PHYSVR: VISION-LANGUAGE MODEL GUIDED INTERFERENCE-AWARE TEMPORAL FEATURE REFINEMENT FOR REMOTE PHYSIOLOGICAL MEASUREMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote photoplethysmography (rPPG) enables contactless physiological measurement from facial videos, yet its subtle pulse-related variations are easily affected by illumination variation, head motion, facial blur, and region-of-interest instability. Existing methods mainly suppress interference during feature learning. However, whether interference persists in the learned temporal features, and how it can be further reduced before rPPG estimation, have received limited attention. To address this limitation, we propose PhysVR, a vision-language model guided interference-aware temporal feature refinement framework for rPPG estimation. Specifically, a physiological backbone produces global temporal features and coarse rPPG prediction, whose local temporal characteristics are used to construct time-resolved physiological reliability evidence. In parallel, a frozen vision-language model processes sampled facial frames under an interference-oriented prompt to provide visual interference evidence. Temporal cross-attention integrates the physiological and visual evidence with the global temporal features to construct interference-aware temporal context. Guided by this context, a shared temporal correction unit performs general refinement, while four interference-specific experts are adaptively routed to suppress different interference. Extensive experiments on five public benchmarks demonstrate that PhysVR consistently outperforms representative methods under both intra-dataset and cross-dataset evaluation protocols.

1 INTRODUCTION

Remote photoplethysmography (rPPG) enables contactless physiological measurement by capturing subtle variations in facial skin color induced by cardiac blood-volume changes. Requiring only a camera and no physical contact, rPPG has been investigated for remote physiological monitoring and clinical care (McDuff, 2023; Yao et al., 2026; Li et al., 2026), affective-state analysis (Casado et al., 2023), and mobile physiological sensing (Liu et al., 2022). However, reliable rPPG estimation remains challenging in practical environments, where weak pulse-related variations can be easily obscured by illumination variation, head motion, facial blur, and region-of-interest (ROI) instability.

Existing deep-learning based rPPG methods mainly follow two feature modeling strategies. STMap-based approaches organize temporal color traces extracted from multiple facial regions into structured spatio-temporal maps, whereas video-based approaches learn spatial and temporal features directly from facial clips (Liang et al., 2024; Chen & McDuff, 2018). Their architectures have evolved from convolutional networks to transformer-based and state-space models (Yu et al., 2022; Zou et al., 2025b). Despite differences in feature extraction and temporal modeling, most existing methods directly regress rPPG signals from the learned features.

Under this direct estimation paradigm, the feature extractor must preserve weak pulse-related variations while suppressing non-physiological changes. Illumination variation alters facial color and intensity distributions, head motion and ROI instability disrupt temporal correspondence across frames, and facial blur attenuates subtle skin-color variations (Wang et al., 2015; Song et al., 2021; Liu et al., 2024b; Zhang et al., 2024). Existing methods generally address these factors during feature learning and subsequently use the learned temporal features for signal estimation. However,

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.

whether these features remain affected by interference after feature learning and how to further suppress such interference before prediction are rarely examined.

Recently, pretrained vision-language models (VLMs) and large language models (LLMs) have been introduced into remote physiological measurement. Existing studies exploit cross-modal knowledge to improve physiological feature learning and signal estimation (Yue et al., 2025; Xie et al., 2026). rPPG-VQA further combines signal-quality analysis with multimodal scene assessment to select suitable videos for unsupervised rPPG training (Dai et al., 2026). These studies demonstrate the potential of language-guided visual understanding for remote physiological measurement. Nevertheless, the potential of VLMs to characterize visual interference for subsequent temporal feature refinement remains largely unexplored.

To address this problem, we propose PhysVR, a vision-language model guided interference-aware temporal feature refinement framework for robust rPPG estimation. Since interference may remain in the learned temporal features after feature learning, PhysVR further refines these features before final rPPG estimation. As illustrated in Fig. 1, most existing methods directly estimate rPPG from the learned temporal features after feature learning. However, under challenging visual conditions, these features may still be affected by interference, leading to deviations in the estimated rPPG signal. PhysVR therefore uses physiological reliability evidence and visual interference evidence to guide further refinement of the learned temporal features before final rPPG estimation. Specifically, the physiological backbone extracts global temporal features and predicts coarse rPPG signal. Based on the local temporal characteristics of the coarse prediction, time-resolved physiological reliability evidence is derived to reflect its reliability at different temporal positions. In parallel, a frozen VLM processes sampled facial frames under an interference-oriented prompt, and an evidence head derives clip-level visual interference evidence to guide subsequent temporal feature refinement. Temporal cross-attention integrates the physiological reliability evidence and visual interference evidence with the global temporal features to construct interference-aware temporal context. Guided by this context, a shared temporal correction unit performs general refinement of the learned temporal features, while four interference-specific experts further suppress different types of interference through adaptive routing. The refined temporal features are then used for final rPPG estimation.

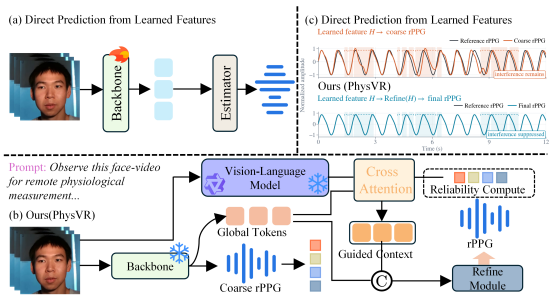


Figure 1: Comparison of direct rPPG estimation and PhysVR. (a) Direct rPPG prediction from learned temporal features. (b) PhysVR refines temporal features using physiological reliability and VLM-derived visual interference evidence. (c) Further refinement improves rPPG prediction under challenging visual conditions.

Our contributions are summarized as follows:

- We propose PhysVR, a vision-language model guided interference-aware temporal feature refinement framework that further refines learned temporal features after feature learning, reducing the influence of interference that may remain under challenging visual conditions.
- We construct an interference-aware temporal context by integrating physiological reliability evidence and VLM-derived visual interference evidence with global temporal features through temporal cross-attention, allowing complementary physiological and visual information to jointly guide the subsequent refinement process.
- We design a context-guided refinement module consisting of a shared temporal correction unit and four interference-specific experts. The shared unit performs general temporal refinement, while the router selectively activates interference-specific experts to further suppress different interference. Experiments on five public benchmarks demonstrate the effectiveness of PhysVR under both intra-dataset and cross-dataset.

2 Related Work

2.1 Remote Physiological Measurement

Early rPPG methods relied on handcrafted color-space projections and signal-processing priors to recover pulse signals while suppressing illumination changes and motion artifacts (De Haan & Jeanne, 2013; Wang et al., 2016). With the development of deep learning, the field gradually shifted toward data-driven spatio-temporal feature learning. Representative architectures include convolutional attention networks (Chen & McDuff, 2018), 3D spatio-temporal networks (Yu et al., 2019), temporal-shift models (Liu et al., 2020), transformer-based models (Yu et al., 2022), and efficient end-to-end networks (Liu et al., 2023a). More recent studies have further explored periodic sparse attention, state-space modeling, diffusion-based modeling, and physics-grounded temporal modeling (Zou et al., 2025a;b; Qian et al., 2025; Zhao et al., 2026b).

Beyond architectural advances, considerable efforts have focused on improving robustness to environmental interference and domain variation. ND-DeeprPPG disentangles environmental noise from pulse-related features using background regions as a reference (Liu & Yuen, 2023). A recent study (Shao et al., 2025) further models complex time-varying interference under real-world and extreme illumination conditions through interference sharing and disentanglement. FLOW improves cross-domain generalization by aligning multi-source features through optimal-transport-driven feature warping (Zhao et al., 2026a). Despite these advances, existing methods mainly suppress or disentangle interference during feature learning, or reduce distribution discrepancies across domains. The learned temporal features are typically used directly for rPPG estimation, without further interference-aware refinement.

2.2 Vision-Language Models

Vision-language models (VLMs) jointly model visual and textual information to support open-vocabulary recognition, visual grounding, image captioning, and multimodal reasoning (Radford et al., 2021; Li et al., 2022; Alayrac et al., 2022; Liu et al., 2024a). Their prompt-conditioned features have also been increasingly adapted to domain-specific visual understanding tasks.

Pretrained multimodal models have recently been introduced into remote physiological measurement for different purposes. VLPhys (Yue et al., 2025) align visual and textual features carrying physiological frequency information to facilitate self-supervised rPPG learning. PhysLLM incorporates a large language model (LLM) into rPPG estimation through cross-modal alignment and task-oriented physiological cues (Xie et al., 2026). rPPG-VQA employs a multimodal large language model (MLLM) to jointly assess signal quality and recording conditions, enabling the selection of suitable videos for unsupervised rPPG training (Dai et al., 2026). Collectively, these studies exploit pretrained multimodal models for physiological feature learning, signal estimation, or training-data selection. However, they primarily use multimodal information to support physiological representation learning or data assessment, rather than to provide visual interference information for further refinement of learned temporal features.

3 Methodology

As illustrated in Fig. 2, PhysVR consists of three main components: a physiological backbone, a VLM-based visual interference perception branch, and a context-guided refinement module. The physiological backbone provides temporal features and physiological reliability evidence for rPPG estimation, while the VLM-based branch provides visual interference evidence. PhysVR follows a two-stage training scheme: the physiological backbone is first trained for coarse rPPG estimation and then frozen, after which the refinement module is optimized to further refine the learned temporal features. The detailed components and training objective are described below.

3.1 Physiological Backbone

The physiological backbone extracts global temporal features $\mathbf{H}$ from frame differences and predicts a coarse rPPG signal $\hat{\mathbf{s}}^c$. The coarse signal is used to derive time-resolved physiological reliability evidence $\mathbf{R}$.

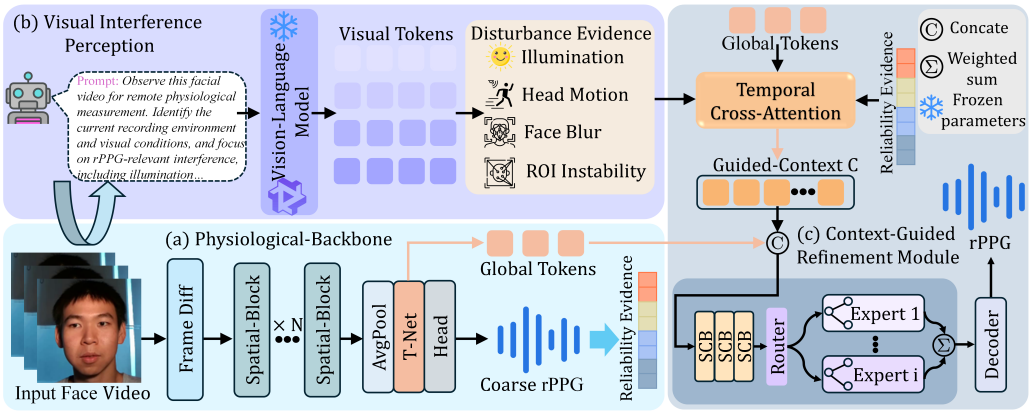


Figure 2: Overview of PhysVR. PhysVR refines learned temporal features using physiological reliability and VLM-derived visual interference evidence, with an interference-aware context guiding a shared correction unit and interference-specific experts through adaptive routing.

Given a facial video $\mathbf{V} = (\mathbf{I}_1, \ldots, \mathbf{I}_T) \in \mathbb{R}^{T\times 3\times H\times W}$, we construct a frame-difference sequence:

$$\begin{aligned}\Delta\mathbf{V} &= \mathcal{D}(\mathbf{V}) \\ &= [\mathbf{0}, \mathbf{I}_2 - \mathbf{I}_1, \ldots, \mathbf{I}_T - \mathbf{I}_{T-1}]\,.\end{aligned} \tag{1}$$

The zero tensor $\mathbf{0}$ has the same dimensions as a video frame and is inserted at the first position to preserve the sequence length. Frame differencing suppresses temporally invariant facial appearance while emphasizing adjacent-frame variations, including subtle pulse-related changes and non-physiological fluctuations caused by visual interference.

Spatial features are first extracted from the difference sequence and then processed along the temporal dimension:

$$\begin{aligned}\mathbf{F} &= E_s\left(\Delta\mathbf{V}\right), \\ \mathbf{H} &= E_T\left(E_p\left(\mathrm{AvgPool}_{\mathrm{sp}}(\mathbf{F})\right)\right).\end{aligned} \tag{2}$$

The spatial encoder $E_s(\cdot)$ consists of stacked Spatial-Blocks. As illustrated in Fig. 3(a), each Spatial-Block uses a residual bottleneck composed of pointwise and spatial convolutions. Spatial average pooling reduces the spatial resolution, while the projection layer $E_p(\cdot)$ adjusts the feature dimension for subsequent temporal modeling. As shown in Fig. 3(b), the $E_T(\cdot)$ contains two Depth-Blocks followed by two Spatial-Blocks to model temporal dependencies. The $\mathbf{H} \in \mathbb{R}^{T\times d}$ denotes the global temporal features.

A signal prediction head $h_s(\cdot)$ maps $\mathbf{H}$ to coarse rPPG signal:

$$\hat{\mathbf{s}}^c = h_s(\mathbf{H}), \tag{3}$$

where $\hat{\mathbf{s}}^c \in \mathbb{R}^T$ denotes the coarse rPPG signal from the frozen backbone. The physiological backbone provides global temporal features and coarse rPPG prediction, which are used to construct physiological reliability evidence.

We construct four time-aligned descriptors to capture local temporal characteristics of the coarse rPPG signal:

$$\begin{aligned}\mathbf{g}_{\mathrm{diff}} &= \left|\nabla_t \hat{\mathbf{s}}^c\right|, \\ \mathbf{g}_{\mathrm{band}} &= \mathcal{M}\left(\left|\mathcal{H}_{\mathrm{bp}}(\hat{\mathbf{s}}^c)\right|\right), \\ \mathbf{g}_{\mathrm{drift}} &= \left|\mathcal{H}_{\mathrm{lp}}(\hat{\mathbf{s}}^c)\right|, \\ \mathbf{g}_{\mathrm{dev}} &= \left[1 - \sigma\left(h_q(\mathbf{H})\right)\right] \odot \left|\hat{\mathbf{s}}^c - \mathcal{M}(\hat{\mathbf{s}}^c)\right|.\end{aligned} \tag{4}$$

The operator ∇_t denotes the first-order temporal difference, with a zero inserted at the first position to preserve the sequence length, and $\mathbf{g}_{\mathrm{diff}}$ measures changes between adjacent signal values. The fixed band-pass operator $\mathcal{H}_{\mathrm{bp}}(\cdot)$ retains frequencies within 0.7–3.0 Hz, while the fixed low-pass operator $\mathcal{H}_{\mathrm{lp}}(\cdot)$ retains frequencies below 0.5 Hz. Both operators are implemented as fixed, differentiable frequency-domain masks. The operator $\mathcal{M}(\cdot)$ denotes a nine-frame moving average

with boundary padding. Accordingly, $\mathbf{g}_{\text{band}}$ describes the local magnitude of the physiological-band component, while $\mathbf{g}_{\text{drift}}$ captures slowly varying components. For $\mathbf{g}_{\text{dev}}$, $h_q(\cdot)$ maps $\mathbf{H}$ to position-wise modulation coefficients through a sigmoid, which adaptively weight the deviation of the coarse signal from its local temporal mean.

The coarse rPPG signal and the four descriptors are concatenated along the feature dimension:

$$\mathbf{G}_r = \text{Concate}\left(\hat{\mathbf{s}}^c, \mathbf{g}_{\text{diff}}, \mathbf{g}_{\text{band}}, \mathbf{g}_{\text{drift}}, \mathbf{g}_{\text{dev}}\right). \tag{5}$$

A one-dimensional temporal convolution maps the concatenated descriptors to the feature dimension of $\mathbf{H}$:

$$\mathbf{R} = h_{\text{rel}}\left(\mathbf{G}_r\right). \tag{6}$$

The resulting sequence $\mathbf{R} \in \mathbb{R}^{T\times d}$ summarizes local temporal characteristics related to the reliability of the coarse prediction and serves as time-resolved physiological reliability evidence.

3.2 VLM-Based Visual Interference Perception

While $\mathbf{R}$ captures time-resolved signal characteristics related to coarse-prediction reliability, it does not explicitly model visual interference in the clip. As shown in Fig. 2(b), the VLM-based branch therefore extracts clip-level visual cues related to rPPG interference from sampled facial frames. A fixed interference-oriented prompt directs the VLM to assess illumination fluctuation, head motion, facial blur, and ROI instability. The same prompt is used for all clips and datasets.

To cover the visual conditions throughout the clip while controlling the computational cost, we uniformly sample N frames:

$$\begin{aligned} \mathcal{S}_N(\mathbf{V}) &= (\mathbf{I}_{\tau_1}, \ldots, \mathbf{I}_{\tau_N}), \\ \tau_i &= 1 + \left\lfloor \frac{(i-1)(T-1)}{N-1} \right\rfloor, \end{aligned} \tag{7}$$

where $2 \le N \le T$. The first and last frames are always selected, and the remaining sampling positions are distributed across the clip while preserving the original temporal order.

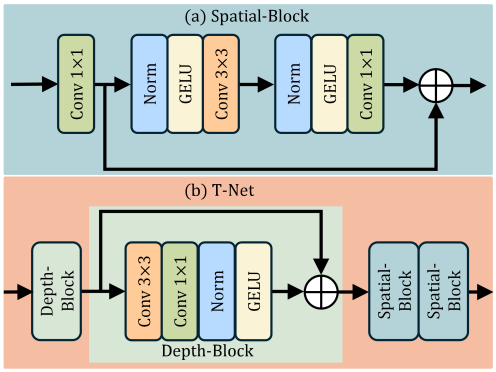


Figure 3: Principal components of the physiological backbone. (a) A Spatial-Block uses a residual bottleneck composed of pointwise and spatial convolutions. (b) The temporal network (T-Net) contains two Depth-Blocks followed by two Spatial-Blocks.

The sampled frames are jointly processed with the fixed prompt p by the frozen VLM:

$$\begin{aligned} \mathbf{Y}_v &= \Phi_{\text{VLM}}\left(\mathcal{S}_N(\mathbf{V}), p\right), \\ \mathbf{z}_v &= \mathbf{Y}_{v,\ell^\star}. \end{aligned} \tag{8}$$

The sampled frames are arranged in temporal order and jointly provided to the VLM together with the prompt. We use the VLM as the feature encoder and instantiate Φ_{VLM} with Qwen2-VL-2B-Instruct (Wang et al., 2024), whose parameters remain frozen throughout training. The matrix $\mathbf{Y}_v \in \mathbb{R}^{L_v\times d_v}$ contains the final-layer hidden states of the valid input sequence, where L_v is the valid sequence length and $\ell^\star$ denotes the final non-padding token.

A trainable evidence head maps $\mathbf{z}_v$ to visual interference evidence:

$$\mathbf{e}_v = \tanh\left(\mathbf{W}_{e,2}\delta\left(\mathbf{W}_{e,1}\,\text{LN}(\mathbf{z}_v) + \mathbf{b}_{e,1}\right) + \mathbf{b}_{e,2}\right). \tag{9}$$

Here, $\mathbf{W}_{e,1}$ and $\mathbf{W}_{e,2}$ are learnable weight matrices, while $\mathbf{b}_{e,1}$ and $\mathbf{b}_{e,2}$ are the corresponding bias vectors. $\text{LN}(\cdot)$ denotes layer normalization, and $\delta(\cdot)$ denotes the GELU activation. The vector $\mathbf{e}_v \in (-1,1)^{d_e}$ summarizes clip-level visual cues related to illumination variation, head motion, facial blur, and ROI instability, and provides visual interference evidence for subsequent temporal feature refinement.

3.3 CONTEXT-GUIDED TEMPORAL FEATURE REFINEMENT

As shown in Fig. 2(c), PhysVR integrates the physiological reliability evidence $\mathbf{R}$ and visual interference evidence $\mathbf{e}_v$ with the global temporal features $\mathbf{H}$ through temporal cross-attention. $\mathbf{H}$ provides the queries, while the time-resolved $\mathbf{R}$ and clip-level $\mathbf{e}_v$ are projected into the key and value spaces. For $\xi \in \{K, V\}$, the context construction is formulated as:

$$\begin{aligned} \mathbf{E}_\xi &= \mathrm{Concate}\left(\mathbf{R}\mathbf{W}_r^\xi, \mathbf{e}_v^\top \mathbf{W}_e^\xi\right), \\ \mathbf{C} &= \mathrm{MHA}\left(\mathbf{H}\mathbf{W}_q, \mathbf{E}_K, \mathbf{E}_V\right)\mathbf{W}_o. \end{aligned} \tag{10}$$

Here, $\mathbf{W}_q, \mathbf{W}_r^\xi \in \mathbb{R}^{d\times d_a}$, $\mathbf{W}_e^\xi \in \mathbb{R}^{d_e\times d_a}$, and $\mathbf{W}_o \in \mathbb{R}^{d_a\times d}$ are learnable projection matrices. The key and value sequences $\mathbf{E}_K, \mathbf{E}_V \in \mathbb{R}^{(T+1)\times d_a}$ consist of T physiological evidence vectors and one clip-level visual evidence vector, yielding the temporal context $\mathbf{C} \in \mathbb{R}^{T\times d}$. Although $\mathbf{e}_v$ is clip-level, its contribution is modulated by the position-dependent queries from $\mathbf{H}$, allowing the visual interference evidence to affect different temporal positions adaptively.

The temporal context is concatenated with the global temporal features and projected to dimension d. A shared temporal correction unit then performs general refinement:

$$\begin{aligned} \mathbf{\Delta}_{\mathrm{shr}} &= R_{\mathrm{shr}}\left(\mathrm{Concate}\left(\mathbf{H}, \mathbf{C}\right)\right), \\ \mathbf{U} &= \mathbf{H} + \lambda_{\mathrm{shr}} \tanh\left(\mathbf{\Delta}_{\mathrm{shr}}\right). \end{aligned} \tag{11}$$

The shared temporal correction unit $R_{\mathrm{shr}}(\cdot)$ first projects the concatenated features from $2d$ to d, followed by three residual Shared Correction Blocks (SCBs) and a pointwise projection. Each SCB contains layer normalization, a depthwise temporal convolution, and two pointwise projections with GELU activation. The temporal convolutions use a kernel size of 5 and dilation rates of 1, 2, and 4, respectively, enabling multi-scale temporal correction. The resulting correction $\mathbf{\Delta}_{\mathrm{shr}} \in \mathbb{R}^{T\times d}$ is scaled and added to $\mathbf{H}$ to obtain $\mathbf{U} \in \mathbb{R}^{T\times d}$.

The shared unit captures temporal deviations common to different interference conditions. To further suppress specific types of interference, we introduce four experts for illumination variation, head motion, facial blur, and ROI instability:

$$\mathcal{G} = \{\mathrm{ill}, \mathrm{mot}, \mathrm{blur}, \mathrm{roi}\}. \tag{12}$$

A temporal router independently controls the activation of the four experts at each temporal position:

$$\begin{aligned} \mathbf{P} &= \sigma\left(h_{\mathrm{rt}}(\mathbf{U})\right), \\ \mathbf{M}_{\mathrm{rt}} &= \mathrm{Gate}(\mathbf{P}) \\ &= \left[\mathbf{m}_{\mathrm{ill}}, \mathbf{m}_{\mathrm{mot}}, \mathbf{m}_{\mathrm{blur}}, \mathbf{m}_{\mathrm{roi}}\right]. \end{aligned} \tag{13}$$

Here, $\mathbf{P} \in (0,1)^{T\times 4}$ contains the position-wise activation probabilities, and each $\mathbf{m}_g \in \{0,1\}^T$ denotes the binary temporal gate of expert g. The four expert gates are independently predicted, allowing multiple experts to be activated simultaneously when different interference factors coexist. The binary gates use a threshold of 0.5, with a straight-through estimator adopted during training.

For each expert, G_g maps $\mathbf{e}_v$ to channel-wise modulation parameters that modulate $\mathbf{U}$ before expert-specific correction:

$$\begin{aligned} \boldsymbol{\gamma}_g, \boldsymbol{\beta}_g &= G_g(\mathbf{e}_v), \\ \mathbf{U}_g &= \mathrm{LN}(\mathbf{U}) \odot \left(1 + \frac{1}{2}\tanh(\boldsymbol{\gamma}_g)\right) + \tanh(\boldsymbol{\beta}_g), \\ \mathbf{\Delta}_g &= \lambda_g \tanh\left(R_g(\mathbf{U}_g)\right), \qquad g \in \mathcal{G}. \end{aligned} \tag{14}$$

Here, G_g and R_g denote the learnable expert-specific modulation and correction mappings. Specifically, G_g generates the channel-wise parameters $\boldsymbol{\gamma}_g, \boldsymbol{\beta}_g \in \mathbb{R}^d$ from $\mathbf{e}_v$, while R_g transforms the modulated features $\mathbf{U}_g$ into the residual correction $\mathbf{\Delta}_g \in \mathbb{R}^{T\times d}$. The four experts use distinct transformations tailored to different interference characteristics. The illumination expert performs channel-wise correction using pointwise projections, whereas the motion expert combines first-order temporal differences with local temporal convolution to capture abrupt variations. The blur expert

Table 1: Intra-dataset comparison of remote HR estimation on benchmark datasets. Best results are shown in bold with light-lavender shading, and second-best results with light-apricot shading.

Method	NIRP-IND			NIRP-DRV			VIPL-HR			BUAA-MIHR			MMPD		
	MAE↓	RMSE↓	ρ ↑	MAE↓	RMSE↓	ρ ↑	MAE↓	RMSE↓	ρ ↑	MAE↓	RMSE↓	ρ ↑	MAE↓	RMSE↓	ρ ↑
POS (Wang et al., 2016)	5.52	6.85	0.40	12.75	15.36	0.34	11.50	17.20	0.30	5.04	7.12	0.63	12.36	17.71	0.18
CHROM (De Haan & Jeanne, 2013)	6.84	8.41	0.32	14.52	17.41	0.18	11.07	17.99	0.27	6.09	8.29	0.51	13.66	18.76	0.08
LGI (Pilz et al., 2018)	8.95	11.01	0.39	12.90	15.51	0.31	12.84	19.02	0.29	6.97	11.33	0.42	17.08	23.32	0.04
DeepPhys (Chen & McDuff, 2018)	3.11	4.44	0.74	13.22	18.39	0.43	11.00	13.80	0.11	4.78	6.74	0.69	22.27	28.92	-0.03
NEST (Lu et al., 2023)	1.08	2.26	0.89	3.61	7.32	0.82	5.52	7.96	0.80	2.88	4.69	0.89	–	–	–
ND-DeeprPPG (Liu & Yuen, 2023)	0.66	1.45	0.89	3.47	6.54	0.85	5.15	7.52	0.78	0.58	1.81	0.95	–	–	–
EfficientPhys (Liu et al., 2023a)	1.37	4.81	0.81	3.67	12.28	0.81	5.23	8.25	0.72	1.43	4.98	0.93	13.47	21.32	0.21
PhysFormer++ (Yu et al., 2023)	0.64	1.39	0.89	3.56	7.59	0.83	4.88	7.62	0.80	0.93	1.66	0.91	–	–	–
RhythmFormer (Zou et al., 2025a)	0.52	1.10	0.91	3.44	6.72	0.82	4.75	7.49	0.82	0.67	1.57	0.94	4.69	11.31	0.60
PhysDiff (Qian et al., 2025)	0.50	1.06	0.91	3.24	6.26	0.83	3.92	6.65	0.85	0.82	1.71	0.93	7.17	9.63	0.78
PHASE-Net (Zhao et al., 2026b)	0.48	0.98	0.91	3.18	6.23	0.86	3.87	6.37	0.86	0.63	1.43	0.94	4.78	8.22	0.71
PhysLLM (Xie et al., 2026)	0.57	1.27	0.90	2.95	6.13	0.86	4.24	6.81	0.85	0.78	1.62	0.94	4.36	10.76	0.65
PhysVR (Ours)	**0.38**	**0.85**	**0.92**	**2.56**	**5.85**	**0.87**	**3.59**	**5.68**	**0.88**	**0.42**	**1.09**	**0.96**	**3.92**	**8.03**	**0.80**

models local deviations from neighboring features, while the ROI-instability expert uses dilated temporal convolutions to capture longer-range temporal inconsistency.

The activated expert corrections are added to $\mathbf{U}$:

$$\begin{aligned} \widetilde{\mathbf{H}} &= \mathbf{U} + \sum_{g \in \mathcal{G}} \mathbf{m}_g \odot \mathbf{\Delta}_g, \\ \hat{\mathbf{s}} &= D_\theta\left(\widetilde{\mathbf{H}}\right). \end{aligned} \tag{15}$$

The decoder $D_\theta(\cdot)$ maps the refined temporal features $\widetilde{\mathbf{H}} \in \mathbb{R}^{T \times d}$ to the final rPPG signal $\hat{\mathbf{s}} \in \mathbb{R}^T$.

3.4 Training Objective

PhysVR follows a two-stage training scheme. For a predicted rPPG signal $\mathbf{y}$ and its corresponding ground-truth signal $\mathbf{s}$, we define:

$$\mathcal{L}_{\text{sig}}(\mathbf{y}, \mathbf{s}) = \lambda_{\text{mse}}\mathcal{L}_{\text{mse}} + \lambda_{\text{np}}\mathcal{L}_{\text{np}} + \lambda_{\text{freq}}\mathcal{L}_{\text{freq}}, \tag{16}$$

where $\mathcal{L}_{\text{np}} = 1 - \rho_{\text{sig}}(\mathbf{y}, \mathbf{s})$ and $\mathcal{L}_{\text{freq}}$ measures the Smooth L_1 distance between their normalized power spectra. The physiological backbone and coarse prediction head are pretrained with $\mathcal{L}_{\text{sig}}(\hat{\mathbf{s}}^c, \mathbf{s})$ and then frozen. The refinement-related modules are trained with:

$$\mathcal{L} = \mathcal{L}_{\text{sig}}(\hat{\mathbf{s}}, \mathbf{s}). \tag{17}$$

4 Experiment

4.1 Datasets and Performance Metrics

We evaluate PhysVR on public benchmark datasets: BUAA-MIHR (Xi et al., 2020), VIPL-HR (Niu et al., 2018; 2019), NIRP-DRV (Nowara et al., 2020), NIRP-IND (Magdalena Nowara et al., 2018), and MMPD (Tang et al., 2023). Following the standard evaluation protocol (Lu et al., 2021), we report mean absolute error (MAE), root mean square error (RMSE), and Pearson's correlation coefficient (ρ). MAE and RMSE are measured in beats per minute (bpm), with lower values indicating better performance, whereas a higher ρ is preferred. More details are in Appendices A and B.

4.2 Intra-dataset Evaluation

Table 1 summarizes the intra-dataset results on five benchmarks. PhysVR consistently achieves the best performance across all datasets, with lower MAE/RMSE and higher correlation coefficients. On NIRP-IND and NIRP-DRV, PhysVR obtains MAEs of 0.38 bpm and 2.56 bpm, outperforming previous methods under both indoor and driving scenarios. It also achieves best performance on VIPL-HR, BUAA-MIHR, and MMPD, with MAEs of 3.59 bpm, 0.42 bpm, and 3.92 bpm, respectively. These consistent improvements demonstrate the effectiveness of PhysVR in suppressing visual interference and improving robust rPPG estimation under diverse conditions.

Table 2: Cross-dataset comparison of remote HR estimation. Best results are shown in bold with light-lavender shading, and second-best results with light-apricot shading.

Method	NIRP → VIPL-HR			VIPL-HR → NIRP-IND			VIPL-HR → NIRP-DRV		
	MAE↓	RMSE↓	ρ ↑	MAE↓	RMSE↓	ρ ↑	MAE↓	RMSE↓	ρ ↑
DeepPhys (Chen & McDuff, 2018)	15.53	17.48	0.41	6.58	9.16	0.52	10.51	12.71	0.56
NEST (Lu et al., 2023)	5.15	8.78	0.72	2.78	4.63	0.78	6.27	8.45	0.65
ND-DeeprPPG (Liu & Yuen, 2023)	5.08	7.92	0.75	1.74	3.28	0.81	6.18	7.75	0.68
EfficientPhys (Liu et al., 2023a)	5.25	11.81	0.71	1.80	6.19	0.76	6.07	11.53	0.64
PhysFormer++ (Yu et al., 2023)	5.09	8.97	0.71	1.79	3.77	0.81	5.78	8.65	0.66
RhythmFormer (Zou et al., 2025a)	4.83	8.05	0.74	1.70	3.43	0.83	5.40	7.93	0.71
PhysDiff (Qian et al., 2025)	4.34	8.12	0.74	1.62	3.35	0.85	5.46	7.96	0.72
PHASE-Net (Zhao et al., 2026b)	4.27	7.96	0.75	1.53	3.22	0.86	5.39	7.71	0.72
PhysLLM (Xie et al., 2026)	4.51	8.37	0.73	1.69	3.49	0.85	5.56	8.15	0.71
PhysVR (Ours)	**3.92**	**7.23**	**0.77**	**1.34**	**3.07**	**0.87**	**5.14**	**7.21**	**0.74**

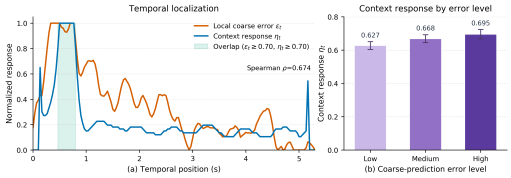


Figure 4: Temporal correspondence between the context response and coarse-prediction errors.

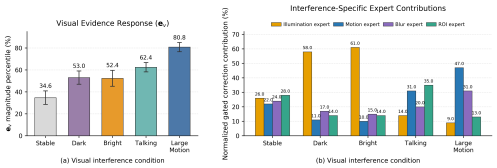


Figure 5: Expert responses to different visual interference conditions.

4.3 Cross-Dataset Evaluation

Table 2 summarizes the cross-dataset evaluation between NIRP and VIPL-HR. PhysVR achieves the best performance across all three transfer settings, demonstrating strong cross-domain generalization. When trained on NIRP and tested on VIPL-HR, PhysVR obtains an MAE of 3.92 bpm and an RMSE of 7.23 bpm. It also achieves competitive performance when transferring from VIPL-HR to NIRP-IND and NIRP-DRV, with MAEs of 1.34 bpm and 5.14 bpm, respectively. These results indicate that PhysVR can effectively mitigate domain variations and maintain robust rPPG estimation across diverse acquisition conditions. Additional qualitative examples of predicted versus ground-truth rPPG signals, together with model complexity, are provided in Appendices D and E.

4.4 Ablation Study

Temporal Response of the Interference-Aware Context. We measure local coarse-prediction error by the normalized deviation from the ground truth and context response by the normalized magnitude of the temporal context. Fig. 4(a) shows that the two temporal sequences exhibit a clear correspondence, achieving a Spearman correlation of 0.674, with larger context responses generally occurring at positions with higher prediction errors. This relationship is also observed at the clip level, where the average context response increases from 0.627 to 0.695 across groups with increasing coarse-prediction RMSE (Fig. 4(b)). These results suggest that the constructed context adaptively emphasizes less reliable coarse predictions, providing targeted guidance for subsequent temporal feature refinement.

Analysis of Visual Evidence and Interference-Specific Experts. Fig. 5 examines the visual evidence and expert responses on VIPL-HR. The magnitude of $\mathbf{e}_v$ generally increases from stable conditions to more severe interference scenarios, suggesting that the learned visual representation is sensitive to clip-level visual disturbances relevant to rPPG. Expert responses also vary with the visual condition: the illumination expert responds most strongly to illumination changes, the motion expert responds most strongly to large motion, while the ROI-instability, motion, and blur experts contribute jointly under talking, where multiple disturbances may occur simultaneously. These observations support a condition-dependent division of refinement roles, rather than a one-to-one assignment between experts and interference types. The optimization pathway underlying these responses is analyzed in Appendix F.

Table 3: Ablation of physiological and visual evidence for temporal feature refinement on NIRP-DRV.

Configuration	Refinement	$\mathbf{R}$	$\mathbf{e}_v$	MAE ↓	RMSE ↓
Coarse prediction	✗	✗	✗	3.79	6.83
Feature-only refinement	✓	✗	✗	3.67	6.72
+ Physiological evidence	✓	✓	✗	3.26	6.31
+ Visual evidence	✓	✗	✓	3.15	6.27
Full model	✓	✓	✓	**2.56**	**5.85**

Table 4: Comparison of visual encoders, prompt designs, and VLM backbones on VIPL-HR.

Encoder	Prompt	MAE ↓	RMSE ↓
DINOv3 (Siméoni et al., 2025)	✗	4.49	6.65
PE (Bolya et al., 2026)	✗	4.27	6.48
Qwen2-VL-2B (Wang et al., 2024)	Generic	3.83	5.92
Qwen2-VL-2B (Wang et al., 2024)	Interf.	3.59	5.68
InternVL2.5-2B (Chen et al., 2024)	Interf.	3.69	5.85
SmolVLM2-2.2B (Marafioti et al., 2025)	Interf.	3.51	5.63
Qwen2-VL-7B (Wang et al., 2024)	Interf.	**3.47**	**5.61**

Ablation of Physiological and Visual Evidence. Table 3 evaluates the contributions of physiological reliability evidence $\mathbf{R}$ and visual interference evidence $\mathbf{e}_v$. Without either evidence source, feature refinement provides only a limited improvement over the coarse prediction, reducing MAE/RMSE from 3.79/6.83 bpm to 3.67/6.72 bpm. Introducing physiological or visual evidence further improves performance, achieving MAE/RMSE values of 3.26/6.31 bpm and 3.15/6.27 bpm, respectively. Combining both evidence sources further reduces the errors to 2.56/5.85 bpm, demonstrating that physiological reliability and visual interference evidence provide complementary guidance for temporal feature refinement.

Effect of Visual Encoder, Prompt Design, and VLM Backbone. Table 4 investigates the impact of visual encoders, prompt designs, and VLM backbones on VIPL-HR. Compared with frozen visual encoders including DINOv3 and PE, the VLM-based evidence branch achieves better performance, indicating that language-guided visual representations provide more effective interference-aware guidance. Using Qwen2-VL-2B, the interference-oriented prompt improves performance over the generic prompt, reducing MAE/RMSE from 3.83/5.92 bpm to 3.59/5.68 bpm. Backbone comparisons show that larger VLMs provide only marginal gains, with Qwen2-VL-7B achieving the best result (3.47 bpm MAE and 5.61 bpm RMSE), suggesting that a compact VLM is sufficient for visual interference perception.

Table 5: Ablation of the context-guided refinement module in PhysVR on VIPL-HR.

Configuration	MAE ↓	RMSE ↓
No feature refinement	4.82	6.87
w/o Shared correction unit	4.75	6.78
w/o Illumination expert	4.37	6.35
w/o Motion expert	4.29	6.27
w/o Blur expert	4.35	6.32
w/o ROI-instability expert	4.41	6.39
w/o Adaptive routing	4.16	6.08
Full model	**3.59**	**5.68**

Ablation of the Context-Guided Refinement Module. Table 5 evaluates the contribution of each component in the context-guided refinement module. Removing the entire refinement module degrades performance, increasing the MAE/RMSE from 3.59/5.68 bpm to 4.82/6.87 bpm, demonstrating the necessity of temporal feature refinement. Component-wise ablations show that the shared correction unit and interference-specific experts provide complementary benefits, while disabling adaptive routing also leads to performance degradation. These results validate the effectiveness of the complete refinement design. Further ablation analyses are provided in Appendix G.

5 Conclusion

In this work, we presented PhysVR, a vision-language model guided interference-aware temporal feature refinement framework for robust rPPG estimation. By integrating physiological reliability evidence and visual interference evidence, PhysVR effectively refines temporal features and suppresses interference before final prediction. Physiological reliability evidence reflects local variations in coarse-prediction, while visual interference evidence characterizes disturbances in the input clip. Extensive experiments on five benchmarks demonstrate consistent improvements in both intra-dataset and cross-dataset settings, while ablations support the effectiveness of the proposed design. Future work will explore lightweight visual encoders and knowledge distillation to further improve inference efficiency while preserving rPPG estimation accuracy.

AI USE STATEMENT

We used generative AI tools to assist with language editing and improve the readability of the manuscript, including grammar correction, wording refinement, and sentence-level polishing. All AI-assisted edits were carefully reviewed by the authors to ensure technical accuracy and consistency with the original research content. The methodology, experimental results, mathematical formulations, and scientific claims presented in this work were verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. *Advances in Neural Information Processing Systems*, 38:60884–60937, 2026.

Constantino Álvarez Casado, Manuel Lage Cañellas, and Miguel Bordallo López. Depression recognition using remote photoplethysmography from facial videos. *IEEE Transactions on Affective Computing*, 14(4):3305–3316, 2023.

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

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Tianyang Dai, Ming Chang, Yan Chen, and Yang Hu. rPPG-VQA: A video quality assessment framework for unsupervised rPPG training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1365–1375, June 2026.

Gerard De Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE transactions on biomedical engineering*, 60(10):2878–2886, 2013.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Zixu Li, Hang Shao, Lei Luo, Jian Yang, et al. Physflow: Frequency decoupled with dual-field rectified flow for remote photoplethysmography. *arXiv preprint arXiv:2606.23226*, 2026.

Qian Liang, Yan Chen, and Yang Hu. Continual learning for remote physiological measurement: Minimize forgetting and simplify inference. In *European conference on computer vision*, pp. 126–144. Springer, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Si-Qi Liu and Pong C Yuen. Robust remote photoplethysmography estimation with environmental noise disentanglement. *IEEE Transactions on Image Processing*, 33:27–41, 2023.

Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020.

Xin Liu, Yuntao Wang, Sinan Xie, Xiaoyu Zhang, Zixian Ma, Daniel McDuff, and Shwetak Patel. Mobilephys: Personalized mobile camera-based contactless physiological sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–23, 2022.

Xin Liu, Brian Hill, Ziheng Jiang, Shwetak Patel, and Daniel McDuff. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5008–5017, 2023a.

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

Xin Liu, Yuting Zhang, Zitong Yu, Hao Lu, Huanjing Yue, and Jingyu Yang. rppg-mae: Self-supervised pretraining with masked autoencoders for remote physiological measurements. *IEEE Transactions on Multimedia*, 26:7278–7293, 2024b.

Hao Lu, Hu Han, and S Kevin Zhou. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12404–12413, 2021.

Hao Lu, Zitong Yu, Xuesong Niu, and Ying-Cong Chen. Neuron structure modeling for generalizable remote physiological measurement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18589–18599, 2023.

Ewa Magdalena Nowara, Tim K Marks, Hassan Mansour, and Ashok Veeraraghavan. Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1272–1281, 2018.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Daniel McDuff. Camera measurement of physiological vital signs. *ACM Computing Surveys*, 55 (9):1–40, 2023.

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

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, 2019.

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, 2020.

Christian S Pilz, Sebastian Zaunseder, Jarek Krajewski, and Vladimir Blazek. Local group invariance for heart rate estimation from face videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1254–1262, 2018.

Wei Qian, Gaoji Su, Dan Guo, Jinxing Zhou, Xiaobai Li, Bin Hu, Shengeng Tang, and Meng Wang. Physdiff: physiology-based dynamicity disentangled diffusion model for remote physiological measurement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6568–6576, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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 Computer Vision and Pattern Recognition Conference*, pp. 10858–10867, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Rencheng Song, Huan Chen, Juan Cheng, Chang Li, Yu Liu, and Xun Chen. Pulsegan: Learning to generate realistic pulse waveforms in remote photoplethysmography. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1373–1384, 2021.

Jiankai Tang, Kequan Chen, Yuntao Wang, Yuanchun Shi, Shwetak Patel, Daniel McDuff, and Xin Liu. 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.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wenjin Wang, Sander Stuijk, and Gerard De Haan. A novel algorithm for remote photoplethysmography: Spatial subspace rotation. *IEEE transactions on biomedical engineering*, 63(9):1974–1984, 2015.

Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016.

Lin Xi, Weihai Chen, Changchen Zhao, Xingming Wu, and Jianhua Wang. 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.

Yiping Xie, Mingtong Dai, Jian-Ping Zhou, Yue Sun, Tao Tan, Weicheng Xie, Linlin Shen, Zitong YU, et al. Physllm: Harnessing large language models for cross-modal remote physiological sensing. In *International Conference on Learning Representations*, volume 2026, pp. 93760–93779, 2026.

Xing Yao, Rencheng Song, Juan Cheng, Chang Li, and Xun Chen. Hrvfusion: Video-based long-term heart rate variability measurement with conditional diffusion models. *IEEE Transactions on Multimedia*, 2026.

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

Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip HS Torr, and Guoying Zhao. 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.

Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Yawen Cui, Jiehua Zhang, Philip Torr, and Guoying Zhao. Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. *International Journal of Computer Vision*, 131(6):1307–1330, 2023.

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(7):4112–4133, 2025.

Xinyu Zhang, Weiyu Sun, Hao Lu, Ying Chen, Yun Ge, Xiaolin Huang, Jie Yuan, and Yingcong Chen. Self-similarity prior distillation for unsupervised remote physiological measurement. *IEEE Transactions on Multimedia*, 26:10290–10305, 2024.

Bo Zhao, Junzhe Cao, Dan Guo, Dongmin Huang, Wenjin Wang, Tao Tan, Yue Sun, and Zitong Yu. Flow: Optimal transport-driven feature warping for generalized remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28481–28491, 2026a.

Bo Zhao, Dan Guo, Junzhe Cao, Yong Xu, Bochao Zou, Tao Tan, Yue Sun, and Zitong Yu. Phasenet: Physics-grounded harmonic attention system for efficient remote photoplethysmography measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21198–21207, 2026b.

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, 2025a.

Bochao Zou, Zizheng Guo, Xiaocheng Hu, and Huimin Ma. Rhythmmamba: Fast, lightweight, and accurate remote physiological measurement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11077–11085, 2025b.

A Datasets

BUAA-MIHR (Xi et al., 2020) is designed for rPPG evaluation under low-illumination conditions. It contains 165 RGB facial videos from 15 subjects recorded at 11 illumination levels ranging from 1.0 to 100.0 lux. Each 60-s video was captured at 30 fps with synchronized finger-clip PPG recorded at 60 Hz. Following prior work, we use recordings with illumination intensities of at least 6.3 lux, as lower-illumination videos are severely underexposed and typically require dedicated low-light processing beyond the scope of standard rPPG estimation.

VIPL-HR (Niu et al., 2018; 2019) contains 3,130 facial videos from 107 subjects collected under less-constrained conditions. The dataset covers substantial variations in illumination, head motion, facial appearance, and acquisition devices, making it challenging for robust remote physiological measurement. Each video lasts approximately 30 seconds, with synchronized finger BVP signals provided as physiological reference.

NIRP-DRV (Nowara et al., 2020) contains 190 RGB and near-infrared facial recordings from 19 subjects collected in realistic driving environments. The recordings involve diverse in-vehicle illumination conditions, natural head movements, and other driving-related visual disturbances, providing a challenging benchmark for evaluating rPPG robustness in practical scenarios. Synchronized pulse-oximeter measurements are provided as physiological reference.

NIRP-IND (Magdalena Nowara et al., 2018) provides a complementary indoor setting with 15 RGB and near-infrared facial recordings from 8 subjects. The dataset includes both stationary and motion conditions under controlled indoor environments, enabling evaluation of physiological measurement under different levels of subject movement. Synchronized pulse-oximeter measurements are provided as physiological reference. Although both RGB and near-infrared recordings are available in NIRP-IND and NIRP-DRV, we use only the RGB modality in all experiments.

MMPD (Tang et al., 2023) is a mobile rPPG dataset containing 660 one-minute RGB facial videos from 33 subjects. It covers Fitzpatrick skin types III–VI, four illumination conditions, and four activities including stationary, head rotation, talking, and walking, resulting in substantial variations in appearance, motion, and recording conditions. Synchronized finger PPG signals are provided as physiological reference, making MMPD suitable for evaluating rPPG methods under diverse mobile and real-world conditions.

B Performance Metrics

Following the standard evaluation protocol (Lu et al., 2021), we use three commonly adopted metrics for heart rate estimation: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and

Pearson's Correlation Coefficient (ρ). Given N test samples, let $\hat{h}_i$ and h_i denote the estimated and ground-truth heart rates of the i-th sample, respectively. MAE and RMSE are defined as:

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{N}\sum_{i=1}^{N}\left|\hat{h}_i - h_i\right|, \\ \mathrm{RMSE} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{h}_i - h_i\right)^2}. \end{aligned} \tag{18}$$

Pearson's correlation coefficient is defined as:

$$\rho = \frac{\sum_{i=1}^{N}\left(\hat{h}_i - \bar{\hat{h}}\right)\left(h_i - \bar{h}\right)}{\sqrt{\sum_{i=1}^{N}\left(\hat{h}_i - \bar{\hat{h}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(h_i - \bar{h}\right)^2}}. \tag{19}$$

Where $\bar{\hat{h}}$ and $\bar{h}$ denote the mean estimated and ground-truth heart rates, respectively. Lower MAE and RMSE indicate smaller estimation errors, while a higher ρ indicates stronger agreement between the estimated and ground-truth heart rates.

C Implementation Details

We implement PhysVR using the rPPG-Toolbox (Liu et al., 2023b) and conduct all experiments on RGB facial videos. Temporal alignment, reference-signal resampling, and HR extraction follow the standard implementation in rPPG-Toolbox. Facial regions are cropped, resized to 128×128, and divided into 160-frame clips. PhysVR follows a two-stage training protocol. In the first stage, the physiological backbone is independently trained to produce coarse rPPG predictions using $\mathcal{L}_{\mathrm{sig}}(\hat{\mathbf{s}}^c, \mathbf{s})$. Once trained, the backbone is frozen, and its global temporal features and coarse prediction are treated as fixed inputs to the refinement branch. In the second stage, optimization is restricted to the physiological reliability mapping, visual evidence head, temporal cross-attention, context-guided refinement module, and final decoder. For the alternative-backbone experiments, each backbone is independently trained on the same data partition and similarly frozen before refinement training. For each clip, eight uniformly sampled frames are processed by the frozen Qwen2-VL-2B-Instruct (Wang et al., 2024) using a fixed interference-oriented prompt. The refinement-related components are jointly optimized for 30 epochs using AdamW with a learning rate of 3×10^{-4}, while both the physiological backbone and Qwen2-VL remain fixed. The shared correction scale λ_{shr} is initialized to 0.10, while all four expert-specific scales are initialized to 0.05. We set $\lambda_{\mathrm{mse}} = 1.00$, $\lambda_{\mathrm{np}} = 0.25$, and $\lambda_{\mathrm{freq}} = 0.05$. All dataset partitions are subject-disjoint. VIPL-HR follows the official five-fold protocol, while BUAA-MIHR, MMPD, and the combined NIRP dataset use a 3:1 train/test split. NIRP-IND and NIRP-DRV are jointly partitioned and evaluated separately. For fair comparison, all deep learning-based baselines are reimplemented using the same preprocessing pipeline and dataset partitions. BUAA-MIHR and MMPD are used only for intra-dataset evaluation. Following prior work (Shao et al., 2025), cross-dataset evaluation is conducted without target-domain training or fine-tuning for NIRP $\rightarrow$ VIPL-HR, VIPL-HR $\rightarrow$ NIRP-IND, and VIPL-HR $\rightarrow$ NIRP-DRV.

Prompt Details: For the prompt-design ablation in Table 4, the generic prompt is "Observe this facial video for remote physiological measurement. Describe its visual content, recording conditions, and overall quality". The interference-oriented prompt is "Observe this facial video for remote physiological measurement. Identify the current recording environment and visual conditions, and focus on rPPG-relevant interference, including illumination fluctuation, head motion, facial blur, and ROI instability".

D Qualitative Comparison With Competing Methods

Fig. 6 provides a qualitative comparison of the predicted rPPG signals and their power spectra on VIPL-HR. For the representative sample, PhysVR estimates an HR of 79.4 bpm, closely matching

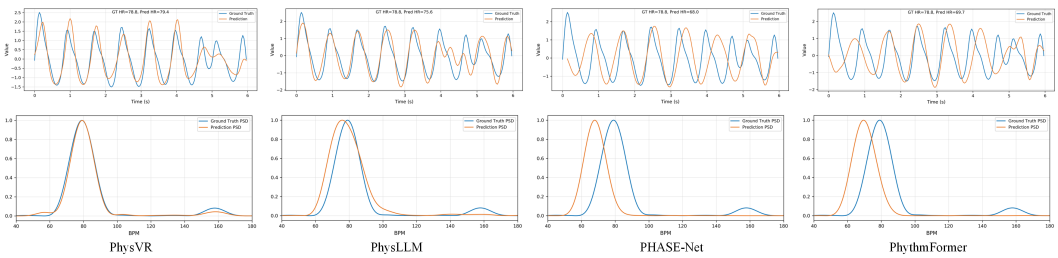


Figure 6: Comparison of rPPG estimation results on VIPL-HR. The top row presents the predicted and ground-truth rPPG signals, while the bottom row shows their corresponding normalized power spectral density distributions. From left to right, the columns correspond to PhysVR, PhysLLM, PHASE-Net, and RhythmFormer.

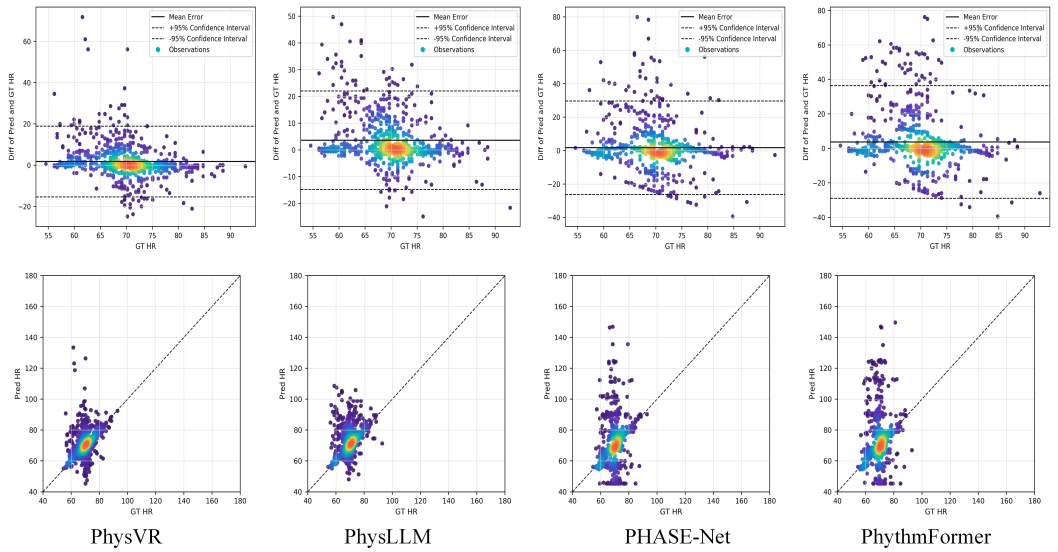


Figure 7: Comparison of HR estimation results on NIRP-DRV. The top row presents the difference between the predicted and ground-truth HR values, while the bottom row shows the corresponding scatter plots of predicted HR against ground-truth HR. The diagonal line represents perfect agreement. From left to right, the columns correspond to PhysVR, PhysLLM, PHASE-Net, and RhythmFormer.

the ground-truth value of 78.8 bpm. Its predicted signal follows the reference periodicity, and the dominant spectral peak is well aligned with the ground-truth HR frequency. In comparison, PhysLLM, PHASE-Net, and RhythmFormer show larger temporal deviations or spectral shifts. Fig. 7 further presents the HR estimation results on NIRP-DRV. PhysVR exhibits a more concentrated error distribution around zero and fewer large deviations, while its predictions are more closely distributed around the identity line. These qualitative results further illustrate the stable estimation performance of PhysVR across different recording conditions.

E Model Complexity Analysis

Fig. 8 further compares HR estimation accuracy and computational complexity on VIPL-HR. PhysVR achieves the lowest RMSE among the compared methods. Although PhysVR employs a frozen 2B-parameter VLM, only eight uniformly sampled frames from each clip are processed by the VLM, and its forward computation is included in the reported MACs. This restricted VLM processing limits the additional computational overhead while retaining the performance gains brought by interference-aware temporal feature refinement.

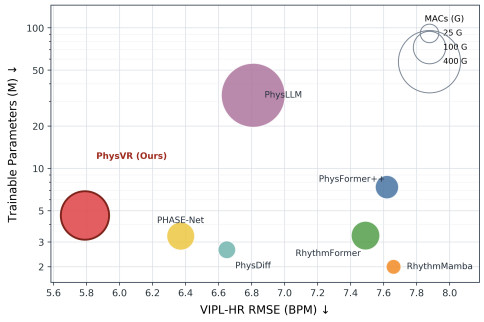


Figure 8: Comparison of HR estimation accuracy and computational complexity on VIPL-HR. The horizontal and vertical axes denote RMSE and trainable parameters, while bubble size represents MACs. PhysVR includes frozen VLM inference on eight sampled frames.

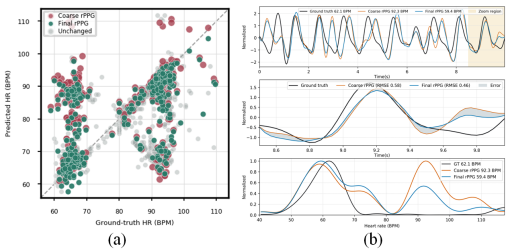


Figure 9: Effect of temporal feature refinement on BUAA-MIHR. (a) Coarse and final HR estimates, with gray markers indicating unchanged estimates. (b) Temporal and spectral comparisons of ground-truth, coarse, and final rPPG signals.

Algorithm 1 Refinement training procedure of PhysVR

Input: facial videos $\mathbf{V}$, ground-truth rPPG signals $\mathbf{s}$, fixed prompt p, and pretrained physiological backbone B_{ϕ^*}
Output: optimized refinement parameters Θ_{R}
Freeze B_{ϕ^*} and Φ_{VLM}
for each training minibatch $(\mathbf{V}, \mathbf{s})$ **do**
  $(\mathbf{H}, \hat{\mathbf{s}}^c) \leftarrow \mathrm{sg}(B_{\phi^*}(\mathbf{V}))$ ▷ fixed coarse predictor
  Construct physiological reliability evidence $\mathbf{R}$ from $\hat{\mathbf{s}}^c$ and $\mathbf{H}$
  Sample $\mathcal{S}_N(\mathbf{V})$ and obtain visual evidence $\mathbf{e}_v$ through the frozen VLM and trainable evidence head
  Construct interference-aware context $\mathbf{C}$ from $\mathbf{H}$, $\mathbf{R}$, and $\mathbf{e}_v$
  Compute shared refinement $\mathbf{U}$ from $\mathbf{H}$ and $\mathbf{C}$
  $\mathbf{P} \leftarrow \sigma(h_{\mathrm{rt}}(\mathbf{U}))$, $\quad \mathbf{M}_{\mathrm{rt}} \leftarrow \mathrm{Gate}(\mathbf{P})$ ▷ temporal routing
  for $g \in \mathcal{G}$ **do**
    $(\boldsymbol{\gamma}_g, \boldsymbol{\beta}_g) \leftarrow G_g(\mathbf{e}_v)$ ▷ evidence conditioning
    Compute evidence-conditioned feature $\mathbf{U}_g$ and expert correction $\boldsymbol{\Delta}_g$
  end for
  $\mathbf{H}^e \leftarrow \mathbf{U} + \sum_{g \in \mathcal{G}} \mathbf{m}_g \odot \boldsymbol{\Delta}_g$
  $\hat{\mathbf{s}} \leftarrow D_\theta(\mathbf{H}^e)$ ▷ final prediction
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{sig}}(\hat{\mathbf{s}}, \mathbf{s})$
  Update only Θ_{R} by back-propagating $\mathcal{L}$
end for

F Gradient Analysis of Evidence Optimazation and Expert Routing without Interference Annotations

PhysVR is trained without explicit annotations of illumination variation, head motion, facial blur, ROI instability, or expert selection. The key question is therefore how the visual evidence head, temporal router, and interference-specific experts can be optimized without such direct supervision. Although no auxiliary supervision is introduced for these components, they all participate in the refinement of temporal features used for the final rPPG prediction and are therefore optimized through the final rPPG objective. In the following, we analyze how this task supervision reaches the visual evidence head, temporal router, and interference-specific experts, explaining how these refinement components can be optimized without explicit interference or expert-selection annotations. For clarity, we denote the final prediction loss $\mathcal{L}_{\mathrm{sig}}(\hat{\mathbf{s}}, \mathbf{s})$ by $\mathcal{L}_f$ in the following derivation. Algorithm 1 summarizes the corresponding training procedure.

F.1 Gradient Path to the Visual Evidence Head

The evidence head in Eq. 9 maps the frozen VLM representation $\mathbf{z}_v$ to the visual interference evidence $\mathbf{e}_v$. No direct target is assigned to $\mathbf{e}_v$. Its supervision instead arises from its effect on the refined temporal feature. Following Eq. 15, this feature is written as:

$$\mathbf{H}^e = \mathbf{U} + \sum_{g \in \mathcal{G}} \mathbf{m}_g \odot \mathbf{\Delta}_g. \tag{20}$$

The shared feature $\mathbf{U}$ depends on $\mathbf{e}_v$ through the temporal context. The routing gates $\mathbf{m}_g$ depend on $\mathbf{e}_v$ through the context-guided shared feature $\mathbf{U}$, while the expert corrections $\mathbf{\Delta}_g$ depend on $\mathbf{e}_v$ through both $\mathbf{U}$ and the expert-specific modulation. Therefore, the total derivative of the refined feature with respect to the visual evidence is:

$$\frac{\mathrm{d}\mathbf{H}^e}{\mathrm{d}\mathbf{e}_v} = \frac{\mathrm{d}\mathbf{U}}{\mathrm{d}\mathbf{e}_v} + \sum_{g \in \mathcal{G}} \left[\frac{\mathrm{d}\mathbf{m}_g}{\mathrm{d}\mathbf{e}_v} \odot \mathbf{\Delta}_g + \mathbf{m}_g \odot \frac{\mathrm{d}\mathbf{\Delta}_g}{\mathrm{d}\mathbf{e}_v} \right]. \tag{21}$$

The three terms describe the gradient paths through the context-guided shared refinement, adaptive routing, and evidence-conditioned expert correction, respectively. The derivative through the binary routing gates is evaluated using the straight-through estimator introduced in the next subsection.

Let Θ_e denote the parameters of the evidence head. Applying the chain rule gives:

$$\nabla_{\Theta_e} \mathcal{L}_f = \left(\frac{\partial \mathbf{e}_v}{\partial \Theta_e} \right)^\top \left(\frac{\mathrm{d}\mathbf{H}^e}{\mathrm{d}\mathbf{e}_v} \right)^\top \nabla_{\mathbf{H}^e} \mathcal{L}_f. \tag{22}$$

This expression identifies the gradient paths by which the loss for the final prediction reaches the evidence head. Hence, $\mathbf{e}_v$ can be optimized through the final rPPG objective without requiring a direct interference target.

F.2 Surrogate Gradient for Binary Temporal Routing

We next examine the routing gradient at temporal position t. Let

$$\mathbf{q}_t = \frac{\partial \mathcal{L}_f}{\partial \mathbf{H}_t^e}$$

denote the gradient arriving at the refined feature. Holding $\mathbf{U}_t$ and $\mathbf{\Delta}_{g,t}$ fixed and temporarily treating $m_{g,t}$ as a continuous mixing coefficient, we obtain:

$$\frac{\partial \mathcal{L}_f}{\partial m_{g,t}} = \langle \mathbf{q}_t, \mathbf{\Delta}_{g,t} \rangle. \tag{23}$$

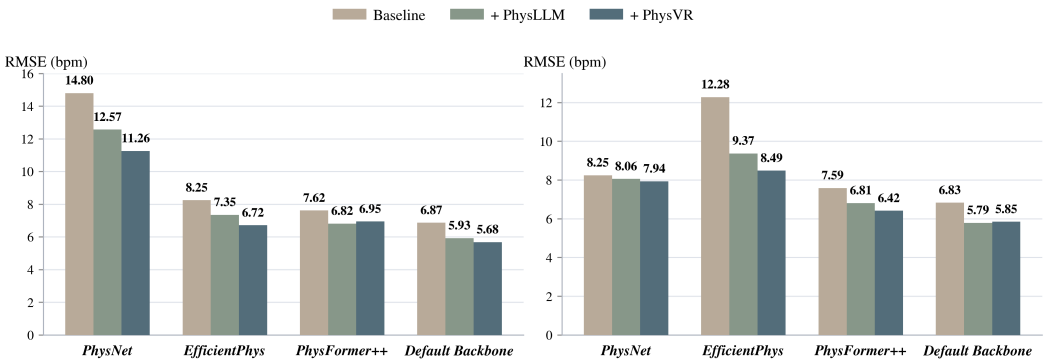
The inner product measures the local effect of changing the contribution of expert g. A negative value indicates that a small increase in its contribution reduces $\mathcal{L}_f$ to first order.

As defined in Eq. 13, the router predicts $p_{g,t} = \sigma(a_{g,t})$ and applies a threshold of 0.5. Since the binary gate is not differentiable, PhysVR uses the straight-through estimator:

$$\begin{aligned} m_{g,t}^{\mathrm{STE}} &= \mathrm{sg}\left(\mathbb{I}[p_{g,t} > 0.5] - p_{g,t}\right) + p_{g,t}, \\ \frac{\widetilde{\partial} \mathcal{L}_f}{\partial a_{g,t}} &= p_{g,t}\left(1 - p_{g,t}\right) \langle \mathbf{q}_t, \mathbf{\Delta}_{g,t} \rangle, \end{aligned} \tag{24}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and $\widetilde{\partial}$ denotes the surrogate derivative. The forward value remains binary, while the backward pass uses the derivative of the routing probability.

For a routing probability that has not saturated, a negative inner product produces a negative surrogate gradient with respect to $a_{g,t}$. Gradient descent therefore increases the corresponding routing logit and its activation tendency. This establishes the local optimization mechanism by which the router favors expert corrections that reduce the final rPPG loss, without requiring an explicit routing target.



(a) VIPL-HR (b) NIRP-DRV

Figure 10: Impact of different rPPG backbones on VIPL-HR and NIRP-DRV in terms of RMSE (bpm).

Table 6: Condition-wise ablation of interference-specific experts on VIPL-HR.

Removed expert	Dark	Bright	Talking	Motion
Illumination	5.02/7.22	4.88/7.05	4.28/6.42	4.76/7.08
Motion	4.15/6.34	3.82/6.03	4.72/6.86	5.63/8.04
Blur	4.17/6.37	3.85/6.06	4.49/6.68	5.24/7.64
ROI-instability	4.20/6.39	3.88/6.10	5.08/7.28	4.97/7.26
Shared correction unit	5.64/7.52	5.14/7.28	5.36/7.55	5.81/8.19
Full model	**3.92/6.05**	**3.61/5.77**	**4.08/6.16**	**4.54/6.83**

F.3 Gradient Allocation to Shared and Expert Refinement

Let Θ_g denote the transformation, modulation, and scaling parameters of expert g. Since the binary gate $m_{g,t}$ multiplies the expert correction in the forward pass, the gradient with respect to the expert parameters is:

$$\nabla_{\Theta_g} \mathcal{L}_f = \sum_t m_{g,t} \left(\frac{\partial \mathbf{\Delta}_{g,t}}{\partial \Theta_g} \right)^\top \mathbf{q}_t. \tag{25}$$

Only temporal positions with $m_{g,t} = 1$ contribute directly to the update of expert g. Consequently, different routing patterns expose different experts to different subsets of temporal prediction errors, allowing them to acquire condition-dependent correction behaviors.

In contrast, the shared feature $\mathbf{U}$ contributes to $\mathbf{H}^e$ at every temporal position without routing selection. The shared correction unit therefore receives an ungated task-gradient path across different visual conditions, whereas each expert is updated only at the temporal positions selected by its gate. This asymmetric gradient allocation encourages the shared unit to capture temporal deviations common to different interference conditions, while the routed experts learn complementary condition-dependent corrections. Together with the task-driven routing mechanism derived above, this provides a gradient-based explanation for the shared–specific refinement behavior without explicit interference annotations or expert-selection targets.

G Additional Ablation Studies

Shared and Interference-Specific Refinement. Table 6 examines the roles of the shared correction unit and interference-specific experts under different visual conditions. Removing the shared correction unit consistently causes substantial performance degradation across all conditions, indicating its general contribution to temporal feature refinement. In contrast, the individual experts exhibit clear condition-dependent effects. The illumination expert contributes most strongly under dark and

Table 7: Sensitivity analysis of physiological reliability parameters on VIPL-HR. The default settings are highlighted in bold.

Parameter	Setting	MAE ↓	RMSE ↓
Band-pass range	0.6–3.0 Hz	3.83	5.92
	0.7–3.0 Hz	**3.59**	**5.68**
	0.7–3.5 Hz	3.74	5.74
Low-pass cutoff	0.4 Hz	3.78	5.83
	0.5 Hz	**3.59**	**5.68**
	0.6 Hz	3.72	5.75

bright conditions, while the motion expert is most influential under large motion, where the blur expert also provides complementary refinement. For talking sequences, the ROI-instability expert plays the most prominent role. These results support the intended shared-specific refinement mechanism: the shared correction unit captures temporal deviations common to different interference conditions, while the interference-specific experts further provide condition-dependent corrections for different types of visual interference.

Effectiveness of Temporal Feature Refinement. To examine whether the temporal features produced by the trained backbone benefit from further refinement, Fig. 9 compares the coarse and final predictions of PhysVR on BUAA-MIHR. As shown in Fig. 9(a), the final HR estimates are closer to the identity line than the coarse estimates. Fig. 9(b) presents a representative example in the temporal and frequency domains. Within the enlarged interval, the local RMSE decreases from 0.58 to 0.46, while the dominant spectral peak shifts from 92.3 to 59.4 bpm, approaching the ground-truth HR of 62.1 bpm. These results show that further temporal feature refinement after backbone feature learning improves rPPG estimation.

Sensitivity to Physiological Reliability Parameters. We further evaluate the sensitivity of the physiological reliability construction to the band-pass range and low-pass cutoff on VIPL-HR. As shown in Table 7, we vary one parameter at a time while keeping the remaining settings fixed. The default configuration achieves the best overall performance, while moderate variations of either parameter result in only limited performance degradation. These results indicate that the physiological reliability construction is robust to reasonable variations in these frequency-related parameters.

Impact of Different Backbones. To assess whether PhysVR depends on a particular coarse predictor, we evaluate it with four pretrained rPPG backbones: PhysNet, EfficientPhys, PhysFormer++, and our default backbone. For each configuration, the backbone is first trained to produce coarse rPPG predictions and pre-head temporal features, and is subsequently kept frozen during refinement training. We also apply PhysLLM to the same predictors under this protocol for comparison. As shown in Fig. 10, PhysVR consistently reduces the RMSE of the unrefined predictors across all eight backbone–dataset configurations. The improvements are particularly pronounced for PhysNet on VIPL-HR and EfficientPhys on NIRP-DRV, with RMSE reductions of 3.54 and 3.79 bpm, respectively. PhysVR also outperforms PhysLLM in six of the eight configurations. PhysLLM performs marginally better with PhysFormer++ on VIPL-HR and the default backbone on NIRP-DRV, but only by 0.13 and 0.06 bpm, respectively. Averaged across all configurations, PhysVR reduces RMSE from 9.06 to 7.41 bpm, compared with 7.84 bpm achieved by PhysLLM. These results demonstrate that PhysVR can effectively refine predictions from diverse pretrained rPPG models without updating their backbone parameters.

H LIMITATIONS

Although PhysVR consistently improves performance across multiple rPPG benchmarks, several limitations remain. The use of a frozen VLM introduces additional inference cost, while clip-level visual evidence from sparsely sampled frames may overlook brief disturbances. Moreover, the predefined physiological frequency priors may limit flexibility across broader physiological ranges. Future work will explore lightweight visual interference modeling, adaptive physiological priors, and validation in more diverse real-world settings.