

# RPM-Distill: Physiology-guided Adaptive Cross-modal Distillation for Robust Remote Physiological Measurement

Jiyao Wang<sup>1</sup>, Qingyong Hu<sup>2</sup>, Duoxun Tang<sup>4</sup>, Xiao Yang<sup>3</sup>, Kaishun Wu<sup>3</sup>, and Jiangbo Yu<sup>1\*</sup>

<sup>1</sup> McGill University, Montreal, Canada

<sup>2</sup> Hong Kong University of Science and Technology, Hong Kong, China

<sup>3</sup> Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

<sup>4</sup> Tsinghua University, Shenzhen, China  
{jiangbo.yu}@mcgill.ca

**Abstract.** Video-based remote physiological measurement (RPM) is highly accessible but remains fragile under varying illumination, skin tones, and motion. Radio frequency (RF) radar is largely invariant to illumination and appearance, providing complementary cardio-respiratory micro-motion cues; however, requiring radar at inference is often impractical due to its limited ubiquity and deployment overhead. We propose **RPM-Distill**, a physiology-guided cross-modal distillation framework that leverages synchronized radar only during training while retaining video-only inference. Our key observation is that although RGB and RF waveforms differ in sensing physics and time-domain morphology, they share similar latent periodic rhythm in the frequency domain. We thus distill physiology-structured spectral evidence to improve robustness, via losses that (i) anchor the fundamental peak, (ii) match the off-peak background distribution, and (iii) preserve spectral morphology and sharpness. To avoid negative transfer under sample-level teacher quality and alignment uncertainty, a spectral policy network predicts sample-level distillation gates and component weights from the student–teacher spectral relation map, learned with a meta bilevel objective on a small labeled validation split. Through extensive experiments in challenging conditions and cross-dataset settings, RPM-Distill reduces MAE by up to 81% and improves correlation by up to 21% over the best video-only baseline. Code is at <https://github.com/WJULYW/RPM-Distill>.

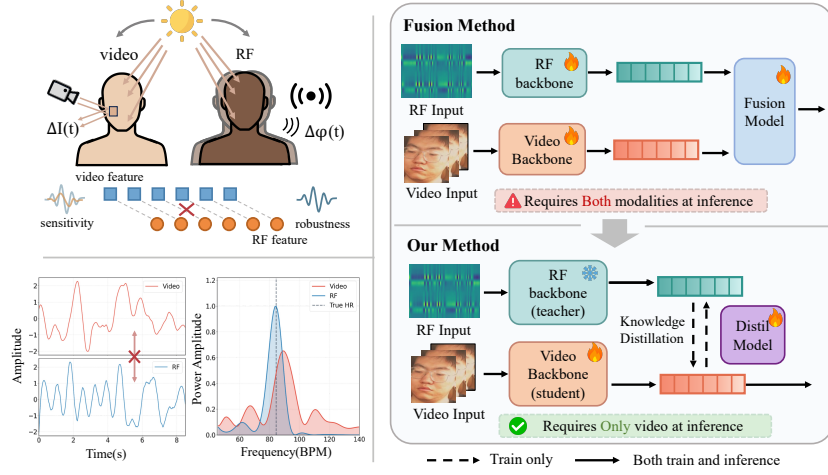
**Keywords:** Cross-modal distillation · Meta learning · Remote photoplethysmography · Radio frequency

## 1 Introduction

Remote physiological measurement (RPM) recovers vital signs, such as heart rate (HR) and blood volume pulse (BVP), without direct skin contact, enabling

---

\* Corresponding author



**Fig. 1:** Motivation for our study. Left: rPPG measures pulsatile reflectance changes, whereas RF radar measures cardio-respiratory micro-displacements; their sensing physics yield mismatched time-domain morphology, making naive feature/time alignment prone to negative transfer. However, within the physiological band, both modalities are modulated by the same periodic rhythm, motivating the distillation on frequency spectra. Right: Instead of fusion that requires both sensors at test time, we distill training-time radar knowledge into a video-only student.

continuous monitoring in extensive applications [55, 59]. Remote photoplethysmography (rPPG) is a representative RGB video-based RPM technique that extracts physiological signals from subtle, periodic appearance variations in facial or skin videos [38, 53]. Compared to contact-based acquisition, rPPG is non-invasive and highly accessible, yet it remains fundamentally fragile under unconstrained real-world conditions. Motion-induced non-rigid deformation and pose variations introduce strong temporal artifacts, while ambient light fluctuations and diverse skin tones significantly degrade the optical signal-to-noise ratio (SNR) [1, 29, 37]. Improving the robustness of video-based RPM against these compounded interferences remains a central focus in RPM.

Radio-frequency (RF) radar provides a complementary sensing mechanism by measuring cardio-respiratory micro-displacements via phase modulation of reflected electromagnetic waves [2, 27]. It is generally invariant to illumination and appearance and can be privacy-friendly [72]. These complementary properties have motivated video-radar fusion methods [6, 11, 31, 54, 66], yet most existing solutions rely on multimodal processing at inference and require sufficiently large paired datasets for training, both of which are difficult to satisfy in practical deployments where cameras are ubiquitous, but radar is not.

This tension motivates a practical question: *how can one exploit radar’s complementary cues without committing to multi-sensor inference?* Cross-modal knowledge distillation (KD) [20, 28] is a natural candidate, but it is particularly non-trivial for RPM due to physiology- and physics-induced modality mismatch.

As summarized in Fig. 1, first, RGB and RF observe the same underlying rhythm through fundamentally different transduction mechanisms, so their intermediate features and time-domain waveforms are not directly commensurate; feature-level alignment therefore suffers from modality/semantic mismatch [4]. Second, RPM distillation must be reliability-aware. Teacher quality and cross-modal alignment can vary across samples (e.g., transient corruption or synchronization shift), so indiscriminate output matching can propagate erroneous spectral evidence and cause negative transfer [5, 61]. Third, RPM targets are inherently periodic, and their failure modes are structured in the frequency domain. Under low SNR, video predictions often exhibit peak suppression/shift and spurious maxima, while nuisance factors elevate the off-peak noise floor and smear the local spectral shape [35, 37, 44, 65]. These considerations suggest that effective cross-modal distillation for RPM should leverage the shared latent rhythm, and be sample-selective to avoid transferring unreliable teacher signals.

In this work, we address these challenges with **RPM-Distill**, a physiology-guided dynamic cross-modal distillation framework in a training-time privileged-information setting [52]. We distill on band-limited frequency spectra where both modalities share the same latent periodicity, and impose a minimal yet complementary set of physiology-structured constraints that directly target dominant RPM failure modes: (i) fundamental peak anchoring for rhythm localization, (ii) off-peak/background matching for noise-floor suppression, and (iii) spectral morphology/sharpness consistency for stable rhythm evidence. To further handle sample-dependent teacher quality and alignment uncertainty, we introduce a spectral policy learner that predicts per-sample distillation gates and component weights from a three-channel student–teacher spectral relation map. Since optimal gating/weighting is unobserved and paired labeled data are limited, the policy is trained via a bilevel meta objective that is anchored by a small held-out validation split. Our contributions are summarized as:

- To the best of our knowledge, we present the first cross-modal distillation framework for RPM, enabling video-only inference while leveraging RF teacher as the training-time privileged information.
- We propose the physiology-guided spectral distillation objective with three components that target complementary spectral properties and mitigate modality-specific artifacts.
- We introduce an adaptive distillation policy network optimized via bilevel meta-learning. By predicting per-sample gates and weights from a three-channel spectral relation map, the policy improves both generalization and distillation selectivity under unreliability in teacher and synchronization.
- Extensive experiments on challenging conditions demonstrate consistent gains over fusion, unimodal, and generic KD baselines, with improved generalization under diverse skin tone, motion, and dynamic lights.

## 2 Related Works

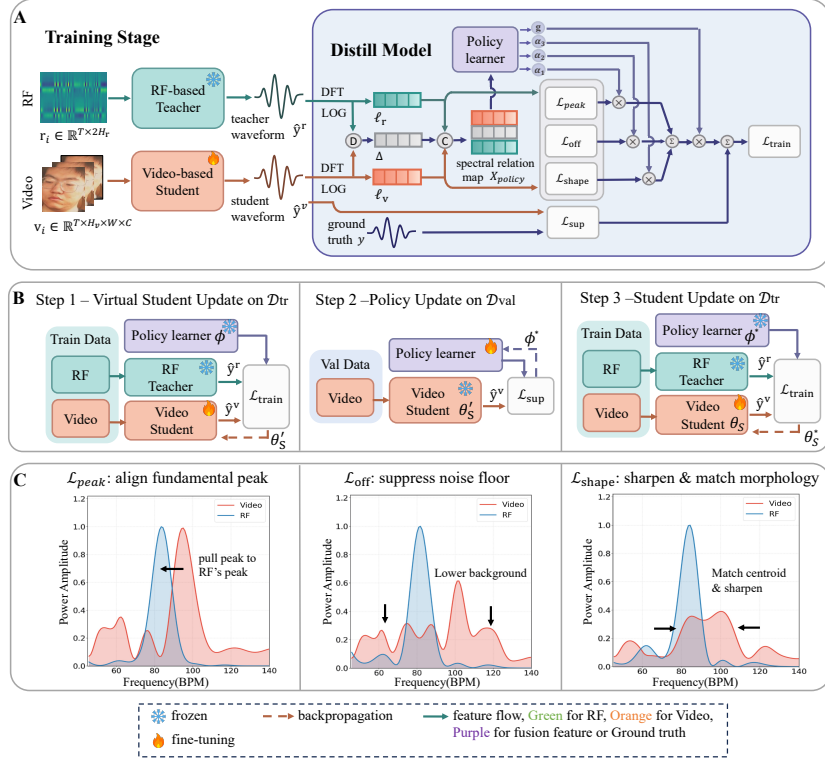
**Remote Physiological Measurement.** RPM has been primarily advanced by video-based methods that capture subtle facial color variations. Early approaches employed signal decomposition [38, 39, 51] and color transformations [63, 64], yet remained constrained by strong priors. Deep learning subsequently enabled robust spatiotemporal modeling via supervised [22, 33, 58, 62, 67, 74] and unsupervised frameworks [13, 45, 49]. Nevertheless, they remain susceptible to illumination changes and motion artifacts [30, 60].

RF-based methods circumvent optical limitations by sensing cardiopulmonary-induced chest vibrations. Initial efforts used handcrafted decomposition [3, 26, 34, 51]; recent deep learning adaptations improved accuracy [18, 19, 24, 73]. However, RF-based models still suffer from electromagnetic interference and occlusion-induced attenuation.

Recognizing their complementary nature, early works explored Video-RF fusion [6, 54], yet struggled with the modality gap and asynchronous sensor degradation. Recent advances propose adaptive strategies, including dynamically weights modalities by uncertainty and noise [11]; integration with cross-modal temporal dynamics [66]; and adapter-based alignment mitigates spatial-temporal mismatches [31]. Nevertheless, some feature-level fusion methods [31] account for alignment and impose dependence on particular backbone architectures of both modalities, which may impair the transferability. Besides, all existing fusion approaches still require multimodal input in inference, which hinders deployment.

**Privileged-Information Learning and Adaptive Distillation.** Learning Using Privileged Information (LUPI) formalizes the training-deployment mismatch where extra cues are available only during training, improving generalization via additional supervision signals [52]. Generalized distillation unifies LUPI and KD by treating the privileged modality as a teacher that provides soft targets to a student with limited inputs at deployment [32]. KD typically transfers "dark knowledge" through softened outputs [16] and has been extended to intermediate features and attention maps [41, 69, 71].

Cross-modal KD instantiates when teacher and student use different modalities. Early work showed that a stronger modality can supervise a weaker one via cross-modal distillation for supervision transfer [15, 17]. In video understanding, auxiliary streams (e.g., depth/optical flow) have been used as training-time privileged information, while keeping RGB-only testing [10]. Recent cross-modal KD methods go beyond static objectives to address modality gaps and sample-dependent transfer quality [14, 20, 28], particularly in VLM pretraining and efficient adaptation [9, 25]. Some works also attempt to distill knowledge between RF and video [40, 42] for human/object detection, while cross-modal KD toward RPM remains underexplored. The two modalities share the same latent rhythm but arise from different sensing physics and exhibit structured, sample-dependent spectral corruption in RPM. Naive global matching might propagate modality-specific artifacts, which motivates physiology-aware distillation on spectral components.



**Fig. 2:** The overview of RPM-Distill. (A) In training, a fixed radar teacher provides guidance to a video student via distillation with three components. A spectral policy network predicts a per-sample gate and component weights from the student–teacher spectral relation map, and is optimized by bilevel meta-learning. (B) The process of bilevel meta optimization. A virtual student update is evaluated on a labeled validation split to update the policy, after which the student is updated using the refined policy. (C) The illustration of our three distillation components in the frequency domain.

### 3 Methodology

#### 3.1 Preliminaries

We consider a multimodal dataset  $\mathcal{D}_{tr} = \{(\mathbf{v}_i, \mathbf{r}_i)\}_{i=1}^N$  during training.  $\mathbf{v}_i \in \mathbb{R}^{T \times H_v \times W \times C}$  is a facial image sequence, where  $T, C$  denote the numbers of frames and channels, and  $H_v$  and  $W$  are height and width, respectively. For RF data, the raw signal is preprocessed into a range–time representation and decomposed into in-phase and quadrature components, resulting in a feature tensor  $\mathbf{r}_i \in \mathbb{R}^{T \times 2H_r}$  (see Supplementary for details). Ground-truth physiological labels  $\mathbf{y}$  may be only available for a small-scale training data.

Our goal is to learn a video-based student network  $S(\cdot; \theta_S)$  that predicts a physiological waveform  $\hat{\mathbf{y}}^v = S(\mathbf{v}_i; \theta_S)$  using only RGB videos at inference. During training, we leverage a pre-trained radar teacher  $R(\cdot; \theta_R)$  to provide

auxiliary guidance  $\hat{\mathbf{y}}^r = R(\mathbf{r}_i; \theta_R)$ . Depending on label availability, we optimize the student with either distillation-only training or joint supervised-and-distilled training. This is summarized by the following label-optional objective:

$$\begin{aligned} \theta_S^* &= \arg \min_{\theta_S} \mathbb{E}_{(\mathbf{v}, \mathbf{r}) \sim \mathcal{D}_{tr}} \left[ \mathcal{L}_{\text{train}}(\hat{\mathbf{y}}^v, \hat{\mathbf{y}}^r, \mathbf{y}) \right], \\ \mathcal{L}_{\text{train}}(\hat{\mathbf{y}}^v, \hat{\mathbf{y}}^r, \mathbf{y}) &= \mathbb{I}_{\{\mathbf{y}\}} \mathcal{L}_{\text{sup}}(\hat{\mathbf{y}}^v, \mathbf{y}) + \lambda_{\text{distill}} \mathcal{L}_{\text{distill}}(\hat{\mathbf{y}}^v, \hat{\mathbf{y}}^r; \phi). \end{aligned} \quad (1)$$

$\mathbb{I}_{\{\mathbf{y}\}}$  is an indicator function that equals 1 if the ground-truth physiological label  $\mathbf{y}$  is available for the current sample, and 0 otherwise.  $\mathcal{L}_{\text{distill}}$  is governed by a meta-policy  $\phi$ . To prevent the policy from propagating modality-specific noise, a separate labeled validation set  $\mathcal{D}_{val}$  is reserved exclusively to meta-optimize  $\phi$ , anchoring the learned policy to supervised physiological recovery. The overview of our proposal is shown in Fig. 2.

### 3.2 Physiology-Structured Spectral Distillation

Physiological signals are fundamentally periodic, yet standard cross-modal distillation is hard to directly apply to RPM. RGB and RF follow different sensing physics, their time-domain waveforms and intermediate features are not commensurate, and indiscriminate alignment in feature or time domains can propagate modality-specific artifacts and amplify errors when the teacher is locally unreliable. We instead exploit the fact that, despite distinct morphology, both modalities encode the same latent cardiac rhythm in the frequency domain. This shared spectral structure provides a more stable interface for supervision and reduces sensitivity to modality-dependent distortions (e.g., radar multipath or optical shading). Moreover, the dominant RPM failure modes manifest in a small set of recurring spectral patterns, including peak mis-localization, off-peak background elevation, and local spectral smearing. We therefore design a new distillation objective by decomposing supervision into three complementary spectral components that explicitly target these modes.

Specifically, for a predicted waveform  $\hat{\mathbf{y}}$  of each modality  $m \in \{v, r\}$ , we compute its band-limited power spectrum  $\mathbf{P}^m \in \mathbb{R}^K$  and log-spectrum  $\ell^m \in \mathbb{R}^K$  that is *z-score normalized per sample* over frequency bins within the physiological band  $\mathcal{B}$ , where  $k \in \mathcal{B}$  indexes discrete frequency bins within the physiological band, and  $K = |\mathcal{B}|$  denotes the number of bins (details in Supplementary):

$$\mathbf{P}^m[k] = |\mathcal{F}(\hat{\mathbf{y}}^m)[k]|^2, \quad \ell^m[k] = \log(\mathbf{P}^m[k] + \epsilon), \quad k \in \mathcal{B}. \quad (2)$$

$\mathcal{F}$  denotes the Discrete Fourier Transform (DFT) and  $\epsilon$  is a small constant for numerical stability. This operation yields  $(\mathbf{P}^v, \ell^v)$  for the student and  $(\mathbf{P}^r, \ell^r)$  for the teacher. Their element-wise absolute discrepancy is  $\Delta = |\ell^v - \ell^r| \in \mathbb{R}^K$ . **Component Decomposition.** To construct differentiable masking mechanisms, we define the soft spectral centroid [48] over the normalized distribution  $\mathbf{s}^m = \text{softmax}(\ell^m) \in \mathbb{R}^K$  as  $\hat{k}^m \triangleq \sum_{k \in \mathcal{B}} k \cdot \mathbf{s}^m[k]$ . To localize supervision around the teacher-provided frequency anchor, we subsequently generate a Gaussian soft window mask centered at the teacher peak  $\hat{k}^r$ .

$$\mathbf{m}_{\text{pk}}[k] = \exp\left(-\frac{(k - \hat{k}^r)^2}{2\sigma^2}\right), \quad (3)$$

where the  $\sigma \approx 0.01K$  controls the window bandwidth. The remaining-band mask is defined as  $\mathbf{m}_{\text{oth}} = \max(\mathbf{0}, \mathbf{1} - \mathbf{m}_{\text{pk}})$ . Utilizing these spatial-frequency masks, we establish three physiological constraints:

1) *Fundamental Peak Alignment* ( $\mathcal{L}_{\text{peak}}$ ). For cardiac RPM, the dominant spectral mode within the physiological band encodes the fundamental rhythm and directly determines HR [45]. In practice, low illumination and skin-tone-dependent attenuation reduce rPPG SNR, causing the student spectrum to exhibit peak shifts or spurious dominant peaks away from the true rhythm [36, 37]. By contrast, when the radar teacher is reliable, its fundamental peak provides a stable frequency anchor even when the optical waveform is temporally distorted. Motivated by this, we enforce agreement specifically within a narrow neighborhood of the teacher peak, so that distillation corrects peak mis-localization without requiring time-domain waveform comparability:

$$\mathcal{L}_{\text{peak}} = \frac{\|\Delta \odot \mathbf{m}_{\text{pk}}\|_1}{\|\mathbf{m}_{\text{pk}}\|_1 + \epsilon}. \quad (4)$$

2) *Background Noise Suppression* ( $\mathcal{L}_{\text{off}}$ ). Another pervasive failure mode of video-based RPM is broadband spectral pollution. Motion, facial expressions, dynamic lighting, and imperfect ROI tracking inject non-physiological dynamics that spread energy across the band and elevate the off-peak noise floor [29, 64]. Such distributed interference is harmful even when the dominant peak is approximately correct, because it destabilizes peak selection and reduces the effective spectral contrast between the rhythm and background. Empirically, when the teacher is reliable, radar spectra tend to concentrate energy more strongly around the cardiac component while exhibiting a lower background level. We therefore match the student and teacher outside the peak window, encouraging the student to suppress non-peak energy:

$$\mathcal{L}_{\text{off}} = \frac{\|\Delta \odot \mathbf{m}_{\text{oth}}\|_1}{\|\mathbf{m}_{\text{oth}}\|_1 + \epsilon}. \quad (5)$$

3) *Spectral Morphology Consistency* ( $\mathcal{L}_{\text{shape}}$ ). Peak frequency alone is insufficient for robust RPM, because nuisance factors often preserve an approximate center frequency while altering the spectral morphology. In particular, motion and illumination transients can broaden the cardiac component in the student spectrum (peak smearing), leading to ambiguous maxima and reduced temporal consistency, whereas a reliable teacher typically exhibits a more concentrated spectral mass [64]. This motivates a morphology-level constraint that captures two complementary properties: (i) peak localization, which penalizes systematic shifts of the spectral centroid, and (ii) spectral sharpness [8], which penalizes overly diffuse distributions and reflects the confidence of the inferred rhythm. Using the normalized spectral distribution  $\mathbf{s}$ , we quantify sharpness via entropy and combine both terms with Smooth L1  $\mathcal{L}_{\text{L1}}$ :

$$H(\mathbf{s}) = -\frac{1}{\log K} \sum_{k \in \mathcal{B}} \mathbf{s}[k] \log(\mathbf{s}[k] + \epsilon),$$

$$\mathcal{L}_{\text{shape}} = \mathcal{L}_{l1}(\hat{k}^v, \hat{k}^r) + \mathcal{L}_{l1}(H(\mathbf{s}^v), H(\mathbf{s}^r)). \quad (6)$$

**Unified Distillation Objective.** The final objective is a policy-weighted summation. A global distillation gate  $g \in (0, 1)$  and component weights  $\boldsymbol{\alpha} \in \Delta^{|\mathcal{J}|-1}$  are dynamically predicted by the meta-policy  $\phi$ . We define  $\mathcal{J} = \{\text{peak}, \text{off}, \text{shape}\}$ :

$$\mathcal{L}_{\text{distill}}(\hat{\mathbf{y}}^v, \hat{\mathbf{y}}^r; \phi) = g \cdot \sum_{j \in \mathcal{J}} \alpha_j \mathcal{L}_j. \quad (7)$$

### 3.3 Data-Driven Spectral Policy Network

As the work [11] reveals, enforcing alignment indiscriminately is highly susceptible to negative transfer. Towards the optimal distillation strategy, we design a data-driven spectral policy network  $\Phi(\cdot; \phi)$  to evaluate inter-modal dependencies and adaptively modulate the distillation process from the spectral topology.

To provide the policy with a comprehensive physiological context, we construct a localized spectral relation map directly from the frequency domain. Specifically, for a given video-radar sample pair, we use the log-PSD vectors of both the student prediction  $\ell^v$  and the teacher prediction  $\ell^r$ . We stack them alongside their absolute discrepancy  $\Delta$  to form a three-channel spectral relation map  $\mathbf{X}_{\text{policy}} = [\ell^v, \ell^r, \Delta]^T \in \mathbb{R}^{3 \times K}$ . This tensor explicitly preserves the localized correlation structure across the frequency dimension  $K$ .

Subsequently,  $\mathbf{X}_{\text{policy}}$  passes through our spectral policy learner. The encoder consists of a 1D convolutional stem and successive residual blocks equipped with group normalization to extract multi-scale spectral representations. Then, inspired by [12, 23], we introduce a matrix-decomposition decoder, which learns basis and coefficient attentions over the frequency axis to extract low-rank quality tokens. These tokens are then fused with globally pooled features and bifurcated through linear projection heads to derive the optimal modulation factors:

$$g, \boldsymbol{\alpha} = \Phi(\mathbf{X}_{\text{policy}}; \phi). \quad (8)$$

Here,  $g$  is bounded by a Sigmoid activation, indicating the distillation’s instantaneous fidelity. Concurrently, the component weighting vector  $\boldsymbol{\alpha}$  is normalized via a Softmax operation to reside on the probability simplex  $\Delta^{|\mathcal{J}|-1}$ .

### 3.4 Bilevel Meta-Optimization

Optimizing the policy network  $\Phi(\cdot; \phi)$  introduces a challenge: we lack explicit ground truth labels for the optimal gate  $g$  and weights  $\boldsymbol{\alpha}$ . Manually tuning these hyperparameters is intractable given the dynamic, sample-dependent nature of unreliable distillation. To resolve this, we formulate the learning of  $\phi$  as a bilevel

optimization problem [21, 43]. The core insight is to treat the distillation policy as a learnable inductive bias.

Let  $\theta_S$  denote the student parameters and  $\phi$  the policy parameters. We aim to find the optimal  $\phi^*$  that minimizes the student’s supervised loss  $\mathcal{L}_{\text{sup}}$  on a held-out validation set  $\mathcal{D}_{\text{val}}$ , subject to the constraint that  $\theta_S$  is optimized on the training set  $\mathcal{D}_{\text{tr}}$  using  $\mathcal{L}_{\text{distill}}$ . In this work, the supervised loss  $\mathcal{L}_{\text{sup}}$  is instantiated as a combination of negative Pearson correlation loss [67] and SNR loss [46] to rigorously quantify the physiological waveform fidelity.

Then, since directly solving this nested optimization is computationally prohibitive, we adopt an online approximation strategy that interleaves the updates of the student and the policy. First, we simulate the student model’s change trajectory when updated under the current policy  $\phi$ . Using  $\mathcal{D}_{\text{tr}}$ , we perform a virtual gradient descent step to obtain temporary student parameters  $\theta'_S$ , where  $\eta$  is the learning rate:

$$\theta'_S \leftarrow \theta_S - \eta \nabla_{\theta_S} \mathcal{L}_{\text{train}}(\theta_S, \phi; \mathcal{D}_{\text{tr}}). \quad (9)$$

We then evaluate the supervised loss of the virtual student  $\theta'_S(\phi)$  on  $\mathcal{D}_{\text{val}}$  and update  $\phi$  to favor distillation decisions that improve supervised generalization. In Eq. (10),  $\beta$  is the meta-learning rate. This critical step explicitly aligns the distillation incentive with the student’s empirical performance.

$$\phi^* \leftarrow \phi - \beta \nabla_{\phi} \mathcal{L}_{\text{sup}}(\theta'_S(\phi); \mathcal{D}_{\text{val}}). \quad (10)$$

Finally, with the optimized policy  $\phi^*$ , we perform the actual update on the student network  $\theta_S$  using  $\mathcal{D}_{\text{tr}}$  to obtain the final  $\theta_S^*$  (Eq. (11)). By coupling the spectral constraints (Sec. 3.2) with this meta-optimization, RPM-Distill establishes a closed-loop system where physiological priors define the solution space, while the data-driven policy navigates the optimization trajectory.

$$\theta_S^* \leftarrow \theta_S - \eta \nabla_{\theta_S} \mathcal{L}_{\text{train}}(\theta_S, \phi^*; \mathcal{D}_{\text{tr}}). \quad (11)$$

## 4 Experiments

### 4.1 Datasets and Baselines

We conducted experiments on four datasets: **EquiPleth** [54] contains 91 participants’ RGB video and radar data with three skin tones and well-synchronized modalities. **PhysDrive** [61] was collected in a naturalistic driving scenario with 48 subjects and comprises RGB video and radar data under varying driver motions, dynamic natural lighting. Its synchronization was on the second level. **PURE** [47] consists of RGB video recordings from 10 subjects under different activities. **MMPD** [50] consists of RGB videos from 33 participants performing four distinct activities and exercise scenarios.

We choose six RGB video-based rPPG baselines, including two traditional methods (CHROM [7], POS [64]), and four deep-based methods (PhysNet [67],

PhysFormer [68], FactorizePhys [22], BeatFormer [33]). Besides, for RF-based methods, following [61], we adopt one traditional FFT-based method [3], and four deep learning-based baselines (IQ-MVED [73], VitaNet [24], mmFormer [18], and the RF predictor in [54]). For the RGB-RF multi-modal fusion baselines, we choose Vilesov et al. [54], Fusion-Vital [6], FusionPhys [66], SATM [31], and EvidentialFusion [11]. Lastly, we compare our proposal with four representative KD baselines, including feature-based FitNets [41], DML with mutual information [71], C<sup>2</sup>KD [20] and MST-Distill [28] for cross-modal KD. For fair comparison, we disable the multimodal specialized teachers in MST-Distill.

## 4.2 Implementation Details

The proposed method is implemented in PyTorch and trained on an NVIDIA A800 GPU. We set the sampling rate to 30 Hz and restrict the physiological band  $\mathcal{B}$  to [45, 180] bpm for all spectral operations and HR estimation. Each training sample is a clip of length  $T = 256$  with a stride of 30 frames. We use MTCNN [70] for face cropping and resize frames to  $72 \times 72$ . We train the student with Adam ( $\eta = 5 \times 10^{-5}$ , weight decay  $10^{-2}$ ) for 5 epochs, and keep a supervised-pretrained radar teacher frozen for producing  $\hat{\mathbf{y}}^r$ . Unless otherwise stated, we instantiate the student with FactorizePhys and the teacher with the RF predictor in [54]. The spectral policy network is optimized with Adam ( $\beta = 10^{-4}$ ), with bilevel updates every 10 student steps. Details are in the Supplementary.

All models aim to recover the waveform. We compute HR for each signal clip using band-pass filtering, detrending, and frequency-domain peak detection. To evaluate the predicted HR error, we report the average standard deviation of the error (STD), mean absolute error (MAE), root mean square error (RMSE), and Pearson’s correlation coefficient (r) of five runs with different random seeds. STD, MAE, and RMSE are reported in beats per minute (bpm).

## 4.3 Experimental Results

**Main results.** We first evaluate RPM-Distill under the cross-dataset protocol, where we train and validate on one dataset, then test and report on the other one. We split the train and validation set with 4:1 in subject level with random seeds. In this experiment, we optimize the student with both the supervised loss and the proposed distillation loss on the full training set. As shown in Table 1, video-only rPPG baselines suffer substantial performance degradation under domain shift. Radar-only methods also exhibit limited transferability, suggesting that RF spectra still may provide unreliable supervision across domains.

Previous KD baselines mitigate the modality gap by feature/output transfer from RF to RGB, yet their improvements remain bounded, particularly when trained on PhysDrive, where dynamic lighting, driving-induced motion, and worse synchronization amplify the optical failure modes. In contrast, RPM-Distill achieves the best overall error metrics on both datasets and maintains strong correlation, even compared to multimodal fusion approaches, which utilize the multimodal data in inference. Notably, it reduces the MAE/RMSE on

**Table 1:** Cross-dataset performance comparison on PhysDrive and EquiPleth dataset. In this and following tables, best results are in **bold** and second best are underlined. Statistically significant ( $p < .005$ ) improvements are marked ‘\*’ using paired  $t$ -tests.

| Modality          | Method                 | PhysDrive [61] |               |               |              | EquiPleth [54] |              |              |              |
|-------------------|------------------------|----------------|---------------|---------------|--------------|----------------|--------------|--------------|--------------|
|                   |                        | STD↓           | MAE↓          | RMSE↓         | $r \uparrow$ | STD↓           | MAE↓         | RMSE↓        | $r \uparrow$ |
| Video             | CHROM [7]              | 17.17          | 16.72         | 23.97         | 0.19         | 10.77          | 7.31         | 17.42        | 0.54         |
|                   | POS [64]               | 17.33          | 17.38         | 24.38         | 0.20         | 11.51          | 4.64         | 15.07        | 0.68         |
|                   | PhysNet [67]           | <u>12.73</u>   | 18.45         | 25.45         | 0.15         | 11.10          | 7.93         | 13.64        | 0.55         |
|                   | PhysFormer [68]        | 17.10          | 18.87         | 21.36         | 0.16         | 12.56          | 9.47         | 13.20        | 0.34         |
|                   | FactorizePhys [22]     | 14.74          | 15.67         | 20.55         | 0.19         | 11.83          | 8.21         | 14.40        | 0.78         |
|                   | BeatFormer [33]        | 16.10          | 17.16         | 21.80         | 0.17         | 13.82          | 9.08         | 13.98        | 0.62         |
| RF                | FFT-based [3]          | 18.21          | 20.60         | 24.29         | 0.10         | 8.77           | 12.82        | 15.53        | 0.32         |
|                   | IQ-MVED [73]           | 16.61          | 18.61         | 20.45         | 0.12         | 8.60           | 13.06        | 15.64        | 0.34         |
|                   | VitaNet [24]           | 18.31          | 16.31         | 21.07         | 0.16         | 8.36           | 11.88        | 14.52        | 0.45         |
|                   | mmFormer [18]          | 15.38          | 15.38         | 27.03         | 0.15         | 12.41          | 13.66        | 13.85        | 0.55         |
|                   | Vilesov et al. [54]    | 13.60          | 15.21         | <u>20.42</u>  | 0.21         | 7.79           | 8.46         | 9.44         | 0.78         |
| Fusion            | Vilesov et al. [54]    | 19.34          | 17.89         | 22.93         | 0.21         | 6.45           | 5.82         | 9.59         | 0.85         |
|                   | FusionPhys [66]        | 15.91          | 15.35         | 22.48         | <b>0.23*</b> | <u>5.24</u>    | 4.88         | 6.48         | 0.89         |
|                   | SATM [31]              | 15.22          | <b>13.79*</b> | 21.99         | <b>0.23*</b> | 5.75           | 3.92         | 6.62         | 0.88         |
|                   | EvidentialFusion [11]  | 15.21          | 16.12         | 22.42         | 0.19         | 7.06           | 7.35         | <u>6.04</u>  | 0.85         |
| Cross-modal<br>KD | FitNets [41]           | 15.26          | 20.20         | 28.31         | 0.09         | 6.60           | 3.09         | 7.23         | 0.85         |
|                   | DML [71]               | 15.25          | 17.96         | 23.58         | 0.13         | 5.44           | 2.57         | 7.98         | 0.87         |
|                   | C <sup>2</sup> KD [20] | 13.21          | 18.25         | 25.37         | 0.12         | 6.92           | <u>2.45</u>  | 6.88         | <u>0.90</u>  |
|                   | MST-Distill [28]       | 14.69          | 16.55         | 23.47         | 0.17         | 5.58           | 2.99         | 7.31         | 0.87         |
|                   | <b>Ours</b>            | <b>12.67*</b>  | <u>14.15</u>  | <b>20.24*</b> | <u>0.22</u>  | <b>4.14*</b>   | <b>1.57*</b> | <b>4.64*</b> | <b>0.94*</b> |

**Table 2:** Cross-dataset comparison between the best RGB-based baseline and our method under different conditions and limited labeled data. In this table, ‘E.M.&D.’ means ‘early morning and dusk’, and ‘C.&R.’ is ‘cloudy and rainy day’.

| Method             | E.M.&D.       |              | Noon          |              | Night         |              | R.&C.         |              | Stationary   |              | Talking       |              |
|--------------------|---------------|--------------|---------------|--------------|---------------|--------------|---------------|--------------|--------------|--------------|---------------|--------------|
|                    | MAE↓          | $r \uparrow$ | MAE↓          | $r \uparrow$ | MAE↓          | $r \uparrow$ | MAE↓          | $r \uparrow$ | MAE↓         | $r \uparrow$ | MAE↓          | $r \uparrow$ |
| FactorizePhys [22] | 15.32         | 0.19         | 13.28         | 0.23         | 20.51         | 0.01         | 13.78         | 0.18         | 13.55        | 0.20         | 18.31         | 0.05         |
| <b>Ours</b> 20%    | 12.42         | <u>0.28</u>  | 14.01         | 0.27         | 18.33         | <u>0.06</u>  | 14.38         | 0.20         | 15.84        | 0.15         | 17.27         | 0.16         |
| 40%                | <b>11.56*</b> | <b>0.31*</b> | <b>11.93*</b> | <b>0.31*</b> | <u>17.87</u>  | <b>0.07*</b> | <b>11.14*</b> | <b>0.29*</b> | <b>13.49</b> | <b>0.23*</b> | <b>14.36*</b> | <b>0.20*</b> |
| 60%                | 12.45         | 0.24         | 13.10         | 0.27         | 18.37         | 0.05         | 13.32         | 0.21         | 14.55        | 0.19         | 16.06         | 0.15         |
| 80%                | <u>12.06</u>  | <u>0.28</u>  | <u>12.70</u>  | 0.30         | <b>17.45*</b> | 0.06         | <u>12.73</u>  | <u>0.22</u>  | 14.11        | <u>0.22</u>  | <u>15.48</u>  | 0.18         |
| 100%               | 12.66         | 0.25         | 13.55         | 0.25         | 18.02         | 0.05         | 13.68         | 0.21         | 15.09        | 0.16         | 15.99         | 0.15         |

EquiPleth to 1.57/4.64, outperforming both fusion and KD baselines. These results support our key design: when training data suffer from lower inter-modal SNR or poorer cross-modal alignment, fusion is exposed to RF corruption and misalignment; while our selective distillation on shared rhythm-relevant regions can effectively avoid negative transfer.

**Performance on challenging conditions and limited labeled data.** To validate our improvements under limited labeled multimodal data, we conduct a label-scarce cross-dataset study by varying the percentage of labeled subjects (randomly selected with five random seeds) used for supervision. The best RGB-only baseline (FactorizePhys) is trained with full-label supervision on the entire EquiPleth training set. In contrast, RPM-Distill uses all paired video-radar samples but only a subset (20–100%) with labels for the joint objective  $\mathcal{L}_{\text{sup}} + \lambda_{\text{distill}} \mathcal{L}_{\text{distill}}$ ; the remaining samples are trained with  $\mathcal{L}_{\text{distill}}$  only. To avoid label leakage, the RF teacher is pretrained only on the labeled subset. Since we

**Table 3:** Ablation study on PhysDrive and EquiPleth with cross-dataset setting.

| Method                           | PhysDrive    |              |              | EquiPleth   |             |              |
|----------------------------------|--------------|--------------|--------------|-------------|-------------|--------------|
|                                  | M.↓          | R.↓          | $r \uparrow$ | M.↓         | R.↓         | $r \uparrow$ |
| w/o $\mathcal{L}_{\text{peak}}$  | 16.28        | 22.51        | 0.17         | 4.30        | 9.92        | 0.75         |
| w/o $\mathcal{L}_{\text{off}}$   | 16.10        | 22.37        | 0.16         | 3.40        | 8.29        | 0.82         |
| w/o $\mathcal{L}_{\text{shape}}$ | 15.95        | 22.29        | 0.16         | 3.13        | 7.93        | 0.83         |
| w/o weight                       | 15.83        | 22.06        | 0.17         | 3.25        | 8.19        | 0.83         |
| w/o gate                         | <u>15.47</u> | <u>21.66</u> | <u>0.18</u>  | 3.45        | 8.53        | 0.81         |
| w/o $\Phi(\cdot; \phi)$          | 15.68        | 21.89        | <u>0.18</u>  | <u>2.78</u> | <u>7.35</u> | 0.85         |
| w/o bilevel                      | 16.15        | 22.54        | 0.16         | 3.07        | 7.74        | 0.84         |
| Full                             | <b>14.15</b> | <b>20.24</b> | <b>0.22</b>  | <b>1.57</b> | <b>4.64</b> | <b>0.94</b>  |

**Table 4:** Evaluation on MMPD and PURE. Models are trained on EquiPleth.

| Method        | MMPD         |              |              | PURE        |              |              |
|---------------|--------------|--------------|--------------|-------------|--------------|--------------|
|               | M.↓          | R.↓          | $r \uparrow$ | M.↓         | R.↓          | $r \uparrow$ |
| CHROM         | 13.63        | 18.75        | 0.08         | 9.79        | <u>12.76</u> | 0.37         |
| POS           | 12.34        | 17.70        | <u>0.17</u>  | 9.82        | 13.44        | 0.34         |
| PhysNet       | 12.52        | 16.38        | 0.14         | 11.99       | 15.20        | 0.44         |
| PhysFormer    | 12.37        | 15.52        | 0.15         | 10.68       | 15.15        | 0.45         |
| FactorizePhys | 12.15        | <u>14.81</u> | <u>0.17</u>  | 9.54        | 14.26        | <u>0.51</u>  |
| BeatFormer    | 12.28        | 15.34        | <u>0.17</u>  | 10.39       | 14.12        | 0.48         |
| <b>Ours</b>   | <b>11.60</b> | <b>12.34</b> | <b>0.19</b>  | <b>8.28</b> | <b>10.33</b> | <b>0.55</b>  |

**Table 5:** Robustness to training-time cross-modal misalignment under the EquiPleth→PhysDrive protocol. The first column denotes the probability of applying such a shift to each training sample.

| Shift Prob. | Ours          |               |               |              | w/o gate     |       |       |              |
|-------------|---------------|---------------|---------------|--------------|--------------|-------|-------|--------------|
|             | STD↓          | MAE↓          | RMSE↓         | $r \uparrow$ | STD↓         | MAE↓  | RMSE↓ | $r \uparrow$ |
| 0           | <b>12.67*</b> | <b>14.15*</b> | <b>20.24*</b> | <b>0.22*</b> | 14.73        | 15.47 | 21.66 | 0.18         |
| 0.2         | 14.50         | <u>14.20</u>  | <u>20.30</u>  | <b>0.22*</b> | 14.65        | 15.68 | 21.74 | 0.18         |
| 0.4         | 14.69         | 14.75         | 20.81         | <u>0.20</u>  | 14.38        | 16.49 | 22.42 | 0.16         |
| 0.6         | 14.65         | 14.77         | 20.80         | <u>0.20</u>  | <u>14.30</u> | 17.04 | 23.50 | 0.15         |
| 0.8         | 14.57         | 14.32         | 20.43         | <u>0.20</u>  | 14.69        | 17.88 | 24.91 | 0.13         |
| 1.0         | 14.77         | 14.80         | 20.91         | 0.19         | 14.91        | 18.08 | 25.64 | 0.12         |

are unable to obtain other datasets that could provide RF and BVP signals, we skip the 0% variant. We evaluate performance in PhysDrive across different illumination and motion conditions.

As reported in Table 2, FactorizePhys degrades substantially on challenging subsets, particularly at Night and under Talking, reflecting the low-SNR and motion-induced failure modes of video-only RPM. In contrast, RPM-Distill remains consistently more robust with limited supervision. Interestingly, the best performance is achieved at 40% labeled data, while using more labels does not monotonically improve results. This is expected in our mixed objective setting: increasing the labeled fraction shifts optimization toward  $\mathcal{L}_{\text{sup}}$  on the source-domain labeled subset, which can amplify domain-specific bias and reduce the regularizing benefit of leveraging a larger portion of paired data through  $\mathcal{L}_{\text{distill}}$ . Thus, 20–40% provides a better balance between supervised anchoring and distillation-driven robustness, yielding the strongest gains across conditions.

**Ablation study.** In the variant w/o weight, we disable the policy learner  $\Phi(\cdot; \phi)$ ’s output weights and use fixed, equal weights for the three distillation components. In the variant w/o gate, we remove the policy learner’s output gate but keep its output weights. In the variant without  $\Phi(\cdot; \phi)$ , we remove the entire policy learner and the associated bilevel meta-optimization. The “w/o bilevel” variant keeps the same policy learner but optimizes it jointly with the student using  $\mathcal{L}_{\text{train}}$ . Table 3 shows that each term in our physiology-structured objective improves robustness, where M./R./ $r$  as MAE/RMSE/Pearson correlation.

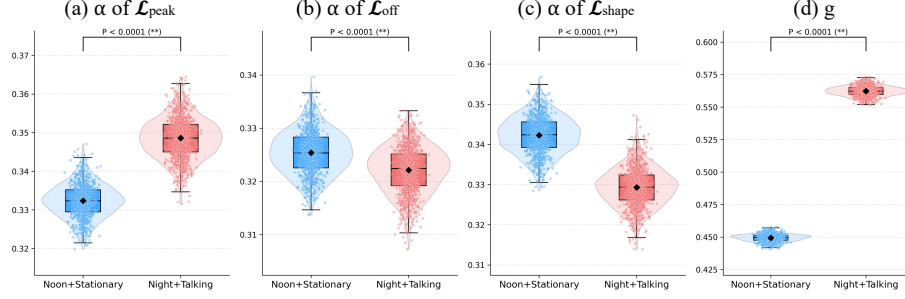
Dropping  $\mathcal{L}_{\text{peak}}$  causes the largest degradation, especially on EquiPleth, indicating that peak mislocalization under challenging illumination is a dominant failure mode and that anchoring the student to the teacher’s fundamental rhythm is crucial. Removing  $\mathcal{L}_{\text{off}}$  or  $\mathcal{L}_{\text{shape}}$  also consistently hurts performance, suggesting that suppressing the off-peak noise floor and constraining spectral morphology provide complementary benefits beyond peak alignment. Besides, using fixed, equal component weights or removing the gate degrades performance, implying that the optimal spectral focus is sample-dependent and indiscriminate supervision can cause negative transfer. Finally, removing  $\Phi(\cdot; \phi)$  leads to additional drops, showing that policy learning is key for strong generalization.

**Case study.** To assess whether the distillation improves the student beyond the multimodal benchmarks, we further perform a case study where models trained on a limited-size dataset (EquiPleth) are directly evaluated on RGB-only datasets. As shown in Table 4, our distilled student consistently outperforms all RGB baselines on both MMPD and PURE, achieving the lowest MAE/RMSE and the highest  $r$ . This suggests that the proposed method transfers a modality-agnostic rhythm prior that remains beneficial even when radar is absent at both training and deployment on the target domain. In particular, the gains on MMPD, which contains diverse activities and more severe motion, indicate that RPM-Distill can leverage RF-guided training to produce a stronger video-only model that generalizes to unseen RGB-only domains.

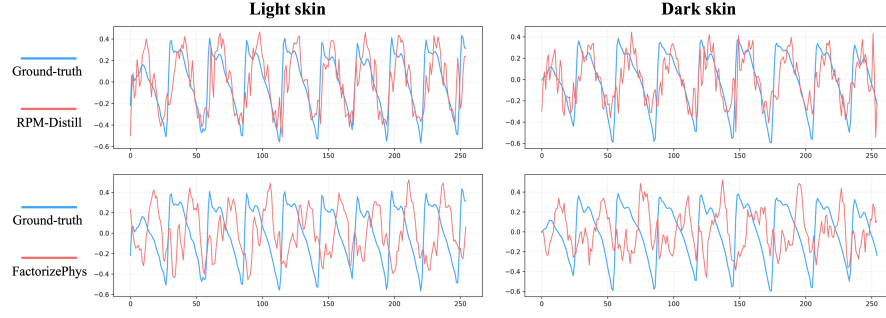
**Robust to unsynchronized multimodal training data.** We evaluate the robustness of the learned gate by explicitly corrupting cross-modal alignment in the well-synchronized EquiPleth training set. For a portion of training pairs, we shift the RF window forward by 30 steps (1s) while keeping the RGB clip unchanged, and vary the shift probability before testing on PhysDrive. As shown in Table 5, the ungated variant degrades steadily as the shift probability increases, indicating accumulated negative transfer from mismatched RF supervision. In contrast, the full model remains relatively stable and consistently outperforms, even when all training pairs are shifted. This suggests that the learned gate can down-weight unreliable student–teacher spectral relations and is effective beyond ideally synchronized training conditions.

#### 4.4 Visualization

We assess the effectiveness of our dynamic distillation policy by visualizing component weights and gate values when training on different scenarios. As shown in Fig. 3, the learned policy shows systematic distributional shifts between an easier subset (Noon+Stationary) and a harder subset (Night+Talking) on PhysDrive, with all changes statistically significant under paired  $t$ -tests. In the harder scenario, the policy assigns a larger gate  $g$ , meaning student training relies more on RF guidance when optical RPM becomes unreliable due to low illumination and stronger motion. The policy also increases the weight of the peak-related component while down-weighting the others. This matches our design motivation: under challenging conditions, video RPM primarily fails via peak suppres-



**Fig. 3:** A visualization of component weights  $\alpha$  and distillation gate  $g$  distributional changes when training in different PhysDrive scenarios. We measure the difference by the paired t-test.



**Fig. 4:** Visualization of pulse waveform examples generated by RPM-Distill and FactorizePhys in light and dark skin tone groups of EquiPleth.

sion/shift and spurious maxima, so emphasizing rhythm anchoring around the fundamental frequency is more effective than time-domain matching.

Besides, Fig. 4 provides additional qualitative evidence from PhysDrive to EquiPleth by comparing representative pulse waveforms from RPM-Distill and an RGB baseline without distillation, across different skin tones. In the light-skin group, both methods recover a periodic structure, but RPM-Distill yields cleaner, more temporally consistent cycles with fewer jagged fluctuations. In the dark-skin group, where optical SNR is lower, RPM-Distill maintains a clearer quasi-periodic pattern with more coherent peak timing, consistent with metric gains. It suggests that our method adapts the distillation focus under difficulty shifts, and that RF-guided spectral distillation improves robustness to skin-tone-induced domain gaps while preserving video-only inference.

## 5 Discussion

Our formulation focuses on band-limited spectral distillation and HR-centric evaluation; extending it to richer waveform-fidelity objectives and additional vi-

tal signs (e.g., RR or multi-task settings) is an important direction. RPM-Distill also assumes synchronized video–RF pairs during training and uses a small labeled validation split for bilevel policy learning, which may limit applicability under severe desynchronization, missing radar segments, or extremely scarce labels. Due to the limited availability of public RF datasets with consistent ground truth, we cannot yet fully validate unsupervised cross-dataset distillation, where an RF teacher pretrained on one dataset transfers to another without target labels. Moreover, our spectral interface presumes a stable periodic rhythm within a predefined physiological band; atypical rhythms, strong nonstationarity, or camera-dependent sampling differences may challenge this assumption. Finally, emerging multimodal foundation models may offer alternative routes for cross-modal alignment, and a careful comparison with physiology-structured distillation remains open.

Beyond accuracy, RPM-Distill offers a practical perspective on how to operationalize complementarity under modality mismatch. For RPM, RGB, and RF are not commensurate in the time domain and may exhibit modality-specific distortions; our results indicate that a more reliable transfer interface is the band-limited spectral rhythm, where heterogeneous sensors share a common periodic substrate. Decomposing distillation into three spectral components is sufficient to target dominant failure modes, while the learned policy highlights that both how much and what to distill are sample-dependent. This design also suggests a transferable route for heterogeneous-sensing tasks beyond RPM. When different modalities capture the same latent process through distinct physical mechanisms, RPM-Distill can be instantiated by defining the task-relevant shared structure, decomposing it into meaningful distillation components, and learning sample-adaptive reliability weights for cross-modal transfer.

## 6 Conclusion

In this paper, we presented **RPM-Distill**, a physiology-guided, adaptive cross-modal distillation framework that leverages RF radar in training to improve video-based RPM while preserving video-only inference. Unlike feature-level transfer or indiscriminate output matching, RPM-Distill distills shared, physiology-structured spectral evidence and uses a data-driven policy to gate and reweight supervision, reducing negative transfer under RF corruption or imperfect synchronization. The bilevel meta-objective further anchors policy learning to generalizable measurement [56, 57]. Extensive tests show consistent improvements.

## Acknowledgements

We gratefully acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant program, the Canada Research Coordination Committee’s New Frontiers in Research Fund (CRCC NFRF) Exploration, and the Canadian Foundation for Innovation (CFI) John R. Evans Leaders Fund (JELF).

## References

1. Acharya, B., Saakyan, W., Hammer, B., Drimalla, H.: The reliability of remote photoplethysmography under low illumination and elevated heart rates. *npj Digital Medicine* (2025)
2. Adib, F., Mao, H., Kabelac, Z., Katabi, D., Miller, R.C.: Smart homes that monitor breathing and heart rate. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*. pp. 837–846 (2015)
3. Alizadeh, M., Shaker, G., De Almeida, J.C.M., Morita, P.P., Safavi-Naeini, S.: Remote monitoring of human vital signs using mm-wave fmcw radar. *IEEE Access* **7**, 54958–54968 (2019)
4. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(2), 423–443 (2018)
5. Cho, J.H., Hariharan, B.: On the efficacy of knowledge distillation. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4794–4802 (2019)
6. Choi, J.H., Kang, K.B., Kim, K.T.: Fusion-vital: Video-rf fusion transformer for advanced remote physiological measurement. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 38, pp. 1344–1352 (2024)
7. De Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. *IEEE transactions on biomedical engineering* **60**(10), 2878–2886 (2013)
8. Elgendi, M., Martinelli, I., Menon, C.: Optimal signal quality index for remote photoplethysmogram sensing. *npj Biosensing* **1**(1), 5 (2024)
9. Feng, Q., Li, W., Lin, T., Chen, X.: Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4178–4188 (2025)
10. Garcia, N.C., Morerio, P., Murino, V.: Modality distillation with multiple stream networks for action recognition. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 103–118 (2018)
11. Ge, J., Sun, Z., Peng, W., Ying, C., Chen, Y., Ren, K., Li, X.: Evidential remote physiological measurement via uncertainty-aware fusion of video and rf. In: *Proceedings of the 33rd ACM International Conference on Multimedia*. pp. 2467–2475 (2025)
12. Geng, Z., Guo, M.H., Chen, H., Li, X., Wei, K., Lin, Z.: Is attention better than matrix decomposition? In: *International Conference on Learning Representations*
13. Gideon, J., Stent, S.: The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 3995–4004 (2021)
14. Guo, S., Zhao, H., Yang, S., Jiang, C., Cheng, Y., Peng, X., Shao, R., Tian, Z.: Multimodal dataset distillation via phased teacher models. In: *The Fourteenth International Conference on Learning Representations*
15. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2827–2836 (2016)
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015)
17. Hoffman, J., Gupta, S., Darrell, T.: Learning with side information through modality hallucination. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 826–834 (2016)

18. Hu, Q., Zhang, Q., Lu, H., Wu, S., Zhou, Y., Huang, Q., Chen, H., Chen, Y.C., Zhao, N.: Contactless arterial blood pressure waveform monitoring with mmwave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **8**(4), 1–29 (2024)
19. Hu, Q., Zhou, Y., Wang, J., Huang, Z., Li, G., Xu, Q., Zhang, Q.: mmtremor: Practical tremor monitoring for parkinson’s disease and essential tremor in daily life. In: *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*. pp. 498–512 (2025)
20. Huo, F., Xu, W., Guo, J., Wang, H., Guo, S.: C2kd: Bridging the modality gap for cross-modal knowledge distillation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16006–16015 (2024)
21. Ji, K., Yang, J., Liang, Y.: Bilevel optimization: Convergence analysis and enhanced design. In: *International conference on machine learning*. pp. 4882–4892. PMLR (2021)
22. Joshi, J., Agaian, S.S., Cho, Y.: Factorizephys: Matrix factorization for multidimensional attention in remote physiological sensing. *arXiv preprint arXiv:2411.01542* (2024)
23. Joshi, J., Cho, Y.: Ibvp dataset: Rgb-thermal rppg dataset with high resolution signal quality labels. *Electronics* **13**(7), 1334 (2024)
24. Khan, U.M., Rigazio, L., Shahzad, M.: Contactless monitoring of ppg using radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **6**(3), 1–30 (2022)
25. Kim, S., Xiao, R., Georgescu, M.I., Alaniz, S., Akata, Z.: Cosmos: Cross-modality self-distillation for vision language pre-training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14690–14700 (2025)
26. Li, C., Lin, J.: Random body movement cancellation in doppler radar vital sign detection. *IEEE Transactions on Microwave Theory and Techniques* **56**(12), 3143–3152 (2008)
27. Li, C., Ling, J., Li, J., Lin, J.: A review on recent advances in doppler radar sensors for noncontact healthcare monitoring. *IEEE Transactions on Microwave Theory and Techniques* **61**(5), 2046–2060 (2013)
28. Li, H., Yang, P., Chen, J., Dong, L., Chen, Y., Wang, Q.: Mst-distill: Mixture of specialized teachers for cross-modal knowledge distillation. In: *Proceedings of the 33rd ACM International Conference on Multimedia*. pp. 1588–1597 (2025)
29. Li, J., Yu, Z., Shi, J.: Learning motion-robust remote photoplethysmography through arbitrary resolution videos. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 37, pp. 1334–1342 (2023)
30. Li, X., Chen, J., Zhao, G., Pietikainen, M.: Remote heart rate measurement from face videos under realistic situations. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4264–4271 (2014)
31. Liang, Q., Geng, R., Chen, J., Wang, H., Chen, Y., Hu, Y.: Spatial alignment and temporal matching adapter for video-radar remote physiological measurement. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 8623–8633 (2025)
32. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643* (2015)
33. Martínez, J.C., Sukno, F.: Beatformer: Efficient motion-robust remote heart rate estimation through unsupervised spectral zoomed attention filters. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 520–530 (2025)

34. Mercuri, M., Liu, Y.H., Lorato, I., Torfs, T., Wieringa, F., Bourdoux, A., Van Hoof, C.: A direct phase-tracking doppler radar using wavelet independent component analysis for non-contact respiratory and heart rate monitoring. *IEEE transactions on biomedical circuits and systems* **12**(3), 632–643 (2018)
35. Niu, X., Shan, S., Han, H., Chen, X.: Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing* **29**, 2409–2423 (2019)
36. Nowara, E.M., McDuff, D., Veeraraghavan, A.: A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 284–285 (2020)
37. Nowara, E.M., McDuff, D., Veeraraghavan, A.: The benefit of distraction: Denoising camera-based physiological measurements using inverse attention. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4955–4964 (2021)
38. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering* **58**(1), 7–11 (2010)
39. Poh, M.Z., McDuff, D.J., Picard, R.W.: Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express* **18**(10), 10762–10774 (2010)
40. Pokhrel, S.R., Kua, J., Satish, D., Williams, P., Zaslavsky, A., Loke, S.W., Choi, J.: Deakin rf-sensing: Experiments on correlated knowledge distillation for monitoring human postures with radios. *IEEE Sensors Journal* **23**(22), 28399–28410 (2023)
41. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets (2015), <https://arxiv.org/abs/1412.6550>
42. Shan, L., Zhang, R., Chilukoti, S.V., Zhang, X., Lee, I., Hei, X.: Identitykd: Identity-wise cross-modal knowledge distillation for person recognition via mmwave radar sensors. In: *Proceedings of the 6th ACM International Conference on Multimedia in Asia*. pp. 1–7 (2024)
43. Sinha, A., Malo, P., Deb, K.: A review on bilevel optimization: From classical to evolutionary approaches and applications. *IEEE transactions on evolutionary computation* **22**(2), 276–295 (2017)
44. Song, R., Chen, H., Cheng, J., Li, C., Liu, Y., Chen, X.: PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics* **25**(5), 1373–1384 (2021)
45. Speth, J., Vance, N., Flynn, P., Czajka, A.: Non-contrastive unsupervised learning of physiological signals from video. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14464–14474 (2023)
46. Špetlík, R., Franc, V., Matas, J.: Visual heart rate estimation with convolutional neural network. In: *Proceedings of the british machine vision conference, Newcastle, UK*. pp. 3–6 (2018)
47. Stricker, R., Müller, S., Gross, H.M.: Non-contact video-based pulse rate measurement on a mobile service robot. In: *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. pp. 1056–1062. IEEE (2014)
48. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 529–545 (2018)
49. Sun, Z., Li, X.: Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In: *European Conference on Computer Vision*. pp. 492–510. Springer (2022)

50. Tang, J., Chen, K., Wang, Y., Shi, Y., Patel, S., McDuff, D., Liu, X.: Mmpd: Multi-domain mobile video physiology dataset. In: 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 1–5. IEEE (2023)
51. Tu, J., Hwang, T., Lin, J.: Respiration rate measurement under 1-d body motion using single continuous-wave doppler radar vital sign detection system. *IEEE Transactions on Microwave Theory and Techniques* **64**(6), 1937–1946 (2016)
52. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. *Neural Networks* **22**(5-6), 544–557 (2009)
53. Verkruyse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. *Optics express* **16**(26), 21434–21445 (2008)
54. Vilesov, A., Chari, P., Armouti, A., Harish, A.B., Kulkarni, K., Deoghare, A., Jalilian, L., Kadambi, A.: Blending camera and 77 ghz radar sensing for equitable, robust plethysmography. *ACM Transactions on Graphics (TOG)* **41**(4), 1–14 (2022)
55. Wang, J., Li, W., Wang, Z., Ayas, S., Donmez, B., He, D., Wu, K.: Drowsydg-phys: Generalizable driver drowsiness estimation in conditional automated vehicles using physiological signals. *Accident Analysis & Prevention* **228**, 108407 (2026)
56. Wang, J., Lu, H., Han, H., Chen, Y., He, D., Wu, K.: Generalizable remote physiological measurement via semantic-sheltered alignment and plausible style randomization. *IEEE Transactions on Instrumentation and Measurement* **74**, 1–14 (2024)
57. Wang, J., Lu, H., Wang, A., Chen, Y., He, D.: Hierarchical style-aware domain generalization for remote physiological measurement. *IEEE Journal of Biomedical and Health Informatics* **28**(3), 1635–1643 (2023)
58. Wang, J., Lu, H., Wang, A., Yang, X., Chen, Y., He, D., Wu, K.: Physmle: Generalizable and priors-inclusive multi-task remote physiological measurement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(6), 4908–4925 (2025)
59. Wang, J., Wang, Y., Jiao, Y., Yang, X., He, D., Jafarnejad, S., Miranda-Moreno, L., Frank, R., Yu, J.: Mild: Mediator agent system with bidirectional perception and multi-layered alignment for human-vehicle collaboration. *arXiv preprint arXiv:2605.01507* (2026)
60. Wang, J., Wei, X., Lu, H., Chen, Y., He, D.: Condiff-rppg: Robust remote physiological measurement to heterogeneous occlusions. *IEEE Journal of Biomedical and Health Informatics* **28**(12), 7090–7102 (2024)
61. Wang, J., Yang, X., Hu, Q., Tang, J., Liu, C., He, D., Wang, Y., Chen, Y., Wu, K.: Physdrive: A multimodal remote physiological measurement dataset for in-vehicle driver monitoring. *Advances in Neural Information Processing Systems* **38** (2026)
62. Wang, J., Yang, X., Lu, H., He, D., Wu, K.: Align the gap: Prior-based unified multi-task remote physiological measurement framework for domain generalization and personalization. *International Journal of Computer Vision* **134**(5), 199 (2026)
63. Wang, W., den Brinker, A.C., De Haan, G.: Discriminative signatures for remote-ppg. *IEEE Transactions on Biomedical Engineering* **67**(5), 1462–1473 (2019)
64. Wang, W., Den Brinker, A.C., Stuijk, S., De Haan, G.: Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering* **64**(7), 1479–1491 (2016)
65. Xi, L., Chen, W., Zhao, C., Wu, X., Wang, J.: Image enhancement for remote photoplethysmography in a low-light environment. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). pp. 1–7. IEEE (2020)

66. Ying, C., Yang, H., Ge, J., Sun, Z., Cheng, X., Ren, K., Li, X.: Fusionphys: A flexible framework for fusing complementary sensing modalities in remote physiological measurement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9363–9373 (2025)
67. Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In: 30th British Machine Vision Conference: BMVC 2019. 9th-12th September 2019, Cardiff, UK. The British Machine Vision Conference (BMVC) (2019)
68. Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P.H., Zhao, G.: Physformer: Facial video-based physiological measurement with temporal difference transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4186–4196 (2022)
69. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (2017)
70. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* **23**(10), 1499–1503 (2016)
71. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4320–4328 (2018)
72. Zhao, M., Tian, Y., Zhao, H., Alsheikh, M.A., Li, T., Hristov, R., Kabelac, Z., Katabi, D., Torralba, A.: Rf-based 3d skeletons. In: Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication. pp. 267–281 (2018)
73. Zheng, T., Chen, Z., Zhang, S., Cai, C., Luo, J.: More-fi: Motion-robust and fine-grained respiration monitoring via deep-learning uwb radar. In: Proceedings of the 19th ACM conference on embedded networked sensor systems. pp. 111–124 (2021)
74. Zou, B., Guo, Z., Hu, X., Ma, H.: Rhythmmamba: Fast, lightweight, and accurate remote physiological measurement. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 11077–11085 (2025)