BRIDGING UNPAIRED MODALITIES WITH UNIFIED PHYSIOLOGICAL CODE LEARNING FOR MULTIMODAL RPPG MEASUREMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal remote photoplethysmography (rPPG) improves physiological measurement by integrating complementary sensing modalities, but existing methods typically require synchronized multimodal recordings. This requirement limits the reuse of unpaired physiological data and complicates the addition of new sensing modalities. In this work, we present the first study of *unpaired multimodal rPPG learning*, which learns from modality-specific recordings without cross-modal correspondence during training. The main challenge lies in learning a shared physiological representation across heterogeneous modalities in the presence of mismatched physiological states and modality-specific sensing characteristics. To address this challenge, we propose **PhysioBridge**, which connects unpaired modalities through a unified physiological code space. Specifically, **Physiological Code Space Construction** learns complementary modality-private and shared codes, enabling physiological information to be represented in a modality-invariant manner while preserving modality-specific characteristics. **Codeword Distribution Alignment** then matches codeword usage across modalities under comparable physiological conditions, encouraging representations from different modalities to utilize a consistent shared code space. At inference, synchronized modalities are adaptively fused based on the relevance of their representations to the shared physiological code space. Experiments on public and self-collected datasets show that PhysioBridge achieves competitive performance without paired training data and generalizes across diverse sensing configurations.

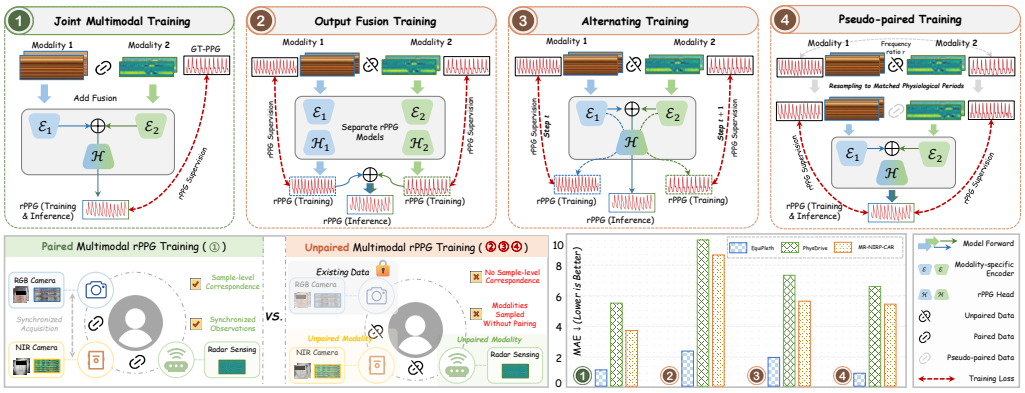
1 INTRODUCTION

Remote photoplethysmography (rPPG) enables non-contact measurement of physiological signals and heart rate (HR) from facial videos (Verkruysse et al., 2008; McDuff, 2023). Recent deep learning methods have substantially improved rPPG performance under diverse recording conditions (Yu et al., 2022; Zou et al., 2025). However, RGB-based rPPG remains sensitive to illumination, motion, skin appearance, and camera noise, motivating the use of complementary modalities such as NIR, radar, and wireless signals (Vilesov et al., 2022; Wang et al., 2025). To improve robustness, recent multimodal rPPG methods have explored integrating complementary sensing modalities (Ying et al., 2025; Liang et al., 2025), but they typically require synchronized recordings with sample-level cross-modal correspondence during training.

Although synchronized acquisition provides natural correspondence for multimodal fusion, it increases the data collection burden: adding a new modality often requires recollecting it together with existing modalities, limiting the reuse of unpaired data. We therefore study *unpaired multimodal rPPG learning*, which enables joint learning from independently sampled modalities without requiring sample-level cross-modal correspondence during training.

However, without cross-modal correspondence, heterogeneous modalities may encode the same physiological information differently, making it challenging to learn a shared representation for multimodal fusion. As illustrated in Fig. 1, we first examine two intuitive solutions. Training separate unimodal models and combining their outputs at inference (Setting 2) avoids paired training but provides no cross-modal representation interaction. Alternating unpaired modalities within a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Paired vs. unpaired multimodal rPPG training data. (b) Performance comparison of different training settings.

Figure 1: Comparison of multimodal rPPG training settings, including their acquisition requirements and HR estimation performance on three public datasets.

shared model (Setting 3) enables parameter sharing, yet mismatched physiological states still hinder effective cross-modal representation learning.

Notably, previous rPPG studies have shown that temporal resampling can controllably modify physiological rate together with its HR label (Gideon & Stent, 2021; Yue et al., 2025). This motivates Setting 4, where recordings from different modalities are resampled to a common HR and jointly trained as pseudo-paired data. While this strategy performs well on EquiPleth (Vilesov et al., 2022), its gains are limited on PhysDrive (Wang et al., 2025) and MR-NIRP-Car (Nowara et al., 2022) under more challenging recording conditions. This gap suggests that HR matching alone cannot resolve differences in subject characteristics, acquisition environments, and sensing mechanisms. Effective unpaired multimodal learning may therefore benefit from distinguishing modality-specific characteristics from physiological information shared across modalities.

To address this problem, we propose **PhysioBridge**, a unified framework for bridging unpaired physiological modalities for effective multimodal rPPG measurement. To the best of our knowledge, PhysioBridge is the first work to study multimodal rPPG learning from unpaired modality data, while achieving competitive multimodal performance without paired training data.

Specifically, the key idea of PhysioBridge is to connect heterogeneous unpaired modalities through a unified discrete physiological code space. We first introduce Physiological Code Space Construction, which learns modality-private and shared codes to capture modality-specific and transferable physiological information. To promote this separation, we use a codeword orthogonalization loss to reduce redundancy between private and shared codes, and cross-modal mutual reconstruction to promote the sharing of physiological information across modalities. Together, the physiological information is encouraged to be represented in a modality-invariant manner while preserving modality-specific characteristics in the private codes. However, the modalities may still use the shared codebook inconsistently during subsequent rPPG learning. We therefore introduce Codeword Distribution Alignment, which matches codeword usage distributions under comparable physiological conditions to promote consistent representations across modalities. Finally, we assess how confidently each modality maps to the shared code space and use the resulting code-confidence scores to adaptively fuse synchronized observations, integrating complementary physiological information.

The main contributions are summarized as follows: 1) We investigate *unpaired multimodal rPPG learning*, removing the need for sample-level paired training data and enabling the reuse of unpaired physiological recordings. 2) We propose Physiological Code Space Construction, which encourages complementary modality-specific characteristics and transferable physiological representations in a unified discrete space, enabling heterogeneous unpaired modalities to share a common physiological representation. 3) We introduce Codeword Distribution Alignment to further align the shared code space by matching codeword usage across modalities under comparable physiological periods, strengthening representation consistency and supporting code-confidence fusion. 4) We collect **MWSHR**, a synchronized multimodal dataset with RGB, NIR, and WiFi recordings under diverse lighting and motion conditions. Extensive experiments show that PhysioBridge achieves competitive performance across diverse sensing configurations and supports cross-dataset transfer without requiring paired multimodal training data.

2 Related Work

2.1 Multimodal rPPG Measurement

Remote photoplethysmography (rPPG) extracts physiological signals from subtle heartbeat-induced variations in facial videos for contact-free physiological measurement. Early methods rely on handcrafted color transformations and signal-processing priors (De Haan & Jeanne, 2013; Wang et al., 2017), while recent deep models learn spatiotemporal physiological representations directly from facial videos (Yu et al., 2022; Chu et al., 2025). Despite substantial progress, RGB-based rPPG remains sensitive to illumination, motion, and other environmental disturbances. This limitation has motivated recent studies to incorporate complementary modalities such as NIR and radar, exploring video–RF fusion, uncertainty-aware integration, cross-modal alignment, and flexible multimodal fusion (Nowara et al., 2022; Vilesov et al., 2022; Choi et al., 2024; Ge et al., 2025; Liang et al., 2025; Ying et al., 2025). However, existing multimodal rPPG methods generally require synchronized recordings during training, motivating multimodal rPPG learning without such correspondence.

2.2 Unpaired Multimodal Learning

Unpaired multimodal learning considers modality-specific samples without sample-level correspondence during training. Existing studies address this setting through knowledge transfer, distribution matching, shared representation learning, and cross-modal transfer (Dou et al., 2020; Xi et al., 2024; Yacobi et al., 2025; Tran et al., 2026). UML (Gupta et al., 2026) transfers complementary information by alternating unpaired modalities within shared model parameters. Related work in heterogeneous collaborative perception further explores non-co-occurring modalities: HEAL (Lu et al., 2024) aligns newly introduced heterogeneous agents to an established feature space, while CodeAlign (Liu et al., 2026) performs codebook-based cross-modal feature translation. These studies demonstrate the feasibility of learning across unpaired modalities, yet how to establish a shared physiological representation for direct multimodal rPPG fusion remains unexplored.

3 Methodology

3.1 Problem Formulation

We consider a set of sensing modalities $\mathcal{M}$ without requiring sample-level correspondence between recordings from different modalities. For modality $m \in \mathcal{M}$, its training set is $\mathcal{T}^m = \{(\mathbf{U}_i^m, \mathbf{y}_i^m, h_i^m)\}_{i=1}^{N_m}$, where $\mathbf{U}_i^m$ denotes a raw recording, $\mathbf{y}_i^m \in \mathbb{R}^T$ is its physiological waveform, and h_i^m is its HR. No sample-level cross-modal correspondence is used during training, whereas synchronized multimodal recordings are available at inference. We describe the bimodal case $\mathcal{M} = \{a, b\}$ below, while the multimodal extension is detailed in the appendix.

Multimodal Physiological Representation. We transform heterogeneous recordings into modality-specific spatio-temporal representations. For RGB and NIR videos, following (Niu et al., 2020; Wang et al., 2025), the face is divided into G regions to extract regional color or intensity traces. The mmWave representation retains in-phase and quadrature responses from R range bins, while the WiFi representation uses the real and imaginary CSI responses from A receiving antennas and S subcarriers. The resulting inputs are

$$\mathbf{X}^{\mathrm{RGB}} \in \mathbb{R}^{3G\times T}, \quad \mathbf{X}^{\mathrm{NIR}} \in \mathbb{R}^{G\times T}, \quad \mathbf{X}^{\mathrm{mmWave}} \in \mathbb{R}^{2R\times \rho T}, \quad \mathbf{X}^{\mathrm{WiFi}} \in \mathbb{R}^{2AS\times T}, \tag{1}$$

where ρ is the temporal sampling-rate ratio between mmWave and video. Each modality-specific encoder $\mathcal{E}^m$ maps its input into a common latent dimensionality, while its decoder $\mathcal{D}^m$ reconstructs the modality-specific input.

Framework Overview. As illustrated in Fig. 2, PhysioBridge consists of two training stages followed by code-confidence fusion. Stage-I constructs modality-private codebooks and a unified shared codebook through codeword queries and cross-modal mutual reconstruction. Stage-II freezes the shared codebook and aligns its conditional usage distributions across independently sampled modalities. At inference, synchronized inputs are mapped into the aligned shared space and fused according to their code confidence.

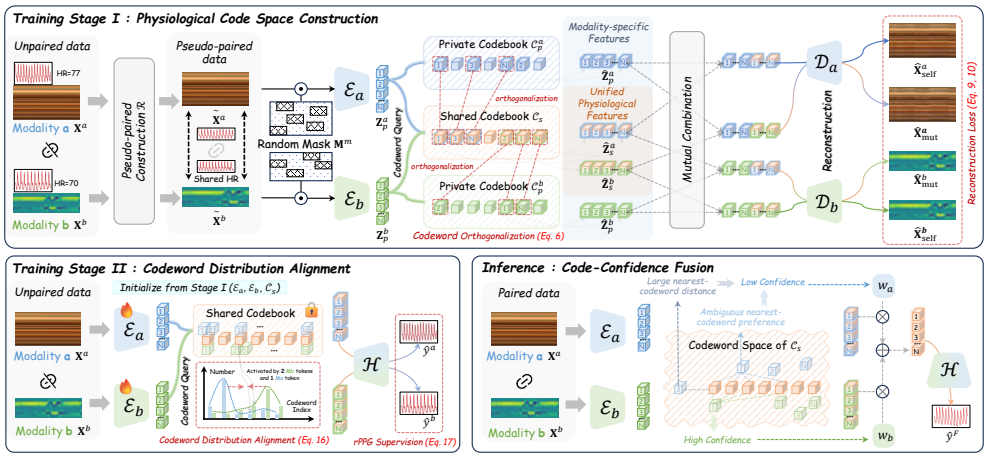


Figure 2: Overview of PhysioBridge for unpaired multimodal rPPG learning, including physiological code space construction, codeword distribution alignment, and code-confidence fusion.

3.2 Stage-I: Physiological Code Space Construction

Stage-I aims to construct a discrete physiological space that can be consistently interpreted across heterogeneous modalities. Since recordings sampled across modalities may differ in physiological states and sensing characteristics, direct cross-modal representation learning is unreliable. We therefore first establish physiological comparability and then learn complementary modality-private and shared representations in a common discrete space.

Pseudo-Pair Construction. Independently sampled recordings may have different HRs, making direct representation exchange unreliable. Since temporal resampling can adjust physiological frequencies (Gideon & Stent, 2021; Yue et al., 2025), we propose to resample independently selected cross-modal recordings to comparable physiological periods. To limit temporal deformation, we first sample one recording from each modality within an HR interval $\mathcal{B}_q$ and choose a target HR $h^* \in \mathcal{B}_q$. Each input is resampled as

$$\widetilde{\mathbf{X}}^m = \mathcal{R}\left(\mathbf{X}^m; \frac{h^*}{h^m}\right), \qquad m \in \{a, b\}, \tag{2}$$

where $\mathcal{R}(\cdot; r)$ denotes temporal resampling with ratio r. These pseudo-pairs have comparable physiological periods but retain no sample-level correspondence.

Private-Shared Code Learning. Although pseudo-pair construction reduces physiological mismatch, substantial discrepancies remain due to different sensing and acquisition conditions. Previous codebook-based rPPG methods have shown that discrete representations can provide a compact basis for physiological dynamics (Chu et al., 2025). We therefore encode each modality into a common latent representation and encourage complementary modality-specific characteristics and transferable physiological representations through private and shared codeword queries. To reduce local reconstruction shortcuts, we apply block-wise masking before encoding:

$$\overline{\mathbf{X}}^m = \widetilde{\mathbf{X}}^m \odot (1 - \mathbf{M}^m), \qquad \mathbf{Z}^m = \mathcal{E}^m(\overline{\mathbf{X}}^m), \tag{3}$$

where $\mathbf{M}^m$ is a modality-specific binary mask and $\mathbf{Z}^m \in \mathbb{R}^{D \times L}$ contains L encoded tokens of dimension D. Each modality has a private codebook $\mathcal{C}_p^m = \{\mathbf{c}_{p,k}^m\}_{k=1}^K$, whereas all modalities use the shared codebook $\mathcal{C}_s = \{\mathbf{c}_{s,k}\}_{k=1}^K$. The nearest-neighbor query of a latent token $\mathbf{z}$ is

$$\mathcal{Q}(\mathbf{z}, \mathcal{C}) = \mathbf{c}_{k^*}, \qquad k^* = \arg\min_k \|\mathbf{z} - \mathbf{c}_k\|_2^2. \tag{4}$$

Applying $\mathcal{Q}$ tokenwise to the same encoded representation yields

$$\widehat{\mathbf{Z}}_p^m = \mathcal{Q}(\mathbf{Z}^m, \mathcal{C}_p^m), \qquad \widehat{\mathbf{Z}}_s^m = \mathcal{Q}(\mathbf{Z}^m, \mathcal{C}_s). \tag{5}$$

The private query captures modality-specific characteristics, while the shared query maps each modality onto a common physiological code space. To reduce redundancy between these representations, we impose an orthogonality constraint on their co-activated codewords:

$$\mathcal{L}_{\text{orth}} = \frac{1}{|\mathcal{M}|L} \sum_{m\in\mathcal{M}} \sum_{l=1}^{L} \left(\frac{(\widehat{\mathbf{z}}^m_{p,l})^\top \widehat{\mathbf{z}}^m_{s,l}}{\|\widehat{\mathbf{z}}^m_{p,l}\|_2 \|\widehat{\mathbf{z}}^m_{s,l}\|_2 + \epsilon} \right)^2, \tag{6}$$

where ϵ is a small constant for numerical stability. This constraint encourages the activated private and shared codewords to capture complementary information.

Cross-Modal Reconstruction. Shared codeword querying alone does not require the resulting codes to carry cross-modally interpretable physiological information. We therefore train each modality's decoder to reconstruct its input using either its own shared codes or those queried by another modality. Self-reconstruction combines private and shared codes from the same modality:

$$\widehat{\mathbf{X}}^m_{\text{self}} = \mathcal{D}^m\left([\widehat{\mathbf{Z}}^m_p; \widehat{\mathbf{Z}}^m_s]\right). \tag{7}$$

Mutual reconstruction retains the target modality's private codes and exchanges only the shared codes:

$$\widehat{\mathbf{X}}^a_{\text{mut}} = \mathcal{D}^a\left([\widehat{\mathbf{Z}}^a_p; \widehat{\mathbf{Z}}^b_s]\right), \qquad \widehat{\mathbf{X}}^b_{\text{mut}} = \mathcal{D}^b\left([\widehat{\mathbf{Z}}^b_p; \widehat{\mathbf{Z}}^a_s]\right). \tag{8}$$

All reconstructions target the complete resampled input $\widetilde{\mathbf{X}}^m$ before masking. Mutual reconstruction encourages the exchanged shared codes to preserve physiological information that can be interpreted across modalities.

We supervise reconstruction using point-wise fidelity and temporal correlation:

$$\ell_{\text{rec}}(\mathbf{A}, \mathbf{B}) = \text{MSE}(\mathbf{A}, \mathbf{B}) + \mathcal{L}_{\text{NP}}(\mathbf{A}, \mathbf{B}), \tag{9}$$

where $\mathcal{L}_{\text{NP}}$ is the negative Pearson loss. The self- and mutual-reconstruction losses are

$$\mathcal{L}_{\text{self}} = \frac{1}{2} \sum_{m\in\{a,b\}} \ell_{\text{rec}}(\widehat{\mathbf{X}}^m_{\text{self}}, \widetilde{\mathbf{X}}^m), \qquad \mathcal{L}_{\text{mut}} = \frac{1}{2} \sum_{m\in\{a,b\}} \ell_{\text{rec}}(\widehat{\mathbf{X}}^m_{\text{mut}}, \widetilde{\mathbf{X}}^m). \tag{10}$$

Following VQ-VAE (van den Oord et al., 2017), we apply codebook and commitment losses to both codeword queries:

$$\mathcal{L}_{\text{VQ}} = \frac{1}{2|\mathcal{M}|} \sum_{m\in\mathcal{M}} \sum_{q\in\{p,s\}} \left(\|\text{sg}[\mathbf{Z}^m] - \widehat{\mathbf{Z}}^m_q\|_2^2 + \beta \|\mathbf{Z}^m - \text{sg}[\widehat{\mathbf{Z}}^m_q]\|_2^2 \right), \tag{11}$$

where $\text{sg}[\cdot]$ denotes stop-gradient and β is the commitment coefficient.

The overall Stage-I objective is

$$\mathcal{L}_{\text{I}} = \mathcal{L}_{\text{self}} + \mathcal{L}_{\text{mut}} + \lambda_{\text{VQ}}\mathcal{L}_{\text{VQ}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}}. \tag{12}$$

3.3 Stage-II: Codeword Distribution Alignment

Stage-I establishes a shared discrete physiological space, but modality-specific adaptation during downstream rPPG learning may still lead to inconsistent usage of this space. We initialize each encoder from Stage-I and keep the unified physiological codebook $\mathcal{C}_s$ frozen. For the i-th independently sampled recording of modality m, we have

$$\mathbf{Z}^m_i = \mathcal{E}^m(\mathbf{X}^m_i), \qquad \widehat{\mathbf{Z}}^m_{s,i} = \mathcal{Q}(\mathbf{Z}^m_i, \mathcal{C}_s), \qquad \widehat{\mathbf{y}}^m_i = \mathcal{H}(\widehat{\mathbf{Z}}^m_{s,i}), \quad m \in \{a, b\}, \tag{13}$$

where i indexes samples independently within each modality.

Despite querying the same codebook, different encoders may activate its codewords differently under comparable physiological conditions. We therefore align codeword usage distributions within local HR intervals. Let $\mathcal{S}^m_q$ contain the samples of modality m whose HR falls in interval $\mathcal{B}_q$. We define a shared soft-assignment function for an encoded token $\mathbf{z}$:

$$\mathcal{A}_k(\mathbf{z}) = \frac{\exp\left(-\|\mathbf{z} - \mathbf{c}_{s,k}\|_2^2 / \tau\right)}{\sum_{j=1}^{K} \exp(-\|\mathbf{z} - \mathbf{c}_{s,j}\|_2^2 / \tau)}, \qquad p^m_{i,l,k} = \mathcal{A}_k(\mathbf{z}^m_{i,l}), \tag{14}$$

where τ controls the assignment sharpness and $p^m_{i,l,k}$ is the assignment of the l-th token in sample i of modality m to the k-th shared codeword. Averaging these assignments across samples and tokens within each HR interval gives the conditional codeword usage distribution:

$$\pi^m_{q,k} = \frac{1}{|\mathcal{S}^m_q| L} \sum_{i \in \mathcal{S}^m_q} \sum_{l=1}^{L} p^m_{i,l,k}. \tag{15}$$

Thus, $\boldsymbol{\pi}^m_q = [\pi^m_{q,1}, \ldots, \pi^m_{q,K}]$ characterizes codeword usage under the physiological condition $\mathcal{B}_q$. We align these distributions between modalities within the same interval:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{I}||\mathcal{P}|} \sum_{q \in \mathcal{I}} \sum_{(m,n) \in \mathcal{P}} \mathcal{D}_{\text{KL}}\left(\boldsymbol{\pi}^m_q \,\|\, \boldsymbol{\pi}^n_q\right), \tag{16}$$

where $\mathcal{I}$ contains HR intervals with samples from both modalities, $\mathcal{P} = \{(a,b),(b,a)\}$ contains both ordered modality pairs, and D_{KL} denotes KL divergence. This conditional alignment encourages consistent codeword usage under comparable physiological periods. Since $\mathcal{C}_s$ is frozen, the differentiable soft assignments guide the modality-specific encoders within the established shared space.

Each modality also receives independent rPPG supervision. Omitting the sample index for clarity, we define

$$\mathcal{L}_{\text{rPPG}} = \frac{1}{2} \sum_{m \in \{a,b\}} \left[\mathcal{L}_{\text{NP}}(\widehat{\mathbf{y}}^m, \mathbf{y}^m) + \mathcal{L}_{\text{SNR}}(\widehat{\mathbf{y}}^m, h^m)\right]. \tag{17}$$

The negative Pearson loss (Yu et al., 2019) constrains waveform similarity, while the SNR loss (Vilesov et al., 2022) promotes spectral concentration. The overall Stage-II objective is

$$\mathcal{L}_{\text{II}} = \mathcal{L}_{\text{rPPG}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}. \tag{18}$$

3.4 Inference: Code-Confidence Fusion

Stage-II encourages heterogeneous modalities to use the shared physiological code space consistently. At inference, synchronized modalities are independently encoded and query the same frozen shared codebook. For the l-th token $\mathbf{z}^m_l$, its soft assignment to the k-th shared codeword is obtained by $p^m_{l,k} = \mathcal{A}_k(\mathbf{z}^m_l)$. As illustrated in Fig. 2, low code confidence can arise from ambiguous codeword preferences or a large distance to the nearest codeword. We assess these complementary aspects using assignment entropy and quantization error:

$$H^m_l = -\sum_{k=1}^{K} p^m_{l,k} \log p^m_{l,k}, \qquad e^m_l = \frac{1}{D} \left\| \mathbf{z}^m_l - \widehat{\mathbf{z}}^m_{s,l} \right\|_2^2. \tag{19}$$

Lower entropy indicates a clearer codeword preference, while lower quantization error indicates a closer match to the shared code space. We combine the two measures into a token-level confidence score and normalize the scores to obtain fusion weights:

$$r^m_l = \frac{1 - H^m_l / \log K}{1 + e^m_l}, \qquad w^a_l = \frac{r^a_l + \epsilon}{r^a_l + r^b_l + 2\epsilon}, \qquad w^b_l = \frac{r^b_l + \epsilon}{r^a_l + r^b_l + 2\epsilon}, \quad m \in \{a,b\}, \tag{20}$$

where ϵ is a small constant for numerical stability. The modality with higher code confidence receives a larger weight for the corresponding token. We then fuse the queried shared features directly:

$$\widehat{\mathbf{z}}^F_l = w^a_l \widehat{\mathbf{z}}^a_{s,l} + w^b_l \widehat{\mathbf{z}}^b_{s,l}. \tag{21}$$

The fused tokens form $\widehat{\mathbf{Z}}^F$, which is passed to the shared rPPG prediction head $\mathcal{H}$ to produce the final signal $\widehat{\mathbf{y}}^F$.

4 Experiments

4.1 Datasets and Evaluation Protocols

Datasets. **EquiPleth** (Vilesov et al., 2022) contains 550 synchronized RGB and mmWave recordings from 91 subjects under relatively controlled conditions. We adopt its official subject-independent splits. **MR-NIRP-Car** (Nowara et al., 2022) provides synchronized RGB and NIR

Table 1: Intra-dataset HR estimation on public multimodal rPPG datasets. NIR/Radar denotes NIR for MR-NIRP-Car and PhysDrive, and mmWave for EquiPleth. Gray, blue, and purple method cells indicate unimodal, paired multimodal, and unpaired multimodal methods, respectively. M and R denote MAE and RMSE. Best and second-best results are shown in **bold** and underlined.

Test Modality	Train Modality	Method	EquiPleth			MR-NIRP-Car			PhysDrive		
			M↓	R↓	r ↑	M↓	R↓	r ↑	M↓	R↓	r ↑
RGB	RGB	CHROM (De Haan & Jeanne, 2013)	7.45	13.4	0.46	8.94	15.6	0.31	12.2	16.0	0.11
		POS (Wang et al., 2017)	2.55	8.93	0.77	7.74	14.3	0.39	12.4	16.2	0.10
		RhythmFormer (Zou et al., 2025)	1.21	4.67	0.93	3.14	4.23	0.82	7.21	9.84	0.45
	RGB+ NIR/Radar (Unpaired)	UMKD (Dou et al., 2020)	3.56	5.12	0.86	5.30	6.58	0.75	11.6	13.4	0.17
		UML (Gupta et al., 2026)	1.79	2.87	0.97	3.73	5.11	0.82	6.45	7.87	0.56
		CodeAlign (Liu et al., 2026)	1.99	3.86	0.89	3.54	5.60	0.82	6.72	7.86	0.61
		PhysioBridge (Ours)	**1.00**	**2.85**	**0.98**	**2.42**	**3.53**	**0.88**	**5.21**	**6.50**	**0.69**
NIR/Radar	NIR/Radar	Contrast-Phys+ (Sun & Li, 2024)	–	–	–	2.68	4.77	0.85	13.7	16.1	0.05
		RhythmFormer (Zou et al., 2025)	–	–	–	**2.37**	3.94	0.90	8.15	9.22	0.16
		Radar-APLANC (Wang et al., 2026)	3.95	9.72	0.64	–	–	–	–	–	–
		FusionPhys (RF) (Ying et al., 2025)	2.18	6.12	0.89	–	–	–	–	–	–
	RGB+ NIR/Radar (Unpaired)	UMKD (Dou et al., 2020)	2.96	6.14	0.88	3.66	5.29	0.78	10.7	12.4	0.17
		UML (Gupta et al., 2026)	2.96	5.28	0.90	3.70	5.67	0.91	7.69	8.74	0.14
		CodeAlign (Liu et al., 2026)	3.25	5.96	0.81	3.91	5.12	0.80	6.81	8.23	**0.22**
		PhysioBridge (Ours)	**1.84**	**3.30**	**0.96**	2.73	**3.44**	**0.94**	**6.37**	**7.98**	0.18
RGB+ NIR/Radar (Paired)	RGB+ NIR/Radar (Paired)	EquiPleth (Vilesov et al., 2022)	1.12	3.42	0.95	3.28	4.65	0.86	6.93	9.26	0.33
		Fusion-Vital (Choi et al., 2024)	3.75	9.30	0.71	2.88	3.96	0.90	7.48	8.50	0.30
		Evidential-Phys (Ge et al., 2025)	**0.79**	2.37	0.98	2.02	3.01	0.92	4.67	**5.47**	0.70
		FusionPhys (Ying et al., 2025)	1.07	3.53	0.96	2.33	3.90	0.93	5.04	5.86	0.69
	RGB+ NIR/Radar (Unpaired)	UMKD (Dou et al., 2020)	3.08	5.11	0.87	4.48	6.06	0.67	7.48	8.50	0.32
		UML (Gupta et al., 2026)	1.15	2.80	0.98	3.32	5.20	0.86	5.85	7.08	0.57
		CodeAlign (Liu et al., 2026)	1.32	3.70	0.95	3.28	5.06	0.82	5.92	7.24	0.55
		PhysioBridge (Ours)	0.81	**2.35**	**0.99**	**1.89**	**2.51**	**0.95**	**4.39**	5.50	**0.74**

recordings under real driving conditions with substantial illumination and motion variations. Following Shao et al. (2025), we adopt a 3:1 training–test split. **PhysDrive** (Wang et al., 2025) contains RGB, NIR, and mmWave recordings from 48 drivers, with an 8:1:1 subject-level training, validation, and test split. Since the synchronized RGB–mmWave data of PhysDrive have not yet been released, we use RGB and NIR for intra-dataset evaluation. **MWSHR** contains synchronized RGB, NIR, and WiFi recordings from 21 subjects. Each subject performs six approximately 60-second tasks under three lighting conditions (dim, natural, and bright) and two motion conditions (stationary and conversation). We use 17 subjects for training and four for testing without subject overlap.

Evaluation Protocols. Paired methods use synchronized multimodal recordings with sample-level correspondence during training, whereas unpaired methods independently sample each modality without accessing cross-modal pairings. At inference, both settings use synchronized multimodal recordings for fusion under the same testing conditions. We report mean absolute error (MAE), root mean squared error (RMSE), and Pearson's correlation coefficient (r) for HR estimation.

Implementation Details. All modalities are processed over a common temporal window corresponding to $T = 256$ video frames. The input shapes are 192×256 for RGB, 64×256 for NIR, 10×1024 for mmWave, and 180×256 for WiFi. For input construction, $G = 64$ (8×8 facial grid), $R = 5$, $\rho = 4$, $A = 3$, and $S = 30$. The modality-specific encoders produce latent representations with a common dimensionality of $D = 256$ and a token count of $L = T/4 = 64$ across all modalities. Each codebook contains $K = 128$ codewords. We use a masking ratio of 0.2, block sizes of 16×32 for video and 8×32 for RF modalities, and 20 bpm HR intervals over 40–180 bpm. Stage-I and Stage-II are trained for 50 and 20 epochs with batch sizes of 4 and 16, respectively, using Adam with a learning rate of 10^{-3}. We set $\beta = 0.25$, $\lambda_{\text{VQ}} = 1.0$, $\lambda_{\text{orth}} = 0.01$, $\lambda_{\text{align}} = 1.0$, and $\tau = 1.0$. The RGB and NIR encoders $\mathcal{E}^m$ adopt the PhysNet encoder (Yu et al., 2019), whereas the mmWave and WiFi encoders follow the EquiPleth encoder (Vilesov et al., 2022). Each modality-specific decoder $\mathcal{D}^m$ mirrors its corresponding encoder. The prediction head $\mathcal{H}$ consists of a point-wise 3D convolutional layer followed by a max-pooling layer.

4.2 Intra-Dataset Evaluation

We compare PhysioBridge with representative unimodal, paired multimodal, and unpaired multimodal methods across different sensing configurations to evaluate its effectiveness.

Table 2: Intra-dataset HR estimation on the MWSHR dataset.

Test Modality	Train Modality	Method	MWSHR		
			MAE↓	RMSE↓	r ↑
RGB	RGB+NIR+WiFi (Unpaired)	UMKD (Dou et al., 2020)	3.05	4.60	0.62
		UML (Gupta et al., 2026)	2.92	4.03	0.82
		CodeAlign (Liu et al., 2026)	2.98	4.11	0.81
		PhysioBridge (Ours)	**2.39**	**3.77**	**0.89**
NIR	RGB+NIR+WiFi (Unpaired)	UMKD (Dou et al., 2020)	5.75	7.38	0.44
		UML (Gupta et al., 2026)	6.54	8.68	0.33
		CodeAlign (Liu et al., 2026)	6.32	7.68	0.37
		PhysioBridge (Ours)	**5.70**	**6.96**	**0.45**
WiFi	RGB+NIR+WiFi (Unpaired)	UMKD (Dou et al., 2020)	18.6	20.9	0.04
		UML (Gupta et al., 2026)	5.69	7.22	**0.54**
		CodeAlign (Liu et al., 2026)	8.47	10.1	0.41
		PhysioBridge (Ours)	**5.20**	**7.20**	0.42
RGB+NIR+WiFi (Paired)	RGB+NIR+WiFi (Paired)	EquiPleth (Vilesov et al., 2022)	3.55	5.69	0.84
		Fusion-Vital (Choi et al., 2024)	2.89	5.28	0.84
		Evidential-Phys (Ge et al., 2025)	3.19	5.52	0.80
		FusionPhys (Ying et al., 2025)	3.75	6.28	0.81
	RGB+NIR+WiFi (Unpaired)	UMKD (Dou et al., 2020)	7.25	8.99	0.39
		UML (Gupta et al., 2026)	3.56	5.17	0.80
		CodeAlign (Liu et al., 2026)	3.32	4.14	0.69
		PhysioBridge (Ours)	**1.93**	**2.40**	**0.90**

Table 3: Cross-dataset HR estimation results.

Method	PhysDrive →MR-NIRP-Car			MR-NIRP-Car →PhysDrive			PhysDrive →EquiPleth		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
EquiPleth (Vilesov et al., 2022)	5.45	6.02	0.76	14.9	17.5	0.07	–	–	–
Fusion-Vital (Choi et al., 2024)	3.94	5.53	0.83	15.2	18.6	0.05	–	–	–
Evidential-Phys (Ge et al., 2025)	2.74	4.37	0.85	10.2	13.5	0.17	–	–	–
FusionPhys (Ying et al., 2025)	3.10	4.91	0.82	11.9	15.5	0.06	–	–	–
UMKD (Dou et al., 2020)	4.43	5.16	0.78	12.1	14.6	0.04	5.66	9.31	0.69
UML (Gupta et al., 2026)	4.98	5.69	0.80	10.8	13.4	0.15	5.30	8.91	0.68
CodeAlign (Liu et al., 2026)	2.75	**4.11**	0.82	9.81	12.6	0.18	5.02	8.95	0.78
PhysioBridge (Ours)	**2.52**	**4.11**	**0.86**	**9.18**	**12.0**	**0.25**	**3.70**	**7.42**	**0.80**

Bimodal Intra-Dataset Evaluation. As shown in Tab. 1, PhysioBridge achieves competitive single-modality performance after unpaired multimodal training, although the improvements are not uniform across sensing modalities. The benefits are more consistent under multimodal inference, where PhysioBridge achieves MAEs of 0.81, 1.89, and 4.39 on EquiPleth, MR-NIRP-Car, and PhysDrive, respectively, outperforming all unpaired methods while remaining competitive with paired methods. Notably, the multimodal results consistently improve upon the corresponding single-modality predictions, indicating that the learned representations retain complementary physiological information across modalities. These results suggest that the shared physiological code space facilitates effective cross-modal integration without requiring paired training data.

Trimodal Intra-Dataset Evaluation. To examine scalability beyond bimodal sensing, we extend PhysioBridge to RGB, NIR, and WiFi on MWSHR, with implementation details provided in the appendix. As shown in Tab. 2, PhysioBridge achieves MAEs of 2.39, 5.70, and 5.20 under RGB-only, NIR-only, and WiFi-only inference, respectively. Compared with UML, PhysioBridge achieves lower MAE for all three modalities, indicating consistent benefits across modalities. Trimodal fusion further reduces MAE from 2.39 to 1.93 compared with RGB-only inference. This result also outperforms all paired multimodal methods, demonstrating the effectiveness of integrating unpaired heterogeneous modalities.

4.3 Cross-Dataset Evaluation

As shown in Tab. 3, we evaluate cross-dataset generalization under different multimodal settings. The two settings involving PhysDrive and MR-NIRP-Car use RGB and NIR, whereas PhysDrive→EquiPleth additionally involves a change in sensing modality. Paired methods are excluded from the latter setting because the synchronized RGB–mmWave data of PhysDrive have not yet been released. PhysioBridge achieves MAEs of 2.52, 9.18, and 3.70 across the three settings, respectively. The improvement is particularly evident on PhysDrive→EquiPleth, where MAE decreases from 5.02 for CodeAlign to 3.70. These results suggest that the learned physiological code

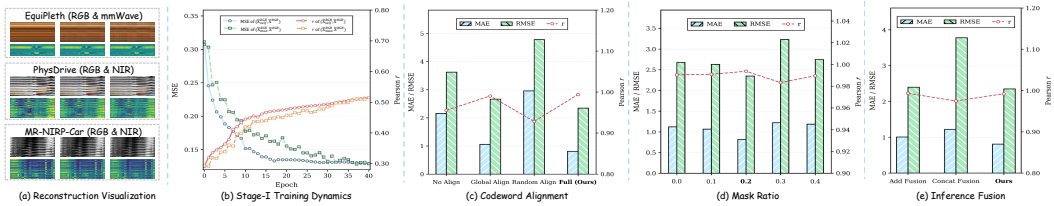


Figure 3: Qualitative and ablation analyses of PhysioBridge. Panel (a) shows the targets, self-reconstructions, and mutual reconstructions from left to right. Panel (b) reports the Stage-I training dynamics of reconstruction. Panels (c)–(e) are evaluated under multimodal inference.

Table 4: Ablation study of physiological code space construction on EquiPleth.

Variant	Code Query Process	Mutual Rec.	$\mathcal{L}_{\text{orth}}$	MAE↓	RMSE↓	r ↑
Continuous Matching	✗	✓	–	4.82	7.13	0.833
w/o $\mathcal{L}_{\text{orth}}$	✓	✓	✗	1.82	4.20	0.965
w/o Mutual Rec.	✓	✗	✓	<u>1.32</u>	<u>3.33</u>	<u>0.982</u>
Full	✓	✓	✓	**0.81**	**2.35**	**0.994**

space supports cross-dataset transfer without requiring sample-level cross-modal correspondence during training. We further evaluate learning from genuinely separate acquisition sources by jointly training on PhysDrive RGB and MR-NIRP-Car NIR, with results reported in Appendix B.1.

4.4 Ablation Study

We evaluate the key components of PhysioBridge on EquiPleth under multimodal inference, as summarized in Tab. 4 and Fig. 3.

Physiological Code Space Construction. As shown in Tab. 4, "Continuous Matching" replaces discrete codeword querying with continuous feature matching, retaining mutual reconstruction but omitting codeword orthogonality and distribution alignment. "w/o Mutual Rec." removes cross-modal reconstruction while retaining self-reconstruction, whereas "w/o $\mathcal{L}_{\text{orth}}$" removes the orthogonality constraint between private and shared codewords. Continuous Matching increases MAE from 0.81 to 4.82, while removing mutual reconstruction and orthogonality increases MAE to 1.32 and 1.82, respectively. These results highlight the contributions of discrete physiological code learning, cross-modal mutual reconstruction, and private–shared representation complementarity. Fig. 3(a) visualizes cross-modal reconstruction across different datasets, while Fig. 3(b) shows consistently decreasing reconstruction errors and increasing correlations in both directions during Stage-I training. Fig. 3(d) further examines the masking ratio, where 0.2 achieves the best performance. Smaller ratios provide insufficient reconstruction difficulty, whereas larger ratios may discard essential physiological information.

Codeword Distribution Alignment. Fig. 3(c) compares four Stage-II variants. "No Align" omits distribution alignment; "Global Align" matches codeword activation distributions across all HR intervals; and "Random Align" applies distribution matching to randomly grouped cross-modal samples. Our method instead aligns samples within comparable HR intervals. Global alignment reduces MAE from 2.18 to 1.08, whereas random alignment increases it to 2.96, even exceeding the error without alignment. This suggests that indiscriminate distribution matching can introduce physiologically inconsistent correspondences. In contrast, HR-conditioned alignment achieves the lowest MAE and RMSE of 0.81 and 2.35, demonstrating the importance of physiological comparability for unpaired cross-modal alignment.

Code-Confidence Fusion. Fig. 3(e) compares our confidence-weighted fusion with direct feature addition ("Add Fusion") and concatenation ("Concat Fusion"). The latter two combine modality features without explicitly accounting for their reliability, yielding MAEs of 1.00 and 1.20, respectively. In contrast, our method achieves an MAE of 0.81 and an RMSE of 2.35 by assigning token-level fusion weights based on codeword assignment entropy and quantization error. This adaptive weighting allows the model to emphasize shared physiological features with higher code confidence while reducing the contribution of lower-confidence features, facilitating the integration of complementary information across modalities.

5 Conclusion and Limitations

In this work, we introduced PhysioBridge, a framework for unpaired multimodal rPPG learning that eliminates the need for sample-level cross-modal correspondence during training. PhysioBridge constructs a unified discrete physiological code space through modality-private and shared representation learning, and employs HR-conditioned codeword distribution alignment to promote consistent codeword usage across unpaired modalities. At inference, code-confidence fusion adaptively integrates synchronized multimodal observations. Extensive experiments on public and self-collected datasets demonstrate competitive performance across different sensing configurations without paired training data. Nevertheless, PhysioBridge still has several limitations, including its reliance on HR annotations and physiological waveform supervision, and the requirement for synchronized multimodal observations at inference. Future work will explore weaker physiological supervision and fusion under imperfect temporal synchronization.

AI use statement

During the preparation of this paper, we used generative AI tools for language refinement, proofreading, and improving the clarity of academic writing. We have not used generative AI tools for generating research ideas, designing methodologies, conducting experiments, analyzing results, or producing scientific claims. All AI-assisted content was manually reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, and experimental results.

Ethics statement

This work involves human participants for the collection of the MWSHR dataset. All participants provided informed consent before data collection. The collected RGB, NIR, WiFi, and physiological recordings are used solely for academic research purposes. Due to privacy considerations involving facial and physiological signals, the dataset will be considered for public release after acceptance following necessary privacy protection procedures.

Reproducibility statement

To facilitate reproducibility, we provide the source code, model implementations, and necessary configurations for training and evaluation in the supplementary materials. The supplementary materials also include detailed descriptions of data preprocessing procedures, implementation settings, and evaluation procedures.

References

Jae-Ho Choi, Ki-Bong Kang, and Kyung-Tae Kim. Fusion-vital: Video-rf fusion transformer for advanced remote physiological measurement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1344–1352, 2024. doi: 10.1609/AAAI.V38I2.27898.

Shuyang Chu, Menghan Xia, Mengyao Yuan, Xin Liu, Tapio Seppänen, Guoying Zhao, and Jingang Shi. Codephys: Robust video-based remote physiological measurement through latent codebook querying. *IEEE Journal of Biomedical and Health Informatics*, 29:4932–4945, 2025.

Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60:2878–2886, 2013.

Qi Dou, Quande Liu, Pheng Ann Heng, and Ben Glocker. Unpaired multi-modal segmentation via knowledge distillation. *IEEE Transactions on Medical Imaging*, 39(7):2415–2425, 2020. doi: 10.1109/TMI.2019.2963882.

Jieyi Ge, Zhaodong Sun, Wei Peng, Chenhang Ying, Yuwei Chen, Kui Ren, and Xiaobai Li. 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. doi: 10.1145/3746027.3754594.

John Gideon and Simon Stent. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3995–4004, 2021.

Sharut Gupta, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. Better together: Leveraging unpaired multimodal data for stronger unimodal models. In *International Conference on Learning Representations (ICLR)*, 2026.

Qian Liang, Ruixu Geng, Jinbo Chen, Haoyu Wang, Yan Chen, and Yang Hu. Spatial alignment and temporal matching adapter for video-radar remote physiological measurement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8623–8633, 2025. doi: 10.1109/ICCV51701.2025.00807.

Changxing Liu, Zichen Chao, and Siheng Chen. Linking modality isolation in heterogeneous collaborative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39765–39774, 2026.

Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Yanfeng Wang, and Siheng Chen. An extensible framework for open heterogeneous collaborative perception. In *International Conference on Learning Representations (ICLR)*, 2024.

Daniel McDuff. Camera measurement of physiological vital signs. *ACM Computing Surveys*, 55: 176:1–176:40, 2023.

Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29: 2409–2423, 2020.

Ewa M. Nowara, Tim K. Marks, Hassan Mansour, and Ashok Veeraraghavan. Near-infrared imaging photoplethysmography during driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(4):3589–3600, 2022. doi: 10.1109/TITS.2020.3038317.

Hang Shao, Lei Luo, Jianjun Qian, Mengkai Yan, Shuo Chen, and Jian Yang. Remote photoplethysmography in real-world and extreme lighting scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10858–10867, June 2025.

Zhaodong Sun and Xiaobai Li. Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:5835–5851, 2024.

Trong Khiem Tran, Anh Duc Chu, Quang Hung Pham, Phi Le Nguyen, and Trong Nghia Hoang. Cross-modal knowledge distillation without paired data: Theoretical foundation and algorithm. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6306–6315, 2017.

Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16:21434–21445, 2008.

Alexander Vilesov, Pradyumna Chari, Adnan Armouti, Anirudh Bindiganavale Harish, Kimaya Kulkarni, Ananya Deoghare, Laleh Jalilian, and Achuta Kadambi. Blending camera and 77 ghz radar sensing for equitable, robust plethysmography. *ACM Transactions on Graphics*, 41 (4), 2022. doi: 10.1145/3528223.3530161.

Jiyao Wang, Xiao Yang, Qingyong Hu, Jack Tang, Can Liu, Dengbo He, Yuntao Wang, Yingcong Chen, and Kaishun Wu. Physdrive: A multimodal remote physiological measurement dataset for in-vehicle driver monitoring. In *Advances in Neural Information Processing Systems*, volume 38, pp. 41134–41155, 2025. doi: 10.52202/085713-1230.

Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering*, 64:1479–1491, 2017.

Ying Wang, Zhaodong Sun, Xu Cheng, Zuxian He, and Xiaobai Li. Radar-aplanc: Unsupervised radar-based heartbeat sensing via augmented pseudo-label and noise contrast. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10270–10278, 2026. doi: 10.1609/aaai.v40i12.37996.

Johnny Xi, Jana Osea, Zuheng Xu, and Jason Hartford. Propensity score alignment of unpaired multimodal data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4480.

Amitai Yacobi, Nir Ben-Ari, Ronen Talmon, and Uri Shaham. Learning shared representations from unpaired data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1552.

Chenhang Ying, Huiyu Yang, Jieyi Ge, Zhaodong Sun, Xu Cheng, Kui Ren, and Xiaobai Li. Fusionphys: A flexible framework for fusing complementary sensing modalities in remote physiological measurement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9363–9373, 2025. doi: 10.1109/ICCV51701.2025.00874.

Zitong Yu, Xiaobai Li, and Guoying Zhao. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *British Machine Vision Conference (BMVC)*, pp. 277, 2019.

Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip H. S. Torr, and Guoying Zhao. Physformer: Facial video-based physiological measurement with temporal difference transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4186–4196, 2022.

Zijie Yue, Miaojing Shi, Hanli Wang, Shuai Ding, Qijun Chen, and Shanlin Yang. Bootstrapping vision-language models for frequency-centric self-supervised remote physiological measurement. *International Journal of Computer Vision*, 133:4112–4133, 2025.

Bochao Zou, Zizheng Guo, Jiansheng Chen, Junbao Zhuo, Weiran Huang, and Huimin Ma. Rhythmformer: Extracting patterned rppg signals based on periodic sparse attention. *Pattern Recognition*, 164:111511, 2025.

APPENDIX

The appendix is organized into three sections. Readers can click on each section number to navigate to the corresponding section:

- Section A provides additional method details, including physiological resampling, network architecture, training procedures, and multimodal extension.
- Section B presents additional experimental analyses of unpaired learning, physiological code learning, distribution alignment, confidence-based fusion, and hyperparameter sensitivity.
- Section C provides dataset protocols, MWSHR collection details, and baseline implementation information.

A ADDITIONAL METHOD DETAILS

This section provides additional implementation details of physiological matching, network architecture, and training procedures, followed by the extension of PhysioBridge to three sensing modalities.

A.1 Physiological Matching and Temporal Resampling

During Stage-I, recordings from different modalities are independently sampled within the same HR interval $\mathcal{B}_q$. A common target HR $h^* \in \mathcal{B}_q$ is selected for each sampled group, and the modality-specific inputs are resampled according to their original HRs:

$$\widetilde{\mathbf{X}}^m = \mathcal{R}\left(\mathbf{X}^m; \frac{h^*}{h^m}\right), \qquad m \in \mathcal{M}, \tag{22}$$

where $\mathcal{R}$ denotes temporal resampling and $\mathcal{M}$ is the set of training modalities.

Temporal resampling is performed on the complete modality-specific sequences using linear interpolation. The resampled sequences may vary in length, from which fixed-length input windows are subsequently extracted. The same procedure is applied to all sensing modalities.

Physiological matching produces inputs with comparable cardiac periods without introducing sample-level cross-modal correspondence. The resampled representations are masked before encoding, whereas self- and mutual-reconstruction losses target the complete resampled representations before masking. Temporal resampling is used only for Stage-I code space construction; Stage-II training and inference do not apply this physiological matching procedure.

A.2 Network Architecture and Optimization

Network Architecture. PhysioBridge employs modality-specific encoders $\mathcal{E}^m$ and decoders $\mathcal{D}^m$. The RGB and NIR encoders adopt the PhysNet encoder (Yu et al., 2019), while the mmWave and WiFi encoders follow the EquiPleth encoder (Vilesov et al., 2022). Each decoder mirrors its corresponding encoder.

The heterogeneous inputs are mapped to latent representations with a common dimensionality $D = 256$ and token count $L = 64$. Each modality maintains a private codebook $\mathcal{C}_p^m$, while all modalities query the unified shared physiological codebook $\mathcal{C}_s$. Each codebook contains $K = 128$ codewords. The shared rPPG prediction head $\mathcal{H}$ consists of a point-wise 3D convolutional layer followed by a max-pooling layer.

Optimization. In Stage-I, the modality-specific encoders and decoders, private codebooks, and shared codebook are jointly optimized using reconstruction, vector-quantization, and orthogonality objectives. We employ a straight-through estimator (STE) to propagate gradients through the hard codeword queries.

In Stage-II, the modality-specific encoders are initialized from Stage-I, while the shared codebook $\mathcal{C}_s$ is frozen. The same STE is used for the hard shared-code queries, allowing the rPPG supervision to update the modality-specific encoders while keeping the shared codebook fixed. The encoders and shared prediction head are optimized using independent rPPG supervision and HR-conditioned codeword distribution alignment. The training hyperparameters are provided in the main paper.

A.3 Training and Inference Procedures

Stage-I Training. Within each HR interval, recordings are independently sampled from the available modalities and resampled to a common target HR. The resulting inputs are masked and processed by their respective encoders. Each latent representation queries its modality-private codebook and the unified shared codebook. Self-reconstruction combines private and shared codes from the same modality, while mutual reconstruction retains the target modality's private codes and exchanges only the shared codes across modalities. Both reconstruction objectives use the complete resampled target representations before masking. The encoders, codebooks, and decoders are jointly optimized using the Stage-I objective.

Stage-II Training. The modality-specific encoders are initialized from Stage-I, and the shared codebook is frozen. Each modality is independently sampled without constructing pseudo-paired inputs or applying Stage-I physiological resampling. The encoded features query the shared codebook and are processed by the shared rPPG prediction head. Independent waveform and frequency supervision is combined with HR-conditioned codeword distribution alignment to optimize the encoders and prediction head.

Inference. Synchronized observations from the available modalities are independently encoded and mapped to the shared physiological codebook. Token-level code confidence is computed using codeword assignment entropy and quantization error. The queried shared codeword features are weighted and fused before being processed by the shared rPPG prediction head. The same trained framework also supports inference from a single available modality.

A.4 Multimodal Extension of PhysioBridge

Although the main formulation considers two modalities, PhysioBridge can be extended to additional sensing modalities without changing its two-stage learning framework.

Physiological Code Space Construction. Let $\mathcal{M} = \{\mathrm{RGB}, \mathrm{NIR}, \mathrm{WiFi}\}$ denote the three sensing modalities. Each modality $m \in \mathcal{M}$ maintains its own encoder $\mathcal{E}^m$, private codebook $\mathcal{C}_p^m$, and decoder $\mathcal{D}^m$, while all modalities query the shared physiological codebook $\mathcal{C}_s$.

During Stage-I, recordings are independently sampled from the three modalities within the same HR interval and resampled to a common target HR. Each modality performs self-reconstruction using its own private and shared codes. For mutual reconstruction, the target modality retains its private codes and directly exchanges shared codes with another modality.

For target modality m and source modality $n \neq m$, mutual reconstruction is defined as

$$\widehat{\mathbf{X}}_{\mathrm{mut}}^{m \leftarrow n} = \mathcal{D}^m \left(\left[\widehat{\mathbf{Z}}_p^m ; \widehat{\mathbf{Z}}_s^n \right] \right). \tag{23}$$

The self- and mutual-reconstruction losses are averaged over the three modalities and all ordered cross-modal combinations:

$$\begin{aligned} \mathcal{L}_{\mathrm{self}}^{(3)} &= \frac{1}{3} \sum_{m \in \mathcal{M}} \ell_{\mathrm{rec}} \left(\widehat{\mathbf{X}}_{\mathrm{self}}^m, \widetilde{\mathbf{X}}^m \right), \\ \mathcal{L}_{\mathrm{mut}}^{(3)} &= \frac{1}{6} \sum_{m \in \mathcal{M}} \sum_{\substack{n \in \mathcal{M} \\ n \neq m}} \ell_{\mathrm{rec}} \left(\widehat{\mathbf{X}}_{\mathrm{mut}}^{m \leftarrow n}, \widetilde{\mathbf{X}}^m \right). \end{aligned} \tag{24}$$

Thus, mutual reconstruction considers all six ordered exchanges among RGB, NIR, and WiFi. The orthogonality and vector-quantization objectives are applied to the private and shared codeword queries of all three modalities, following their definitions in the main paper. The resulting Stage-I objective is

$$\mathcal{L}_{\mathrm{I}}^{(3)} = \mathcal{L}_{\mathrm{self}}^{(3)} + \mathcal{L}_{\mathrm{mut}}^{(3)} + \lambda_{\mathrm{VQ}} \mathcal{L}_{\mathrm{VQ}}^{(3)} + \lambda_{\mathrm{orth}} \mathcal{L}_{\mathrm{orth}}^{(3)}. \tag{25}$$

Codeword Distribution Alignment. During Stage-II, the three modality-specific encoders are initialized from Stage-I, and the shared codebook $\mathcal{C}_s$ is frozen. Each modality receives independent rPPG supervision, while its conditional codeword usage distribution $\boldsymbol{\pi}_q^m$ is estimated within HR interval $\mathcal{B}_q$.

We extend the alignment objective to all ordered modality pairs:

$$\mathcal{L}_{\mathrm{align}}^{(3)} = \frac{1}{|\mathcal{I}||\mathcal{P}|} \sum_{q \in \mathcal{I}} \sum_{(m,n) \in \mathcal{P}} D_{\mathrm{KL}} \left(\boldsymbol{\pi}_q^m \| \boldsymbol{\pi}_q^n \right), \tag{26}$$

where

$$\mathcal{P} = \{(m,n) \in \mathcal{M} \times \mathcal{M} \mid m \neq n\}, \qquad |\mathcal{P}| = 6, \tag{27}$$

and $\mathcal{I}$ contains HR intervals with available samples from all three modalities. The six ordered pairs account for both directions of KL divergence for each of the RGB–NIR, RGB–WiFi, and NIR–WiFi combinations.

The independent rPPG supervision is averaged over the three modalities:

$$\mathcal{L}_{\mathrm{rPPG}}^{(3)} = \frac{1}{3} \sum_{m \in \mathcal{M}} \left[\mathcal{L}_{\mathrm{NP}} \left(\widehat{\mathbf{y}}^m, \mathbf{y}^m \right) + \mathcal{L}_{\mathrm{SNR}} \left(\widehat{\mathbf{y}}^m, h^m \right) \right]. \tag{28}$$

Table 5: HR estimation with independently acquired training modalities on PhysDrive and MR-NIRP-Car.

Method	PhysDrive			MR-NIRP-Car		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
UMKD (Dou et al., 2020)	7.14	8.13	0.43	8.15	11.2	0.20
UML (Gupta et al., 2026)	5.53	6.37	0.51	5.95	10.7	0.36
CodeAlign (Liu et al., 2026)	6.91	7.63	0.44	6.35	8.87	0.50
PhysioBridge	**5.17**	**6.23**	**0.66**	**2.54**	**4.53**	**0.83**

The overall Stage-II objective becomes

$$\mathcal{L}_{\mathrm{II}}^{(3)} = \mathcal{L}_{\mathrm{rPPG}}^{(3)} + \lambda_{\mathrm{align}} \mathcal{L}_{\mathrm{align}}^{(3)}. \tag{29}$$

This extension promotes consistent codeword usage across the three modalities under comparable physiological conditions without requiring sample-level cross-modal correspondence during training.

Code-Confidence Fusion. At inference, let $\mathcal{M}_{\mathrm{avail}} \subseteq \mathcal{M}$ denote the synchronized modalities available for fusion. For each modality m and token l, the confidence score r_l^m is computed from assignment entropy and quantization error, following the bimodal formulation in the main paper.

The confidence weights are normalized over the available modalities:

$$w_l^m = \frac{r_l^m + \epsilon}{\sum_{n \in \mathcal{M}_{\mathrm{avail}}} (r_l^n + \epsilon)}, \qquad m \in \mathcal{M}_{\mathrm{avail}}. \tag{30}$$

The queried shared codeword features are then fused as

$$\widehat{\mathbf{z}}_l^F = \sum_{m \in \mathcal{M}_{\mathrm{avail}}} w_l^m \widehat{\mathbf{z}}_{s,l}^m, \qquad \widehat{\mathbf{y}}^F = \mathcal{H}\left(\widehat{\mathbf{Z}}^F\right). \tag{31}$$

Here, $\widehat{\mathbf{z}}_{s,l}^m$ denotes the shared feature obtained through the hard codeword query. When all three modalities are available, the fused representation combines RGB, NIR, and WiFi shared features. The same formulation supports bimodal inference using any two available modalities. For unimodal inference, the shared features of the available modality are passed directly to $\mathcal{H}$.

B Additional Experimental Analysis

B.1 Learning from Independently Acquired Modalities

The main experiments use modality-specific training samples without their available cross-modal correspondence. We further test a setting in which the training modalities come from separate acquisition sources.

Experimental Setting. We train one PhysioBridge model using PhysDrive RGB recordings and MR-NIRP-Car NIR recordings. No sample-level cross-modal correspondence is available between these sources. Evaluation uses synchronized RGB and NIR recordings from each dataset's test set. Thus, one branch encounters a different acquisition source at inference on either dataset.

PhysioBridge achieves the lowest MAE on both test sets in Tab. 5. Its 5.17 bpm MAE on PhysDrive improves upon the closest baseline by 0.36 bpm. Its 2.54 bpm MAE on MR-NIRP-Car improves upon the closest baseline by 3.41 bpm. RMSE and HR correlation likewise favor PhysioBridge. The corresponding intra-dataset MAEs in the main paper are 4.39 bpm and 1.89 bpm. The increase under independently acquired training is consistent with a more difficult source combination. This comparison also changes the training data, so it does not isolate acquisition-source shift. The result nevertheless shows that a single model can learn from separate modality sources, then use synchronized inputs from either test set.

Table 6: Effect of physiological matching on multimodal HR estimation.

Sampling Strategy	EquiPleth			MR-NIRP-Car			PhysDrive		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
Random Sampling	2.58	5.20	0.94	2.94	5.31	0.73	6.16	7.96	0.54
HR-bin Sampling	1.24	3.01	0.98	2.33	4.21	0.87	4.71	5.52	0.72
Physiological Matching	**0.81**	**2.35**	**0.99**	**1.89**	**2.51**	**0.95**	**4.39**	**5.50**	**0.74**

Table 7: Additional component ablations on PhysDrive and MR-NIRP-Car.

Variant	PhysDrive			MR-NIRP-Car		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
Continuous Matching	7.37	8.38	0.38	5.86	8.96	0.19
w/o $\mathcal{L}_{\text{orth}}$	5.69	6.97	0.49	3.87	5.75	0.88
w/o Mutual Rec.	5.28	6.46	0.59	4.02	6.16	0.76
Full PhysioBridge	**4.39**	**5.50**	**0.74**	**1.89**	**2.51**	**0.95**

B.2 Physiological Code Space Construction

We examine physiological matching and the principal components of Stage-I across the available sensing configurations.

B.2.1 Physiological Matching Analysis

We compare three Stage-I sampling strategies. *Random Sampling* selects recordings without an HR constraint. *HR-bin Sampling* selects recordings within the same HR interval. *Physiological Matching* additionally resamples these independently selected recordings to a common target HR within that interval. The architecture, codeword queries, reconstruction objectives, Stage-II training, and inference protocol remain fixed.

Each reconstruction target is the target modality's own complete representation before masking. Under Physiological Matching, this representation is obtained after resampling. No variant reconstructs a physiological ground-truth waveform during Stage-I. Stage-II operates at the original physiological time scale.

Tab. 6 separates the effect of HR-bin selection from the additional effect of temporal resampling. Relative to Random Sampling, HR-bin Sampling lowers MAE from 2.58 to 1.24 bpm on EquiPleth, from 2.94 to 2.33 bpm on MR-NIRP-Car, and from 6.16 to 4.71 bpm on PhysDrive. Physiological Matching further lowers these errors to 0.81, 1.89, and 4.39 bpm. The matching strategy also gives the best RMSE and correlation on every dataset. Comparable HR conditions therefore improve Stage-I learning even before resampling. Matching the physiological period provides an additional benefit without creating sample-level correspondence.

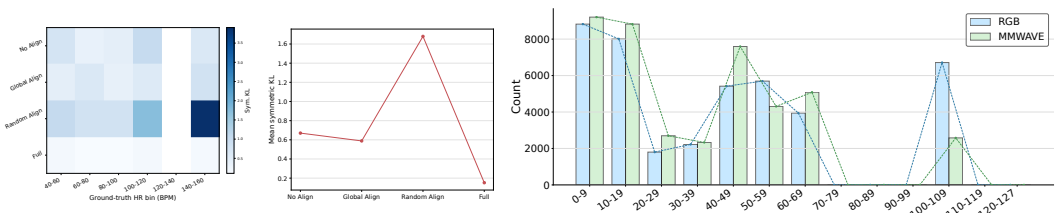
B.2.2 Component Ablations on Additional Datasets

We repeat the component ablations on PhysDrive and MR-NIRP-Car. Removing $\mathcal{L}_{\text{orth}}$ retains the other objectives. Removing mutual reconstruction retains self-reconstruction. Continuous Matching replaces discrete codeword queries with continuous feature matching. It also omits codeword orthogonality and distribution alignment, so it evaluates a broader alternative representation scheme rather than quantization alone.

Full PhysioBridge achieves 4.39 bpm on PhysDrive and 1.89 bpm on MR-NIRP-Car. Removing mutual reconstruction raises MAE by 0.89 bpm and 2.13 bpm, respectively. Removing orthogonality raises it by 1.30 bpm and 1.98 bpm. Both components contribute across these datasets, although their relative effects differ. Continuous Matching reaches 7.37 bpm and 5.86 bpm. Its larger errors support the complete discrete-code framework, but cannot be attributed solely to quantization because this variant changes several mechanisms.

B.3 Codeword Distribution Alignment

We compare four Stage-II strategies on EquiPleth. *No Align* omits distribution alignment. *Global Align* matches codeword usage without HR conditioning. *Random Align* uses randomly grouped



(a) (b) (c)

Figure 4: Shared codeword usage on EquiPleth. Panel (a) shows symmetric KL divergence between RGB and mmWave usage distributions across valid HR intervals. HR intervals without valid test samples are left blank or omitted. Panel (b) reports the mean divergence over valid HR intervals for each alignment strategy. Panel (c) shows activation counts by codeword index range under Full PhysioBridge.

cross-modal samples. *Full* aligns usage within the same HR interval. For each valid test interval, we compute symmetric KL divergence between the RGB and mmWave codeword usage distributions. Valid intervals contain samples from both modalities. We also inspect shared codeword activation counts under Full.

Full has the lowest conditional divergence across the reported HR intervals in panel a of Fig. 4. Panel b gives mean symmetric KL values of approximately 0.15 for Full, 0.67 for No Align, 0.59 for Global Align, and 1.68 for Random Align. Global Align yields a modest reduction relative to No Align. Random Align instead produces the largest divergence. The corresponding main-paper MAEs are 0.81, 2.18, 1.08, and 2.96 bpm in the same order. These results connect improved HR estimation with more consistent codeword usage under comparable physiological conditions. They also show that the choice of alignment condition matters.

Panel c of Fig. 4 shows overlapping activation in several codeword ranges under Full, including indices 0–19 and 40–69. Activation remains nonuniform across the codebook. This panel describes usage by the trained model. The conditional KL measurements provide the direct comparison among alignment strategies. Neither visualization establishes sample-level cross-modal correspondence.

B.4 Code-Confidence Fusion

We assess whether code-confidence weights track relative unimodal prediction quality. We then evaluate inference with different modality subsets on MWSHR.

B.4.1 Confidence and Prediction Error Analysis

Fusion weights derive from codeword assignment entropy and quantization error. For synchronized test sample i, the mean weight of modality m is

$$\bar{w}_i^m = \frac{1}{L} \sum_{l=1}^{L} w_{i,l}^m. \tag{32}$$

Let $e_i^m = |\hat{h}_i^m - h_i|$ be its unimodal HR error. We compare relative weights with relative errors through

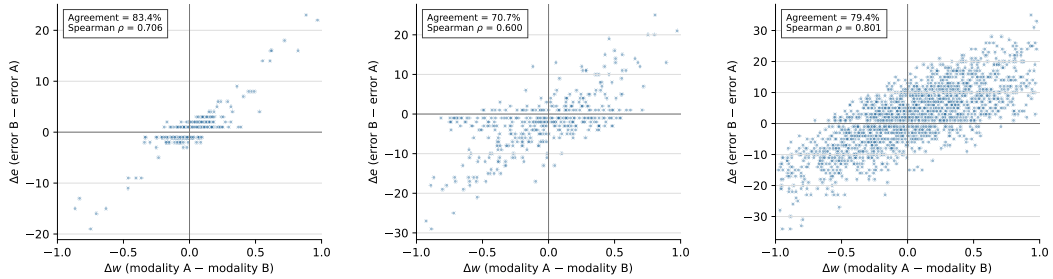
$$\Delta w_i = \bar{w}_i^a - \bar{w}_i^b, \qquad \Delta e_i = e_i^b - e_i^a. \tag{33}$$

A positive $\Delta w_i \Delta e_i$ means that the modality receiving more weight has the smaller unimodal error. Agreement is the percentage of samples with a positive product after excluding tied weights or tied errors. Spearman's ρ measures the monotonic association between the signed differences. These statistics assess relative prediction reliability. They do not measure probabilistic calibration.

Tab. 8 reports agreement rates of 83.4% on EquiPleth, 70.7% on MR-NIRP-Car, and 79.4% on PhysDrive. The respective Spearman correlations are 0.71, 0.60, and 0.80. Thus, the higher-weight modality has the lower unimodal HR error for most non-tied samples on every dataset. The positive correlations show that relative weight differences generally track relative error advantages.

Table 8: Correspondence between code-confidence weights and unimodal HR estimation errors.

Dataset	Agreement (%)↑	Spearman ρ ↑
EquiPleth	83.4	0.71
MR-NIRP-Car	70.7	0.60
PhysDrive	79.4	0.80



(a) EquiPleth. (b) MR-NIRP-Car. (c) PhysDrive.

Figure 5: Relationship between modality weight difference Δw and unimodal HR error difference Δe. Each point represents one synchronized test sample. Points in the first or third quadrant indicate agreement between the higher-weight modality and the lower-error modality.

Fig. 5 exposes the sample-level variation behind these summaries. Agreement is weaker on MR-NIRP-Car than on the other two datasets, yet remains above 70%. The results support code confidence as a useful reliability signal for adaptive fusion. Individual disagreements remain expected because assignment entropy and quantization error do not directly measure HR prediction error.

B.4.2 Fusion Strategy and Modality Subset Analysis

We evaluate all available modality subsets on MWSHR using one PhysioBridge model trained with three unpaired modalities. At inference, we retain the selected branches, then normalize confidence scores over those branches. No subset-specific retraining is performed.

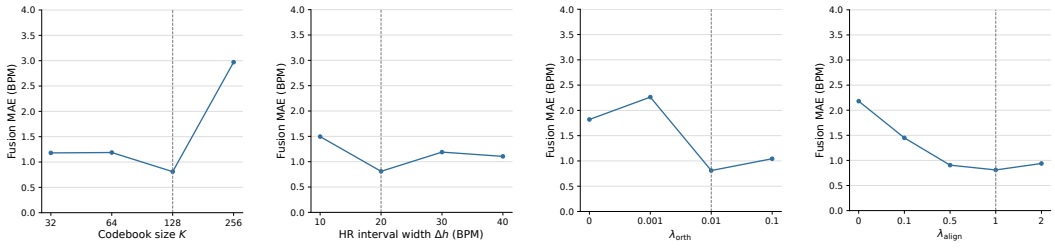
RGB is the strongest single modality in Tab. 9, with an MAE of 2.39 bpm. Adding NIR reduces it to 2.08 bpm. Adding WiFi to RGB yields 2.33 bpm. NIR and WiFi improves MAE relative to either modality alone, reaching 4.57 bpm, although its RMSE rises to 7.78 bpm. The three-modality combination gives the best overall result at 1.93 bpm MAE, 2.40 bpm RMSE, and 0.90 correlation. The 0.15 bpm MAE gain over RGB and NIR indicates a further contribution from WiFi in the complete fusion setting. The mixed pairwise results also show that an additional modality need not improve every metric. All seven configurations use the same trained model.

B.5 Hyperparameter Sensitivity Analysis

We vary codebook size K, HR interval width Δh, orthogonality weight λ_{orth}, and alignment weight λ_{align} on EquiPleth. Each experiment changes one parameter while retaining the other default settings. Fig. 6 shows that the default settings achieve the lowest MAE in all four sweeps. For codebook size, $K = 128$ yields 0.81 bpm, compared with 1.18, 1.19, and 2.97 bpm at $K = 32, 64, 256$. The marked increase at $K = 256$ indicates that additional codewords do not necessarily improve physiological representation learning. For HR interval width, $\Delta h = 20$ bpm achieves 0.81 bpm. Widths of 10, 30, and 40 bpm yield 1.50, 1.19, and 1.11 bpm, respectively. This favors an intermediate interval that provides comparable physiological conditions while retaining sufficient training samples. The default orthogonality weight of 0.01 achieves 0.81 bpm. Setting λ_{orth} to 0, 0.001, and 0.1 increases MAE to 1.82, 2.26, and 1.04 bpm. The alignment sweep shows a similar optimum at $\lambda_{\text{align}} = 1$. Its MAE of 0.81 bpm compares with 2.18, 1.45, 0.91, and 0.94 bpm at weights of 0, 0.1, 0.5, and 2. These results support both objectives, while showing that their weights require appropriate balance.

Table 9: HR estimation with different modality combinations on MWSHR.

Test Modality	MAE↓	RMSE↓	r ↑
RGB	2.39	3.77	0.89
NIR	5.70	6.96	0.45
WiFi	5.20	7.20	0.42
RGB + NIR	2.08	3.06	0.89
RGB + WiFi	2.33	3.49	0.79
NIR + WiFi	4.57	7.78	0.52
RGB + NIR + WiFi	**1.93**	**2.40**	**0.90**



(a) (b) (c) (d)

Figure 6: Hyperparameter sensitivity on EquiPleth. Panel a varies codebook size K. Panel b varies HR interval width Δh. Panel c varies orthogonality weight λ_{orth}. Panel d varies distribution alignment weight λ_{align}.

C Dataset Collection and Baseline Implementation Details

This section provides additional details of the MWSHR dataset and the implementation protocols of the compared baseline methods.

C.1 MWSHR Data Collection

Participants and Recording Protocol. MWSHR contains synchronized RGB, NIR, and WiFi recordings from 21 subjects. Each subject participates in six recording tasks, with each task lasting approximately 60 seconds. The six tasks are constructed by combining two motion conditions with three illumination conditions. The motion conditions include stationary and conversational motion. For the conversational condition, subjects are instructed to naturally interact during recording, resulting in spontaneous head movements and head rotations. The illumination conditions include dim, natural, and bright environments. We adopt a subject-independent split with 17 subjects for training and 4 subjects for testing.

Acquisition Setup and Synchronization. Figure 7 illustrates the acquisition setup and representative multimodal recordings. RGB and NIR videos are captured using a WSD-2133-V1-RGB/IR camera with a resolution of 1080P at 30 fps. The camera provides synchronized RGB and NIR observations from the same recording session.

WiFi CSI signals are collected using two Lenovo X201 laptops equipped with Intel 5300 network interface cards. One laptop serves as the transmitter and the other serves as the receiver. The CSI measurements are acquired under the 802.11n HT20 configuration. Each measurement contains three receiving antennas and 30 OFDM subcarriers, represented by complex-valued responses with real and imaginary components. The CSI acquisition rate is approximately 300 Hz.

A CONTEC CMS60C pulse oximeter is used to obtain reference pulse signals. All modalities are synchronized according to their timestamps. The overlapping temporal range among RGB, NIR, WiFi CSI, and reference signals is first identified. The WiFi CSI sequences are then temporally resampled to the video temporal grid through linear interpolation, yielding temporally aligned multimodal observations.

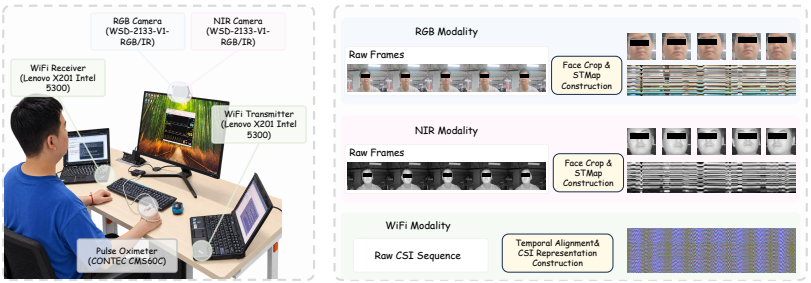


Figure 7: Overview of the MWSHR dataset. (a) Acquisition setup showing a subject and the sensing devices during recording. (b) Example synchronized RGB, NIR, and WiFi representations from the same recording.

Data Preprocessing. For RGB and NIR modalities, facial regions are extracted from the recorded videos and transformed into spatio-temporal representations for model training. For the WiFi modality, the temporally aligned CSI measurements are represented using their real and imaginary components. Measurements from three receiving antennas and 30 subcarriers are reorganized into a $180 \times T$ temporal representation for model input.

All recordings are segmented into synchronized temporal windows. Samples from different modalities within the same window share the same temporal range while preserving their modality-specific characteristics. The resulting multimodal representations are used for subsequent training and evaluation.

C.2 Baseline Implementation Details

Evaluation Protocol. We evaluate all baseline methods using the dataset splits, modality configurations, and HR estimation metrics described in the main paper. For reproduced learning-based methods, the input branches are adapted to the corresponding modality-specific representations. Predicted physiological waveforms are converted into HR using the same evaluation procedure.

Unimodal Baselines. CHROM (De Haan & Jeanne, 2013) and POS (Wang et al., 2017) follow their respective signal-processing procedures. Learning-based unimodal methods are trained and evaluated using their designated sensing modalities, with input adaptations where necessary to accommodate the corresponding RGB, NIR, or mmWave representations.

Paired Multimodal Baselines. We reproduce EquiPleth (Vilesov et al., 2022), Fusion-Vital (Choi et al., 2024), Evidential-Phys (Ge et al., 2025), and FusionPhys (Ying et al., 2025) using synchronized multimodal recordings. For the RGB–NIR–WiFi setting on MWSHR, we extend their modality-specific input branches to accommodate the third modality while retaining their respective fusion strategies.

Unpaired Multimodal Baselines. We adapt UMKD (Dou et al., 2020), UML (Gupta et al., 2026), and CodeAlign (Liu et al., 2026) to physiological waveform prediction. UMKD applies modality-specific waveform supervision and additional cross-modal knowledge distillation through HR spectral distributions. UML employs modality-specific input adapters, a shared temporal encoder, and a shared waveform prediction head, with latent representations averaged for multimodal inference. For CodeAlign, we construct a shared discrete code space through modality-specific self-reconstruction and cross-modal code-statistics alignment. We then train explicit code-to-code translators between modalities, supervising each translated prediction with the source recording's physiological waveform and HR annotations. At inference, the native and translated waveform predictions are averaged to obtain the multimodal output.

All unpaired baselines are trained using independently sampled modality-specific recordings without accessing sample-level cross-modal correspondence. At inference, synchronized observations from the available modalities are used for multimodal prediction.