑
GREEN	19.735	22.213	31.380	0.081	17.468	21.775	25.045	0.421
ICA	**7.749**	**8.924**	**14.763**	**0.647**	10.881	12.720	18.126	0.389
POS	18.461	20.659	30.452	0.095	**9.423**	**11.565**	**14.845**	**0.520**
OMIT	11.011	12.835	18.053	0.448	18.511	23.168	26.219	0.419

MAE = mean absolute error in HR estimation (beats/min), RMSE = root mean square error in HR estimation (beats/min), MAPE = mean percentage error (%), ρ = pearson correlation in HR estimation.

To quantify clinical difficulty, we compared heart-rate–variability (HRV) descriptors between the two subsets (Table 12). Patients in the Clinic subset exhibit significantly higher SDNN, SDSD, and SD2, indicating more irregular and less autonomically regulated rhythms than healthy volunteers. In such cases, relying on a *single* peripheral site increases the risk of under-counting beats (pulse deficit). By capturing two vascular beds simultaneously—the facial region via the front camera and the fingertip via the rear camera with flash—M^{3}PD provides exactly the



(a) Lab Dataset (Face inputs).



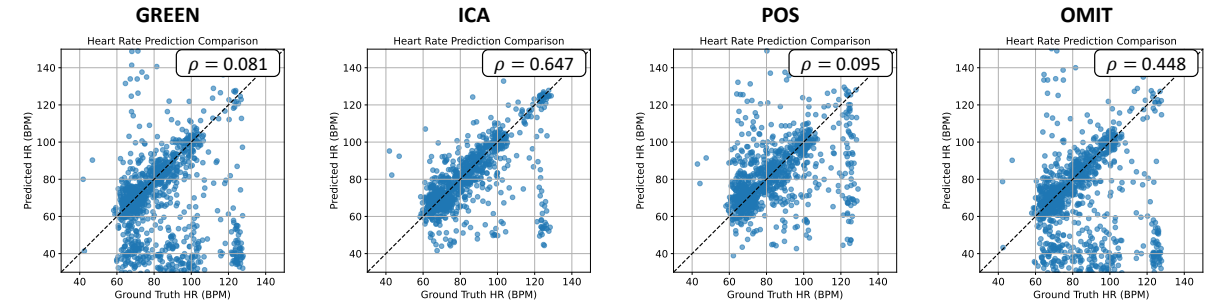
(b) Clinic Dataset (Face inputs).

Fig. 12. **Scatter plots of estimated HR vs. reference HR for face inputs.** Top: Lab; Bottom: Clinic. The diagonal indicates ideal correlation.

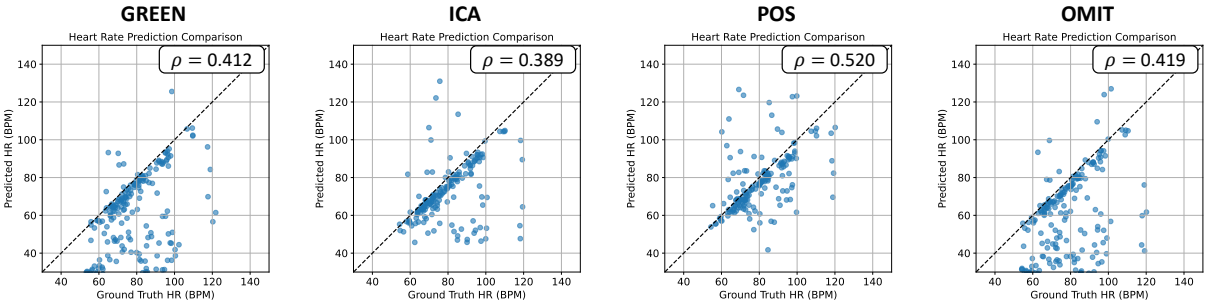
data needed to design fusion models that remain reliable when one view loses pulsatility. This is why we release M^{3}PD as a *dual-view* and *patient-inclusive* mobile dataset, rather than only a clean laboratory collection.

Table 12. Comparison of HRV metrics between healthy subjects and cardiovascular patients. Values are Mean ± SD.

Indicator	Lab	Clinic	p-value	Note
SDNN	34.823 ± 8.269	41.393 ± 7.360	0.013	Overall HRV: higher here mainly reflects rhythm irregularity in patients.
SDSD	23.348 ± 6.345	26.687 ± 3.133	0.015	Short-term/beat-to-beat variability: more unstable cycles in clinic group.
SD2	23.401 ± 5.677	32.286 ± 6.177	< 0.001	Long-term variability: suggests slower, irregular modulation.
PPA	2698.219 ± 996.072	4174.720 ± 1276.353	< 0.001	Poincaré area: more scattered RR pattern, harder for camera PPG.



(a) Lab Dataset (Fingertip inputs).



(b) Clinic Dataset (Fingertip inputs).

Fig. 13. **Scatter plots of estimated HR vs. reference HR for fingertip inputs.** Top: Lab; Bottom: Clinic. The diagonal indicates ideal correlation.