LEARNING WHAT TO TRUST: RELIABILITY-CONDITIONED MASKED MODELING FOR PHYSIOLOGICAL TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked time-series models usually choose reconstruction targets without considering whether the observations are trustworthy. This is a poor fit for wearable physiological signals, where motion and contact artifacts create short, source-specific failures inside otherwise usable ECG and PPG recordings. We introduce QRAFT, a self-supervised learner in which a weakly supervised local reliability proxy determines which patches remain visible and which become reconstruction targets. Visible tokens are processed by a shared temporal encoder, allowing reliable evidence from either source to support prediction of the withheld segment in the other. A temporal, spectral, and contrastive objective trains the representation. To distinguish reliability-aware localization from generic hard-example mining, we introduce difficulty-matched random masking. At the same 38.5% mask ratio, QRAFT reduces PPG reconstruction RMSE by 23.2% relative to random masking and by 10.0% relative to the difficulty-matched control. The representation yields lower blood-pressure errors across three transfer datasets. On synchronized ECG–PPG data, fusion gains rise from 5.5% when both streams are high quality to 18.6% when both are degraded. A factorial analysis further separates gains due to pretraining from those due to the fusion rule. These results show that local reliability can serve as supervision for masked representation learning, not merely as a filter applied before training.

1 INTRODUCTION

Physiological monitoring is increasingly used for continuous health assessment, but wearable recordings remain vulnerable to motion, contact variation, low perfusion, and electrode displacement (Victor et al., 2025; Gambhir et al., 2021). Practical wearable systems therefore need to preserve useful observations while identifying local failures rather than treating an entire record as reliable or unusable (Chen et al., 2025; Deng et al., 2023). ECG records electrical activation whereas PPG reflects peripheral pulsatile responses, and their quality can change differently within the same recording (Scardulla et al., 2023; Joutsen et al., 2024).

This creates a gap between source-level fusion and local evidence selection. Multiview BP models and hybrid waveform predictors show the value of combining physiological sources, but their fusion decisions are generally made after signal features have been formed (Liu et al., 2024b; Tian et al., 2025). A sample-level weight cannot identify which patches remain informative and which should instead be reconstructed or supported by the synchronized source, even though artifact duration can be shorter than the analysis window.

Self-supervised learning can exploit such recordings without dense clinical labels. Contrastive objectives learn invariance across augmented views, whereas masked reconstruction predicts withheld waveform content (Le-Khac et al., 2020; Krishnan et al., 2022; Ding & Wu, 2024; Zhang et al., 2024). Random targets may be unaffected or redundant while corrupted patches remain visible; record-level filtering discards usable intervals and does not teach local prediction. The learning problem is therefore to select both the evidence entering fusion and the content requested from the model.

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.

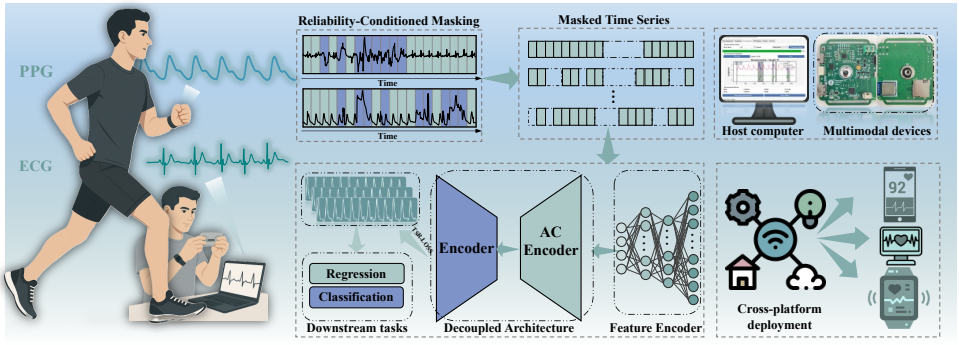


Figure 1: Overview of QRAFT. Fine-grained quality estimates partition each ECG and PPG stream into visible evidence and reconstruction targets. The shared encoder then integrates the available temporal and cross-source context before source-specific decoders predict withheld waveform content.

QRAFT performs these decisions inside masked pretraining. A weakly supervised local reliability proxy produces sub-second scores: reliable patches remain visible, while low-reliability patches become reconstruction targets before shared self-attention. Visible ECG and PPG tokens are processed in one temporal sequence, allowing reliable context from one source to support prediction in the other. A temporal, spectral, and contrastive objective trains waveform fidelity and transferable structure, with reliability acting as a control variable rather than a post hoc confidence score. The resulting evidence partition and shared fusion path are summarized in Fig. 1.

To separate reliability-aware localization from hard-example mining, we use difficulty-matched random masking based on local variance, sample entropy, and high-frequency energy. A factorial comparison separates pretraining from fusion-rule effects, and quality-regime analysis tests whether complementarity changes with source degradation. At the matched 38.5% mask ratio, QRAFT reduces PPG reconstruction RMSE from 0.82 to 0.63 and retains a 10.0% advantage over DM-Rand. It yields lower BP errors on MIMIC-III, UCI, and IHC, while synchronized ECG–PPG fusion gains increase from 5.5% for jointly high-quality streams to 18.6% for jointly degraded streams. These experiments test whether quality can localize targets, improve transfer, and condition source participation. The main contributions are:

- Reliability-conditioned target selection turns sub-second signal quality into an explicit evidence partition: trustworthy patches remain available to the encoder, whereas unreliable patches become reconstruction targets.
- Synchronized ECG–PPG context is integrated before downstream prediction, enabling the shared representation to predict withheld waveform content with support from the complementary source.
- Controlled evaluation separates reliability-aware localization from hard-example mining and isolates the roles of pretraining, cross-source fusion, and source degradation through transfer and reconstruction analyses.

2 Related Work

Self-supervised learning for physiological time series. Contrastive representation learning has become a standard way to exploit augmented views when labels are scarce, whereas masked and generative objectives learn by reconstructing withheld signal content (Le-Khac et al., 2020; Krishnan et al., 2022; Ding & Wu, 2024; Zhang et al., 2024). Physiological adaptations include similarity-based PPG learning for heart-rate estimation and artifact-oriented self-supervision for PPG (Bhattachrjee et al., 2023; Le et al., 2024). Masked autoencoding has also been used for remote physiological measurements and ECG spatiotemporal structure (Liu et al., 2024c; Na et al., 2024); multiview information bottlenecks, masked ECG encoders, and distribution-aware ECG evaluation further target transfer across arrhythmia and related tasks (Yang et al., 2024; Zhang et al., 2023; Soltanieh et al., 2024). These studies establish that morphology, rhythm, and augmented-view agreement can support transfer across downstream tasks, consistent with broader ECG processing and arrhythmia-learning

literature (Ansari et al., 2023; Ardeti et al., 2023). Their masking or augmentation policies, however, are generally independent of local signal quality. A corrupted patch can therefore remain visible while another patch is selected as the target, leaving the pretext task without an explicit distinction between trustworthy evidence and content that should be predicted.

Signal quality assessment and reliability modeling. Signal quality assessment has been studied with handcrafted indices, attractor reconstruction, and robustness analyses for PPG and ECG (Schmith et al., 2023; Pereira et al., 2019; Rahman et al., 2022). Feature selection and learned discriminators improve the ranking of usable segments, but they typically remain external to the representation objective (Mohagheghian et al., 2022; Zhou et al., 2021). Deep PPG and ECG quality models have also explored convolutional, recurrence-based, and attentional designs (Shin, 2022; Roh & Shin, 2021; Jin et al., 2023). QRAFT uses the quality estimate differently: local filtering and artifact-aware preprocessing motivate the reliability signal, but the estimate is placed inside the training loop before encoding (Hao et al., 2024; Liu et al., 2024a; Liang et al., 2018; Liu et al., 2025). It then determines the visible set and reconstruction targets, preserving usable portions of a record while making quality relevant to representation learning even when downstream labels are unavailable.

Multimodal fusion under missing or degraded evidence. Recent multimodal frameworks improve robustness through imputation, dynamic weighting, and self-supervised learning (Wu et al., 2025; Lage Cañellas et al., 2025). These methods address important failures such as missing modalities, source imbalance, and noisy medical data. Their decisions typically occur at the modality, sample, or encoded-feature level. QRAFT addresses a complementary problem: both sources may be present, yet only some local regions are reliable. It therefore combines source-specific quality estimates with feature-level fusion before noisy waveform evidence influences shared attention. The cross-source encoder allows ECG and PPG to support one another according to the local evidence set created by masking. This local control is distinct from PTT-based BP pipelines, which use a timing surrogate and require a separate calibration and fiducial-detection protocol (Kachuee et al., 2015).

3 Reliability-Conditioned Masking

3.1 From Quality Scores to Reconstruction Targets

Let $X^r \in \mathbb{R}^T$ denote source $r \in \mathcal{R}$, partitioned into patches $s_i^r \in \mathbb{R}^\omega$. A fine-grained quality estimator g_ψ^r outputs point scores $q_t^r \in [0,1]$, aggregated as $Q_i^r = \omega^{-1} \sum_t q_t^r$. We define

$$m_i^r = \mathbb{1}(Q_i^r < \theta), \qquad \bar{z}_i^r = (1 - m_i^r) e_i^r + m_i^r (t_{\text{mask}}^r + p_i^r). \tag{1}$$

The binary gate follows from a patch assignment view. If visible evidence has utility $U_{\text{vis}} = Q_i$ and a reconstruction target has utility $U_{\text{rec}} = \lambda_q (1 - Q_i)$, then

$$m_i^* = \arg \max_{m \in \{0,1\}} (1-m) U_{\text{vis}} + m U_{\text{rec}} = \mathbb{1}\left(Q_i \le \frac{\lambda_q}{1 + \lambda_q} \right). \tag{2}$$

The gate is used as an operational decision rule; an entropy-regularized relaxation approaches this hard limit as its temperature decreases.

The mask also defines the evidence available to the shared encoder. Let

$$\begin{aligned} \mathcal{V} &= \{(r,i) : m_i^r = 0\}, \qquad \mathcal{M} = \{(r,i) : m_i^r = 1\}, \\ Z_{\mathcal{V}} &= f_\phi(\{\bar{z}_i^r : (r,i) \in \mathcal{V}\}), \qquad \hat{s}_i^r = d_r(Z_{\mathcal{V}}, p_i^r). \end{aligned} \tag{3}$$

Thus, $\mathcal{V}$ contains the measured evidence that can participate in cross-source encoding, whereas $\mathcal{M}$ contains positions whose observed waveforms are predicted from that evidence. The decoder is queried at masked positions but cannot access their original waveform values; the partition therefore couples target selection and source participation without introducing an additional modality-level weight.

FGQA is trained from coarse expert labels by multi-instance learning. It divides each record into 0.5 s instances and aligns pooled instance predictions with the record label. Multi-scale convolution captures local morphology, while later self-attention stages provide longer context. In the dedicated held-out QA evaluation, the estimator reaches 0.969 PPG and 0.964 ECG accuracy; the separate baseline-comparison protocol is reported in Supplementary Section B.

The available supervision consists of coarse expert labels rather than dense artifact masks. Each five-second record has a coarse expert decision, but the estimator must localize short failures inside it. Let a_j be the predicted acceptability of instance j. Training aligns the pooled prediction $\bar{a} = M^{-1} \sum_j a_j$ with the record label while a regression head processes the instance statistics. The pooled term preserves compatibility with expert supervision, and the instance head discourages every subwindow from receiving the same score. The output is used as a weakly supervised local reliability ranking for target selection.

Labels are aggregated from three independent experts by majority vote, with unresolved cases adjudicated. Pairwise agreement ranges from 0.74 to 0.82 for PPG and from 0.71 to 0.79 for ECG; Fleiss' κ is 0.78 and 0.74. Thresholds are selected on a held-out calibration split rather than the reconstruction test set. Perturbing agreement over $\kappa \in [0.6, 0.9]$ produces reconstruction RMSE values within 0.60–0.71, covering the operating result reported below.

3.2 Shared Temporal and Cross-Source Encoding

Each patch uses convolution branches with kernel sizes $\{1, 3, 5, 7\}$ and a linear amplitude projection, following common 1D architectures for biomedical sequences (Kiranyaz et al., 2021). Fixed sinusoidal, fixed periodic, and learnable positional components are then added. For synchronized sources, visible ECG and PPG tokens are concatenated into one sequence. A head-specific attention block from source s to source r is

$$A_h^{r \leftarrow s} = \mathrm{softmax}_{(u,j)} \left(\frac{Q_h^r (K_h^s)^\top}{\sqrt{d_h}} + B_h^{r,s} \right), \quad O_h^r = A_h^{r \leftarrow r} V_h^r + \sum_{s \neq r} A_h^{r \leftarrow s} V_h^s. \tag{4}$$

Reliability does not alter attention logits; it determines which patches can participate before attention is computed. This separates QRAFT from late quality weighting, which first encodes every patch and only then rescales source-level output. A low-quality patch keeps its position through the mask token and positional encoding, but its measured waveform cannot enter shared attention. The corresponding encoder–decoder path is summarized in Fig. 2. We evaluate this fusion mechanism with bidirectional reconstruction, quality-stratified BP analysis, and a factorial comparison that holds encoder capacity and fine-tuning data fixed while changing pretraining and fusion rules. These controls identify when complementary evidence is useful.

The encoder applies a gated two-branch feed-forward block after attention:

$$E \leftarrow E + \gamma \odot \left(W_2 \sigma(W_1 E) + W_4 \, \mathrm{GELU}(W_3 E) \right), \qquad \gamma = \mathrm{sigmoid}(W_g [E; W_s E]). \tag{5}$$

Stochastic depth regularizes the encoder. Reconstruction heads are lightweight and source specific; they are discarded after pretraining, and the encoder is reused for linear evaluation or end-to-end fine-tuning.

3.3 Temporal–Spectral Reconstruction

For masked set $\mathcal{M}$, the waveform loss combines ℓ_2 and ℓ_1 residuals. A contrastive term compares independently augmented views, and a spectral term compares log-magnitude spectra:

$$\mathcal{L}_t = |\mathcal{M}|^{-1} \sum_{(r,i) \in \mathcal{M}} \left(\alpha \| \hat{s}_i^r - s_i^r \|_2^2 + \beta \| \hat{s}_i^r - s_i^r \|_1 \right), \tag{6}$$

$$\mathcal{L}_f = |\mathcal{M}|^{-1} \sum_{(r,i) \in \mathcal{M}} \left\| \log(|\mathcal{F}(\hat{s}_i^r)| + \epsilon) - \log(|\mathcal{F}(s_i^r)| + \epsilon) \right\|_1, \tag{7}$$

$$\mathcal{L} = \mathcal{L}_t + \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_f. \tag{8}$$

$\mathcal{L}_f$ is not equivalent to time-domain squared error under Parseval's identity because it operates on log magnitudes rather than complex DFT residuals. Here s_i^r denotes the original observed patch, including any artifact present in that observation. The objective predicts observed waveform content, while cross-source support is evaluated through representation quality and downstream prediction. We use $\alpha = 0.75$, $\beta = 0.25$, $\lambda_1 = 0.2$, and $\lambda_2 = 0.1$.

Each pretraining step has three stages. FGQA first scores every source and constructs $\mathcal{M}^r = \{i : Q_i^r < \theta\}$. Visible and mask tokens from all available sources are then concatenated and encoded jointly, after which each decoder reconstructs only its source's masked patches. Finally, losses are

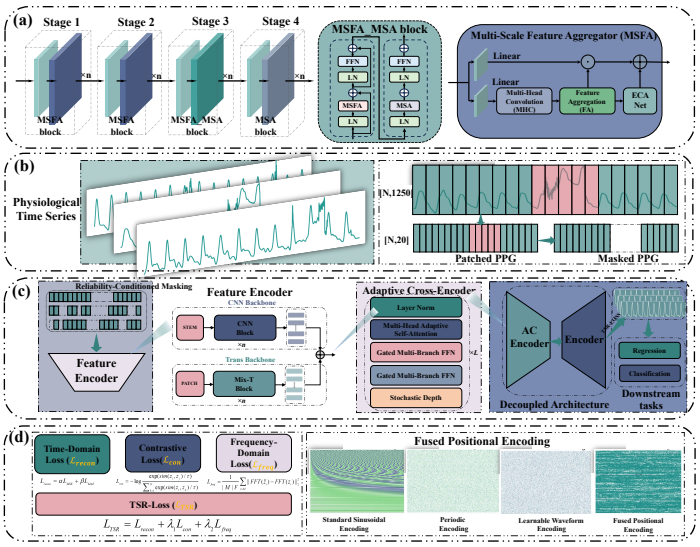


Figure 2: Architecture. FGQA generates patch reliability, masking defines the evidence set, shared attention integrates available temporal and cross-source tokens, and modality-specific heads reconstruct masked patches.

normalized by the masked count and optimized together. Positive contrastive pairs are independently augmented views of the same segment; negatives are other records in the batch. Synchronized ECG and PPG from one record are never used as negatives.

If $Z = f(X_v)$ and X_m is the masked set, a Gaussian decoder gives the familiar reconstruction interpretation

$$I(Z; X_m) \geq C - \mathbb{E}\|X_m - d(Z)\|_2^2. \tag{9}$$

This bound interprets target placement as changing the information requested from the encoder; it does not add a mutual-information term to the training objective.

4 Experiments

4.1 Setup

VitalDB provides synchronized ECG, PPG, and arterial BP recordings from 1,000 subjects; QRAFT pretraining uses ECG and PPG as encoder inputs, while arterial BP is not an input to the representation or reconstruction pathway. Subject-independent transfer is evaluated on MIMIC-III, UCI, and a 32-participant in-house feasibility cohort (Lee et al., 2022; Kachuee et al., 2015; Johnson et al., 2016). Baselines share preprocessing, encoder capacity, optimization budget, and pretraining subjects. The comparison covers supervised, generic self-supervised, physiology-specific, and QA-filtered masked-learning models. The main controls are DM-Rand, quality-agnostic cross-attention, and Soft-QW, which applies FGQA only at late fusion. We report reconstruction RMSE, PRD, and SNR; BP MAE and R^2; and QA accuracy, ECE, and Brier score. Ten-seed comparisons use exact two-sided paired Wilcoxon tests, whereas five-run transfer results are descriptive and paired by run rather than by window. Cohort, preprocessing, and implementation details are in Supplementary Section A.

4.2 Masking Controls

All policies in Table 2 mask 38.5% of PPG patches. DM-Rand scores difficulty by normalized local variance, sample entropy, and log high-frequency energy, then samples from the hardest bin. With the mask ratio and encoder budget held constant, this comparison isolates where targets are selected rather than how much input is withheld. Reliability-conditioned masking outperforms both random and DM-Rand. Its 0.63 RMSE is a 23.2% reduction from random masking and a 10.0% reduction from DM-Rand. On the lowest-quality quartile, RMSE decreases from 1.08 to 0.92; on the

Table 1: Cross-dataset BP estimation under subject-independent transfer, compared with supervised and self-supervised baselines. Values are mean±std across five runs. This transfer protocol is distinct from the synchronized multimodal fusion protocol in Table 5.

Model	MIMIC-III			UCI			IHC		
	MAE	SD	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
Supervised	5.68±0.25	7.53±0.38	0.86±0.02	5.19±0.24	6.97±0.36	0.88±0.02	6.16±0.29	8.17±0.45	0.83±0.02
Linear evaluation									
SimCLR	6.35±0.32	8.41±0.45	0.81±0.03	5.87±0.29	7.84±0.42	0.84±0.02	6.79±0.35	8.97±0.53	0.78±0.03
TS-TCC	6.12±0.29	8.13±0.43	0.83±0.02	5.63±0.27	7.59±0.40	0.85±0.02	6.58±0.32	8.72±0.51	0.80±0.03
MoCo v3	6.82±0.34	8.95±0.56	0.79±0.03	6.25±0.31	8.37±0.49	0.82±0.03	7.23±0.38	9.54±0.61	0.76±0.04
MAE	5.97±0.28	7.92±0.41	0.84±0.02	5.41±0.25	7.28±0.38	0.87±0.02	6.39±0.31	8.48±0.48	0.81±0.03
CLOCS	5.82±0.26	7.74±0.40	0.85±0.02	5.28±0.24	7.12±0.37	0.87±0.02	6.25±0.30	8.29±0.47	0.82±0.03
ST-MEM	5.74±0.26	7.62±0.39	0.86±0.02	5.21±0.23	7.05±0.37	0.87±0.02	6.18±0.29	8.21±0.46	0.82±0.02
QA-filtered MAE	5.63±0.25	7.48±0.38	0.86±0.02	5.09±0.22	6.89±0.35	0.88±0.02	6.04±0.28	8.02±0.44	0.83±0.02
QRAFT	**4.93±0.21**	**6.57±0.35**	**0.89±0.01**	**4.52±0.19**	**6.12±0.32**	**0.91±0.01**	**5.43±0.24**	**7.28±0.41**	**0.87±0.01**
Fine-tuning									
SimCLR	5.03±0.21	6.67±0.37	0.89±0.01	4.41±0.18	5.98±0.33	0.91±0.01	5.42±0.24	7.21±0.39	0.87±0.02
TS-TCC	5.21±0.23	6.91±0.39	0.88±0.02	4.68±0.20	6.31±0.35	0.90±0.01	5.67±0.25	7.53±0.42	0.86±0.02
MoCo v3	5.47±0.25	7.25±0.42	0.87±0.02	4.93±0.21	6.64±0.36	0.89±0.02	5.89±0.27	7.84±0.44	0.84±0.04
MAE	4.58±0.19	6.05±0.34	0.91±0.01	3.97±0.16	5.39±0.30	0.93±0.01	5.07±0.22	6.74±0.36	0.89±0.01
CLOCS	4.72±0.20	6.26±0.35	0.90±0.01	4.13±0.17	5.60±0.31	0.92±0.01	5.18±0.23	6.89±0.37	0.88±0.01
ST-MEM	4.65±0.19	6.15±0.34	0.91±0.01	4.06±0.16	5.50±0.30	0.92±0.01	5.12±0.22	6.81±0.36	0.88±0.01
QA-filtered MAE	4.47±0.19	5.92±0.33	0.92±0.01	3.86±0.15	5.24±0.29	0.93±0.01	4.93±0.21	6.57±0.35	0.90±0.01
QRAFT	**4.09±0.17**	**5.20±0.32**	**0.93±0.02**	**3.52±0.14**	**4.67±0.28**	**0.94±0.01**	**4.55±0.16**	**5.78±0.30**	**0.92±0.02**

NOTE. IHC is an in-house feasibility cohort; its results are not population-level clinical validation.

Table 2: Reconstruction performance of masking strategies at the matched 38.5% ratio.

Subset	Policy	RMSE↓	PRD↓	SNR↑
Full	Random	$0.82 \pm .11$	18.7 ± 1.94	$14.3 \pm .62$
Full	Contiguous	$0.95 \pm .21$	21.2 ± 2.71	$13.1 \pm .48$
Full	Hybrid	$0.88 \pm .17$	19.8 ± 1.45	$14.0 \pm .55$
Full	DM-Rand	$0.70 \pm .09$	16.4 ± 1.15	$15.6 \pm .50$
Full	**QRAFT**	$\mathbf{0.63 \pm .08}^{\ddagger}$	$\mathbf{14.9 \pm .77}$	$\mathbf{16.5 \pm .45}$
High quality	Random	$0.38 \pm .04$	$8.6 \pm .51$	21.3 ± 3.12
High quality	Contiguous	$0.43 \pm .06$	$9.8 \pm .53$	20.2 ± 2.54
High quality	Hybrid	$0.46 \pm .09$	$10.5 \pm .58$	19.6 ± 1.95
High quality	DM-Rand	$0.40 \pm .05$	$9.2 \pm .53$	20.7 ± 1.82
High quality	**QRAFT**	$0.41 \pm .06$	$9.4 \pm .54$	20.5 ± 1.38
Low quality	Random	$1.44 \pm .26$	32.6 ± 3.07	$9.7 \pm .59$
Low quality	Contiguous	$1.67 \pm .43$	37.8 ± 3.46	$8.5 \pm .48$
Low quality	Hybrid	$1.38 \pm .41$	31.2 ± 3.19	$10.1 \pm .62$
Low quality	DM-Rand	$1.08 \pm .15$	24.9 ± 1.83	$12.3 \pm .57$
Low quality	**QRAFT**	$\mathbf{0.92 \pm .03}^{\ddagger}$	$\mathbf{20.8 \pm 1.80}$	$\mathbf{13.6 \pm .51}$

$\ddagger p < 0.01$ versus DM-Rand over ten seeds.

highest-quality quartile, the two methods are tied within 0.01. Reliability therefore changes target location beyond the effect of selecting difficult patches.

The remaining policies clarify the role of mask geometry. Contiguous masking is hardest because long gaps remove nearby temporal support. Hybrid masking partly alleviates that failure but remains worse than random masking at the matched ratio. DM-Rand improves substantially, confirming that hard targets are useful. Its residual gap to QRAFT shows that difficulty and reliability are not interchangeable: variance or entropy may be high in informative morphology, whereas FGQA is trained to determine whether that morphology is usable.

The reconstruction results also preserve waveform features relevant to downstream use; qualitative examples in Fig. 3 show the corresponding masked-prediction patterns, and the complete feature-level analysis is provided in Supplementary Section C.

Quality stratification provides a complementary transfer check. Across high-, medium-, and low-quality PPG subsets, QRAFT maintains the lowest errors: SBP/MAP/DBP MAE is 4.01/3.02/2.57 mmHg for high-quality signals, 4.18/3.17/2.72 mmHg for medium-quality signals, and 4.36/3.33/2.84 mmHg for low-quality signals (Table 3). The advantage is modest when PPG is clean and grows as

Table 3: Quality-stratified BP performance.

Method	High quality signal			Medium quality signal			Low quality signal		
	SBP↓	MAP↓	DBP↓	SBP↓	MAP↓	DBP↓	SBP↓	MAP↓	DBP↓
I: Coarse-grained	5.54±0.50	4.06±0.32	3.62±0.30	6.53±0.63	4.98±0.50	4.00±0.50	8.82±0.78	7.02±0.68	6.13±0.56
II: No quality perception	5.73±0.48	4.32±0.40	3.75±0.35	7.26±0.69	5.62±0.44	4.83±0.53	10.38±0.89	8.24±0.79	6.92±0.60
III: Without reconstruction	5.68±0.55	4.27±0.45	3.71±0.27	6.45±0.59	4.92±0.47	4.21±0.49	8.14±0.79	6.35±0.65	5.37±0.55
QRAFT	**4.01**	**3.02**	**2.57**	**4.18**	**3.17**	**2.72**	**4.36**	**3.33**	**2.84**
PI	27.6%	25.6%	29.0%	35.2%	35.6%	35.4%	46.4%	47.6%	47.1%

NOTE. Quality strata are defined on PPG subsets. PI is QRAFT's relative improvement over the best non-QRAFT method in each column, using fixed-protocol regime-level estimates.

the input becomes less reliable, supporting the interpretation that quality conditioning changes source participation rather than applying a uniform gain to every regime.

4.3 Loss and Mechanism Ablations

The objective and pretraining–fusion decomposition in Supplement C separates the objective terms and the two design factors. The time-domain term is indispensable: using frequency and contrastive terms without it gives 0.75 RMSE. Adding spectral alignment to reconstruction improves RMSE more than adding contrastive regularization, whereas the contrastive term provides the larger reduction in downstream SBP MAE. The three terms therefore play different roles, and the complete objective performs best on both tasks.

The factorial block tests a separate question. Under random-mask pretraining, replacing generic cross-attention with the QRAFT fuser improves MAE by 5.8%. Under QRAFT pretraining, the same replacement improves it by 7.9%. Conversely, quality-conditioned pretraining improves both fusion columns. Target selection and source participation therefore make distinct contributions.

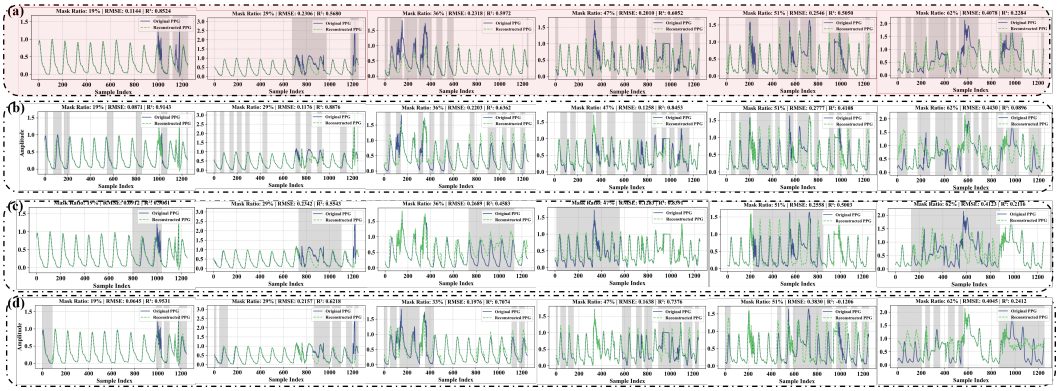


Figure 3: Representative masked predictions under the five compared masking policies.

4.4 Transfer Across Datasets

QRAFT obtains lower SBP error under both frozen linear evaluation and end-to-end fine-tuning (Table 1). Fine-tuned MAE reaches 4.09 mmHg on MIMIC-III, 3.52 on UCI, and 4.55 on the in-house cohort. The QA-filtered MAE comparison isolates the role of target selection: filtering low-quality records before random-mask pretraining helps, but remains worse than using quality to define local targets inside the pretraining loop.

Frozen evaluation and fine-tuning answer different questions. Linear evaluation measures whether pretraining already organizes the representation around BP-relevant structure. Fine-tuning measures whether this structure adapts with task labels. QRAFT ranks first in both regimes. The two public datasets provide the primary transfer evidence, while IHC provides a feasibility evaluation.

The auxiliary classification transfer experiment evaluates the encoder under random-mask pretraining because fine-grained quality scores are unavailable for those datasets. Table 4 reports the complete accuracy and F1 comparison, while Fig. 4 provides complementary confusion-matrix and feature-space

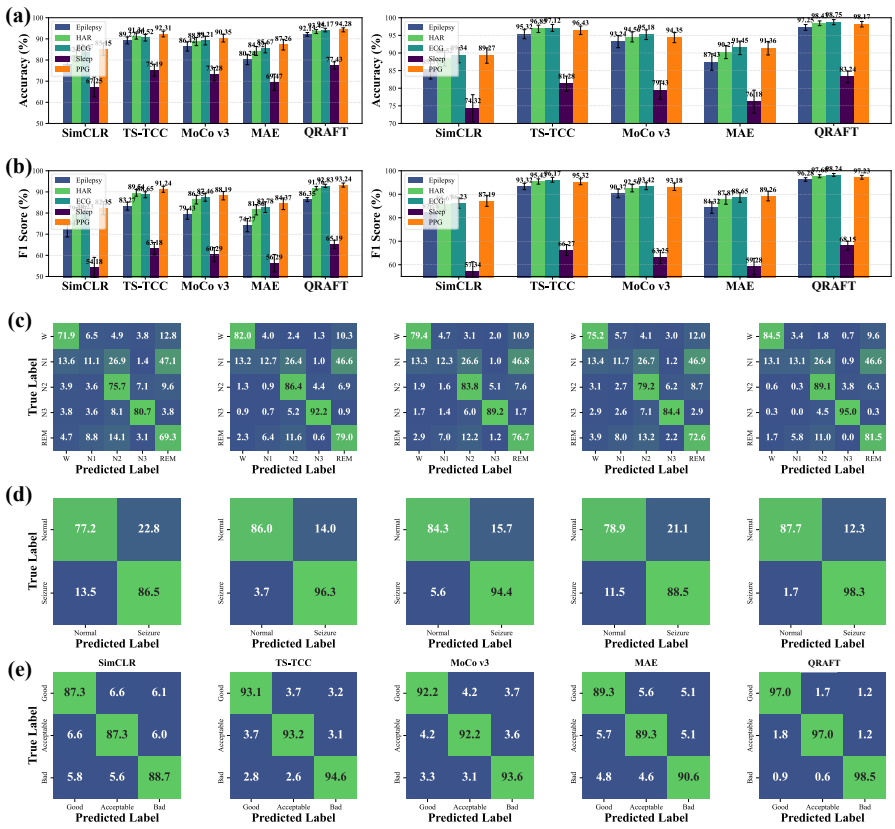


Figure 4: Auxiliary transfer under quality-agnostic random-mask pretraining.

Table 4: Auxiliary classification transfer under random-mask pretraining. Values are mean±std across repeated runs.

Method	Linear evaluation / Accuracy					Fine-tuning / Accuracy				
	Epilepsy	HAR	ECG-SQA	Sleep-EDF	PPG-SQA	Epilepsy	HAR	ECG-SQA	Sleep-EDF	PPG-SQA
SimCLR	78.02±3.54	82.18±2.87	76.36±3.42	67.25±4.73	85.15±2.95	85.37±2.76	84.24±2.45	86.43±2.63	74.32±3.84	89.27±2.18
TS-TCC	89.27±1.62	89.54±1.38	86.42±1.47	75.19±2.84	92.31±1.35	95.32±1.28	91.25±1.42	94.17±1.35	81.28±2.16	96.43±1.21
MoCo v3	86.42±2.18	87.39±1.96	83.17±2.05	73.28±3.17	90.35±1.82	93.24±1.75	89.47±1.93	92.36±1.78	79.43±2.65	94.35±1.43
MAE	80.24±2.47	83.35±2.35	78.53±2.81	69.47±3.86	87.26±2.53	87.43±2.34	85.68±2.17	88.25±2.42	76.18±3.27	91.36±1.93
QRAFT	**92.13±0.87**	**91.25±0.93**	**88.37±0.76**	**77.43±1.72**	**94.28±0.95**	**97.25±0.82**	**92.37±0.89**	**95.18±0.92**	**83.24±1.64**	**98.17±0.78**

Method	Linear evaluation / F1					Fine-tuning / F1				
	Epilepsy	HAR	ECG-SQA	Sleep-EDF	PPG-SQA	Epilepsy	HAR	ECG-SQA	Sleep-EDF	PPG-SQA
SimCLR	72.34±3.68	80.19±3.25	79.27±3.47	54.18±4.85	82.35±3.17	82.26±2.83	83.15±2.64	83.42±2.73	57.34±3.92	87.19±2.34
TS-TCC	83.27±1.87	89.24±1.63	88.31±1.76	63.18±2.93	91.24±1.42	93.32±1.35	91.19±1.53	91.28±1.47	66.27±2.32	95.32±1.26
MoCo v3	79.43±2.34	86.17±2.12	85.28±2.25	60.29±3.35	88.19±1.98	90.37±1.85	89.26±1.82	89.42±1.93	63.25±2.86	93.18±1.57
MAE	74.27±3.12	82.34±2.76	81.18±2.91	56.29±4.13	84.37±2.75	84.32±2.46	85.24±2.32	85.19±2.54	59.28±3.45	89.26±2.08
QRAFT	**86.35±0.92**	**91.18±0.85**	**90.27±0.88**	**65.19±1.98**	**93.24±0.97**	**96.28±0.76**	**92.24±0.83**	**93.17±0.87**	**68.15±1.85**	**97.23±0.82**

views. This experiment isolates transfer of the encoder under a quality-agnostic pretraining protocol; the quality-conditioned masking mechanism is evaluated in the masking and fusion experiments.

4.5 When Does Cross-Source Context Help?

Multimodal input reduces ECG reconstruction RMSE from 0.76 to 0.64 and PPG RMSE from 0.61 to 0.52. The bidirectional reconstruction table and its analysis are provided in Supplementary Section C. We then stratify synchronized MIMIC-III segments by ECG and PPG reliability. Fusion gains are smallest when both sources are clean, increase when one source is degraded, and peak when both are degraded. This pattern ties the measured fusion gain to the local quality variable used by QRAFT and

Table 5: Synchronized multimodal fusion on MIMIC-III; mean±std across ten seeds.

Method	SBP			DBP			MAP		
	MAE↓	SD↓	R^2 ↑	MAE↓	SD↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑
PPG-PM	4.09±0.15	5.20±0.28	0.93±0.02	2.63±0.07	3.95±0.19	0.94±0.02	3.12±0.10	4.35±0.18	0.93±0.02
Early Fusion	3.95±0.17	5.31±0.19	0.92±0.02	2.58±0.05	3.87±0.15	0.93±0.02	3.04±0.09	4.28±0.19	0.93±0.01
Late Fusion	3.83±0.13	5.19±0.24	0.92±0.01	2.52±0.05	3.80±0.21	0.94±0.01	2.96±0.07	4.17±0.13	0.94±0.02
Cross-Att Fusion	3.65±0.12	5.02±0.22	0.93±0.01	2.41±0.04	3.68±0.17	0.94±0.01	2.84±0.06	4.02±0.11	0.94±0.01
Soft-QW Fusion	3.57±0.11	4.89±0.19	0.93±0.01	2.35±0.04	3.58±0.13	0.94±0.01	2.78±0.06	3.91±0.10	0.95±0.01
QRAFT	**3.36±0.10**	**4.58±0.17**	**0.94±0.01**	**2.22±0.04**	**3.41±0.07**	**0.95±0.00**	**2.63±0.05**	**3.72±0.09**	**0.95±0.00**
Improvement over PPG-PM	17.8%	11.9%	1.1%	15.6%	13.7%	1.1%	15.7%	14.5%	2.2%

shows that complementarity varies across quality regimes. The corresponding representation-space visualization is shown in Fig. 5.

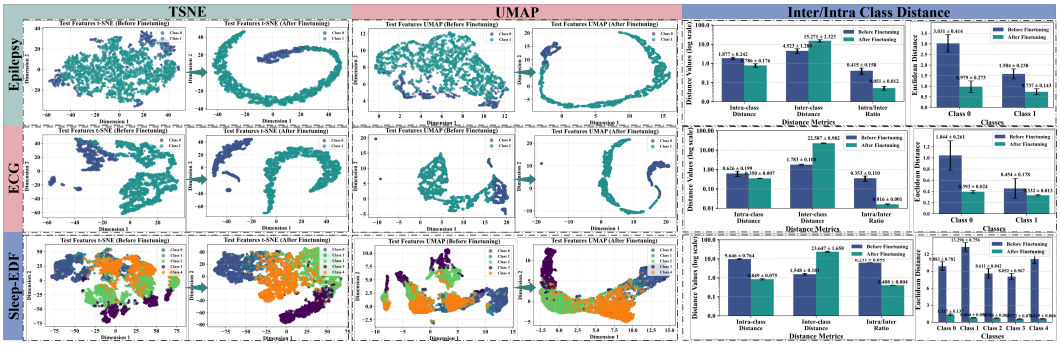


Figure 5: Representation structure before and after fine-tuning across auxiliary physiological tasks.

Under this synchronized multimodal protocol, QRAFT reduces SBP, DBP, and MAP MAE from 3.57, 2.35, and 2.78 to 3.36, 2.22, and 2.63 mmHg versus Soft-QW (Table 5). The 3.36 mmHg SBP result belongs to this fusion setting, whereas the 4.09 mmHg result in Table 1 comes from cross-dataset transfer. The factorial analysis in Supplementary Section C separates pretraining from fusion under matched downstream conditions. Consistent ranking across all three pressure outputs supports a common quality-conditioned participation effect.

5 Discussion and Limitations

The matched-difficulty, quality-regime, and factorial analyses identify reliability as the operative variable: they separate target selection from hard-example mining, show larger gains under degradation, and distinguish pretraining from fusion. A source can contain intervals that should not enter shared attention; hard gating is effective for the masked objective, while continuous weighting remains available for other tasks. Quality therefore selects patches, reconstruction uses remaining context, and shared attention integrates supported signals. Across controlled comparisons, the benefit depends on where evidence enters the learning path: target selection changes what the encoder must infer, whereas cross-source attention determines which visible context supports that inference. This distinction explains the larger gain in jointly degraded regimes and why matching masking difficulty alone does not reproduce the full improvement.

6 Conclusion

QRAFT incorporates local signal reliability into physiological self-supervision by using weakly supervised ECG and PPG scores to define visible evidence and reconstruction targets, while synchronized cross-source context supports prediction of withheld observations. Matched-difficulty and factorial analyses separate hard-example effects from target-selection and fusion-rule effects. Across the evaluated datasets, QRAFT obtains lower reconstruction and blood-pressure errors, with larger fusion gains when the sources are degraded. The results support reliability-conditioned source participation as a practical design principle for multimodal physiological representation learning, in which evidence selection and cross-source support are learned within a shared pretraining objective.

AI USE STATEMENT

Generative AI tools were used for language polishing. All data analyses, experiments, figures, tables, and scientific conclusions were produced and verified by the authors.

ETHICS STATEMENT

Public datasets are used under their repository access terms. Participants in the in-house cohort provided written informed consent, and the cohort was collected under approval from the relevant institutional ethics committee using coded identifiers. The model is evaluated as a representation-learning method and is not presented as a clinically validated blood-pressure device.

REPRODUCIBILITY STATEMENT

Code and configuration files for the QRAFT model, training protocols, and evaluation scripts are available at `https://anonymous.4open.science/r/QRAFT-1/`. The repository is accompanied by the configuration files and documentation needed to reproduce the masking, reconstruction, fusion, and transfer experiments; dataset access remains subject to the original repository terms.

REFERENCES

Y. Ansari, O. Mourad, K. Qaraqe, and E. Serpedin. Deep learning for ecg arrhythmia detection and classification: an overview of progress for period 2017-2023. *Frontiers in Physiology*, 14, 2023.

V. A. Ardeti, V. R. Kolluru, G. T. Varghese, and R. K. Patjoshi. An overview on state-of-the-art electrocardiogram signal processing methods: Traditional to ai-based approaches. *Expert Systems with Applications*, 217, 2023.

S. Bhattachrjee, H. Li, J. Xia, and W. Xu. Simppg: Self-supervised photoplethysmography-based heart-rate estimation via similarity-enhanced instance discrimination. *Smart Health*, 28:100396, 2023.

S. C. Chen, Q. L. Ouyang, X. L. Meng, Y. B. Yang, C. Li, X. B. Miao, Z. H. Chen, G. G. Zhao, Y. G. Lei, B. Ghanem, S. Gautam, J. L. Cheng, and Z. Yan. Starfish-inspired wearable bioelectronic systems for physiological signal monitoring during motion and real-time heart disease diagnosis. *Science Advances*, 11(14), 2025.

Z. Y. Deng, L. H. Guo, X. M. Chen, and W. W. Wu. Smart wearable systems for health monitoring. *Sensors*, 23(5), 2023.

C. Ding and C. Wu. Self-supervised learning for biomedical signal processing: A systematic review on ecg and ppg signals. *medRxiv*, 2024.

S. S. Gambhir, T. J. Ge, O. Vermesh, R. Spitler, and G. E. Gold. Continuous health monitoring: An opportunity for precision health. *Science Translational Medicine*, 13(597), 2021.

Z. H. Hao, J. N. Wang, G. Zhang, L. Z. Gao, X. M. Zhang, J. Liu, X. L. Zhang, X. L. Yang, and Z. X. Lai. Ppg heart rate extraction algorithm based on the motion artifact intensity classification and removal framework. *Biomedical Signal Processing and Control*, 94, 2024.

Y. R. Jin, Z. Y. Li, C. J. Qin, J. L. Liu, Y. Q. Liu, L. Q. Zhao, and C. L. Liu. A novel attentional deep neural network-based assessment method for ecg quality. *Biomedical Signal Processing and Control*, 79, 2023.

A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

A. Joutsen, A. Cömert, E. Kaappa, K. Vanhatalo, J. Riistama, A. Vehkaoja, and H. Eskola. Ecg signal quality in intermittent long-term dry electrode recordings with controlled motion artifacts. *Scientific Reports*, 14(1), 2024.

M. Kachuee, M. M. Kiani, H. Mohammadzade, and M. Shabany. Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time. In *IEEE International Symposium on Circuits and Systems*, pp. 1006–1009, 2015.

S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman. 1d convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, 151, 2021.

R. Krishnan, P. Rajpurkar, and E. J. Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022.

M. Lage Cañellas, C. Álvarez Casado, L. Nguyen, and M. Bordallo López. A self-supervised multimodal framework for 1d physiological data fusion in remote health monitoring. *Information Fusion*, 124:103397, 2025.

T.-D. Le, C. Macabiau, K. Albert, P. Jouvet, and R. Noumeir. A novel transformer-based self-supervised learning method to enhance photoplethysmogram signal artifact detection. *IEEE Access*, 12:159860–159874, 2024.

P. H. Le-Khac, G. Healy, and A. F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.

H.-C. Lee, Y. Park, S. B. Yoon, S. M. Yang, D. Park, and C.-W. Jung. Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9:279, 2022.

Y. B. Liang, M. Elgendi, Z. C. Chen, and R. Ward. An optimal filter for short photoplethysmogram signals. *Scientific Data*, 5:180076, 2018.

J. Liu, S. Hu, Y. Wang, Q. Hu, D. Wang, and C. Yang. A lightweight hybrid model using multiscale markov transition field for real-time quality assessment of photoplethysmography signals. *IEEE Journal of Biomedical and Health Informatics*, 28(2):1078–1088, 2024a.

J. Liu, S. C. Hu, Y. A. Wang, W. Xiang, Q. H. Hu, and C. W. Yang. Personalized blood pressure estimation using multiview fusion information of wearable physiological signals and transfer learning. *Applied Soft Computing*, 167:112390, 2024b.

J. Liu, S. Hu, Y. Wang, Q. Hu, D. Wang, W. Xiang, X. Feng, and C. Yang. Leaf-net: A real-time fine-grained quality assessment system for physiological signals using lightweight evolutionary attention fusion. *Expert Systems with Applications*, 274:126995, 2025.

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

F. Mohagheghian, D. Han, A. Peitzsch, N. Nishita, E. Ding, E. L. Dickson, D. DiMezza, E. M. Otabil, K. Noorishirazi, J. Scott, D. Lessard, Z. Wang, C. Whitcomb, K.-V. Tran, T. P. Fitzgibbons, D. D. McManus, and K. H. Chon. Optimized signal quality assessment for photoplethysmogram signals using feature selection. *IEEE Transactions on Biomedical Engineering*, 69(9):2982–2993, 2022.

Y. Na, M. Park, Y. Tae, and S. Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International Conference on Learning Representations*, 2024.

T. Pereira, C. Ding, K. Gadhoumi, N. Tran, R. A. Colorado, K. Meisel, and X. Hu. Deep learning approaches for plethysmography signal quality assessment in the presence of atrial fibrillation. *Physiological Measurement*, 40(12), 2019.

S. Rahman, C. Karmakar, I. Natgunanathan, J. Yearwood, and M. Palaniswami. Robustness of electrocardiogram signal quality indices. *Journal of the Royal Society Interface*, 19(189), 2022.

D. Roh and H. Shin. Recurrence plot and machine learning for signal quality assessment of photoplethysmogram in mobile environment. *Sensors*, 21(6), 2021.

F. Scardulla, G. Cosoli, S. Spinsante, A. Poli, G. Iadarola, R. Pernice, A. Busacca, S. Pasta, L. Scalise, and L. D'Acquisto. Photoplethysmographic sensors, potential and limitations: Is it time for regulation? a comprehensive review. *Measurement*, 218:113150, 2023.

J. Schmith, C. Kelsch, B. C. Cunha, L. R. Prade, E. A. Martins, A. L. Keller, and R. M. de Figueiredo. Photoplethysmography signal quality assessment using attractor reconstruction analysis. *Biomedical Signal Processing and Control*, 86, 2023.

H. Shin. Deep convolutional neural network-based signal quality assessment for photoplethysmogram. *Computers in Biology and Medicine*, 145, 2022.

S. Soltanieh, J. Hashemi, and A. Etemad. In-distribution and out-of-distribution self-supervised ecg representation learning for arrhythmia detection. *IEEE Journal of Biomedical and Health Informatics*, 28(2):789–800, 2024.

Z. H. Tian, A. P. Liu, G. K. Zhu, and X. Chen. A paralleled cnn and transformer network for ppg-based cuff-less blood pressure estimation. *Biomedical Signal Processing and Control*, 99, 2025.

G. Victor, K. Shishani, E. Vellone, and E. S. Froelicher. The global burden of cardiovascular disease in adults: A mapping review. *Journal of Cardiovascular Nursing*, 40(6):523–537, 2025.

Y. Wu, J. Chen, L. Hu, H. Xu, H. Liang, and J. Wu. Omnifuse: A general modality fusion framework for multi-modality learning on low-quality medical data. *Information Fusion*, 117:102890, 2025.

S. X. Yang, C. Lian, Z. G. Zeng, B. R. Xu, Y. X. Su, and C. Y. Xue. Masked self-supervised ecg representation learning via multiview information bottleneck. *Neural Computing and Applications*, 36(14):7625–7637, 2024.

H. Zhang, W. Liu, J. Shi, S. Chang, H. Wang, J. He, and Q. Huang. Maefe: Masked autoencoders family of electrocardiogram for self-supervised pretraining and transfer learning. *IEEE Transactions on Instrumentation and Measurement*, 72:1–15, 2023.

K. Zhang, Q. Wen, C. Zhang, R. Cai, M. Jin, Y. Liu, J. Y. Zhang, Y. Liang, G. Pang, D. Song, and S. Pan. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6775–6794, 2024.

X. Zhou, X. Zhu, K. Nakamura, and M. Noro. Electrocardiogram quality assessment with a generalized deep learning model assisted by conditional generative adversarial networks. *Life*, 11 (10):1013, 2021.

SUPPLEMENTARY MATERIAL
LEARNING WHAT TO TRUST: RELIABILITY-CONDITIONED MASKED MODELING FOR PHYSIOLOGICAL TIME SERIES

Anonymous authors
Paper under double-blind review

A. ADDITIONAL EXPERIMENTAL PROTOCOL

All experiments use the same subject-independent partitions, 125 Hz resampling, ten-second segments, and 1,000-subject VitalDB pretraining pool described in the main paper. Fine-grained QA uses 0.5 s instances. MIMIC-III contains 1,000 selected subjects with 400 records per subject, UCI contains 12,000 records from 942 patients, and the IHC feasibility cohort contains 32 adults aged 25–78 years (mean 52.4, SD 14.7; 18 male and 14 female). IHC tests use run-level aggregate metrics rather than treating windows as independent inferential units. The matched masking comparison uses threshold $\theta = 0.7$ and a 38.5% mask ratio; the unconstrained threshold sweep reaches its best reported reconstruction point at $\theta = 0.8$. Results based on five runs are descriptive, while symbol-marked comparisons use ten paired runs and exact two-sided Wilcoxon tests.

A.1 REPRODUCIBLE METHOD SPECIFICATION

QRAFT executes five steps: quality estimation, patch assignment, shared encoding, source-specific reconstruction, and joint loss optimization. Quality scores determine only the visible evidence set; they are not inserted as an additional attention weight. Source-specific patch positions remain available through mask tokens and positional encodings, so each decoder can query a withheld location without receiving its original waveform.

Table S1: Notation and tensor flow used by QRAFT.

Symbol or object	Definition and role
X^r, s_i^r	Source r and its ith waveform patch of length ω.
q_t^r, Q_i^r	Pointwise FGQA score and its patch average, $Q_i^r = \omega^{-1} \sum_t q_t^r$.
m_i^r	Binary target indicator; $m_i^r = 1$ when $Q_i^r < \theta$.
$\mathcal{V}, \mathcal{M}$	Visible and masked index sets supplied to the shared encoder and decoders.
$\bar{z}_i^r$	Embedded waveform token or learned mask token with positional encoding.
$Z_{\mathcal{V}}$	Shared temporal and cross-source representation computed from visible tokens.
d_r	Source-specific decoder that predicts $\hat{s}_i^r$ only at masked positions.
$\mathcal{L}_t, \mathcal{L}_c, \mathcal{L}_f$	Temporal reconstruction, contrastive, and log-magnitude spectral terms.

The matched-ratio experiments calibrate θ on the held-out calibration split so that the policy masks 38.5% of PPG patches. The unconstrained threshold sweep is reported separately and is not used to change the matched-ratio comparison. In the DM-Rand control, the quality scores are removed and the same mask ratio is obtained by sampling from the highest difficulty bin defined by normalized local variance, sample entropy, and log high-frequency energy.

The four FGQA stages are kept explicit below because the QA estimator supplies the reliability signal used by the masking policy. The table records the feature-map resolutions and the block types, while the stage ablation in Section B reports their measured contribution.

Algorithm S1 One QRAFT pretraining step.

Require: Synchronized source segments $\{X^r\}_{r\in\mathcal{R}}$, FGQA estimators g^r_ψ, threshold θ, patch length ω

Ensure: Updated encoder, decoder, and objective parameters

1: **for all** $r \in \mathcal{R}$ **do**
2: Partition X^r into patches s^r_i and compute $q^r_t \leftarrow g^r_\psi(X^r)_t$
3: Aggregate $Q^r_i \leftarrow \omega^{-1}\sum_{t\in i} q^r_t$ and set $m^r_i \leftarrow \mathbb{1}(Q^r_i < \theta)$
4: Form $\bar{z}^r_i \leftarrow (1-m^r_i)e^r_i + m^r_i(t^r_{\text{mask}} + p^r_i)$
5: **end for**
6: Concatenate visible and mask tokens from all available sources and apply the shared temporal encoder
7: Apply cross-source attention only to the resulting token sequence; retain source and temporal positions
8: **for all** (r, i) with $m^r_i = 1$ **do**
9: Predict $\hat{s}^r_i \leftarrow d_r(Z_\mathcal{V}, p^r_i)$ with the source-specific decoder
10: **end for**
11: Compute $\mathcal{L}_t$, $\mathcal{L}_c$, and $\mathcal{L}_f$ on masked targets and augmented views
12: Update parameters using $\mathcal{L} = \mathcal{L}_t + \lambda_1\mathcal{L}_c + \lambda_2\mathcal{L}_f$

Table S2: Fixed configuration and evaluation protocol.

Component	Fixed choice or protocol
Sampling and windows	125 Hz resampling; ten-second segments; 0.5 s fine-grained QA instances.
Pretraining pool	VitalDB, 1,000 subjects; ECG and PPG are QRAFT inputs, while arterial BP is not an encoder input.
Matched masking	$\theta = 0.7$ and 38.5% masked PPG patches; all compared policies use the same ratio.
Patch encoder	Convolution branches with kernel sizes $\{1, 3, 5, 7\}$, amplitude projection, and fused positional encoding.
Fusion path	Visible ECG and PPG tokens are concatenated before shared temporal and cross-source attention.
Reconstruction heads	Lightweight source-specific decoders reconstruct only the withheld patches.
Loss weights	$\alpha = 0.75$, $\beta = 0.25$, $\lambda_1 = 0.2$, and $\lambda_2 = 0.1$.
Transfer protocol	Subject-independent transfer; linear evaluation and end-to-end fine-tuning are reported separately.
Statistics	Ten-seed primary comparisons use exact two-sided paired Wilcoxon tests; five-run transfer tables are descriptive.

Table S3: Stage-wise specification of the FGQA estimator.

Stage	Output resolution	Operations retained in the implementation
Stem	$[256, 250, 64]$	Conv1D stem with batch normalization and amplitude projection.
I	$[256, 250, 64]$	Local Conv1D block followed by a mixed-convolution attention block.
II	$[256, 125, 128]$	Strided Conv1D downsampling, channel expansion, and multi-scale aggregation.
III	$[256, 63, 256]$	Multi-scale feature extraction followed by self-attention for longer temporal context.
IV	$[256, 32, 512]$	Global self-attention, adaptive average pooling, flattening, and sigmoid quality output.

Table S4: Dataset roles and evaluation units.

Dataset or subset	Role	Primary unit	Reported use
VitalDB	Self-supervised pretraining (ECG/PPG input)	Subject and ten-second segment	Encoder pretraining
MIMIC-III	Main synchronized ECG–PPG evaluation	Subject-independent segment and run	Reconstruction, BP, and fusion
UCI	Cross-dataset transfer	Patient-independent record and run	BP transfer
IHC	Feasibility transfer cohort	Run-level aggregate for inference	Descriptive BP transfer
Auxiliary tasks	Encoder transfer under random-mask pretraining	Dataset-specific split and run	Classification accuracy and F1

B. Quality Estimation and Calibration

Fig. S1 reports held-out coarse QA confusion matrices. PPG accuracy is 0.969 and ECG accuracy is 0.964 in the dedicated QA evaluation. Pairwise Cohen's κ ranges from 0.74 to 0.82 for PPG and from 0.71 to 0.79 for ECG; Fleiss' κ is 0.78 and 0.74. ECE is 0.046 and 0.058 before temperature scaling and 0.031 and 0.042 after scaling, while Brier score is 0.071 and 0.083. These results support use of the FGQA output as an operational reliability score, not as a clinical diagnostic probability.

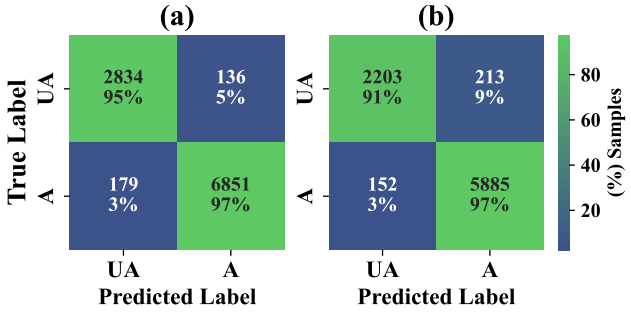
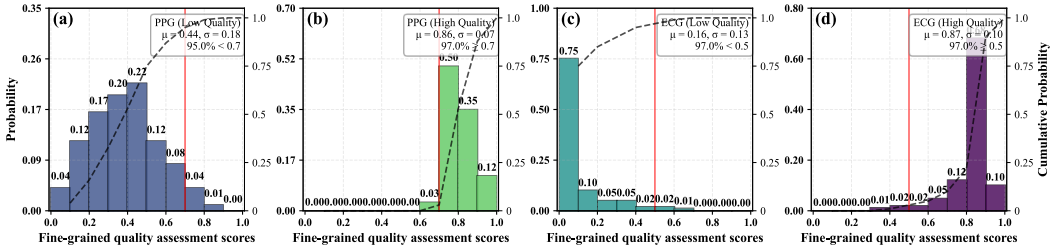


Figure S1: Held-out coarse QA confusion matrices for PPG and ECG.



Signal Type	Accuracy	Sensitivity (Class A)	Sensitivity (Class UA)	Threshold	Test Set Size
PPG	0.969	0.950	0.970	0.7	10,000 (10s segments)
ECG	0.964	0.970	0.910	0.5	10,000 (24.2% UA)

Figure S2: Fine-grained QA score distributions and threshold calibration for PPG and ECG.

The staged FGQA ablation in Table S5 separates the contribution of progressively added local, multi-scale, and global processing blocks. The table is an architectural QA control; it does not change the reconstruction or fusion objective used by QRAFT.

The baseline and lightweighting tables complete the QA evidence chain. The first pair reports modality-specific classification metrics; the second reports the measured cost of the QA variants and reference backbones under the same implementation environment.

The ECG accuracy in this baseline-comparison table is 0.957, whereas the dedicated held-out QA evaluation in Section B reports 0.964; the two values correspond to separate evaluation protocols.

Table S5: FGQA stage ablation on ECG and PPG quality assessment.

Structure	ECG			PPG		
	ACC	PPV	F1	ACC	PPV	F1
Stage I	0.87	0.91	0.91	0.91	0.93	0.94
Stage I+II	0.89	0.92	0.92	0.93	0.94	0.95
Stage I+II+III	0.94	0.96	0.95	0.95	0.98	0.96
Stage I+II+III+IV	**0.97**	**0.97**	**0.97**	**0.97**	**0.98**	**0.98**

Table S6: FGQA baseline comparison on ECG quality assessment.

Model	ACC	Spec. A	Sens. A	PPV A	F1 A
ViT-small	0.913	0.810	0.955	0.926	0.940
ViT-tiny	0.875	0.633	0.972	0.869	0.917
SWT-small	0.858	0.862	0.857	0.934	0.896
SWT-tiny	0.866	0.854	0.871	0.937	0.903
ResNet-50	0.917	0.836	0.950	0.918	0.957
ResNet-34	0.893	0.800	0.800	0.921	0.926
EfficientNet-B0	0.843	0.952	0.798	0.976	0.859
FGQA-Net	**0.957**	0.912	**0.978**	0.965	**0.970**

Table S7: FGQA baseline comparison on PPG quality assessment.

Model	ACC	Spec. A	Sens. A	PPV A	F1 A
ViT-small	0.930	0.808	0.977	0.928	0.952
ViT-tiny	0.926	0.808	0.973	0.927	0.950
SWT-small	0.940	0.792	0.991	0.923	0.960
SWT-tiny	0.942	0.809	0.995	0.924	0.960
ResNet-50	0.969	0.895	0.995	0.958	0.978
ResNet-34	0.961	0.952	0.964	0.979	0.972
EfficientNet-B0	0.915	0.709	0.991	0.894	0.944
FGQA-Net	**0.969**	**0.954**	0.974	**0.980**	**0.978**

Table S8: Lightweighting metrics for FGQA variants and reference models.

Model	Training (s/epoch)	Inference (s/segment)	Params (M)	FLOPs (Gmac)
ViT-base	51	1.66	85.1	1.79
ViT-small	34	1.12	54.1	1.19
ViT-tiny	17	0.38	28.4	0.60
SWT-small	21	0.52	45.8	0.74
SWT-tiny	16	0.29	25.5	0.34
ResNet-50	36	0.28	15.9	0.79
ResNet-34	15	0.21	7.2	0.35
EfficientNet-B0	16	0.24	3.7	0.13
MACN	61	0.43	8.2	2.59
FGQA-S	19	0.33	21.0	0.06
FGQA-T	14	0.23	10.7	0.03
FGQA-M	10	0.21	7.2	0.02
FGQA-N	8	0.18	3.6	0.01

Across the two modality-specific comparisons, FGQA-Net gives the strongest ECG accuracy and F1 and matches the best PPG accuracy and F1. The lightweight variants reduce parameter and compute costs progressively; FGQA-N has 3.6M parameters and 0.01 GMac, providing a compact operating point for quality estimation. These results separate the accuracy of the full estimator from the deployment trade-off of its reduced variants.

C. Extended Reconstruction Analysis

The threshold sweep shows that the unconstrained optimum is $\theta = 0.8$, with mask ratio 45.3% and RMSE 0.61. The matched-ratio study uses $\theta = 0.7$ to compare policies at the same 38.5% ratio. Fig. S3 visualizes mask placement, while Fig. S6 summarizes robustness under Gaussian noise, motion artifact, baseline drift, and mixed corruption. The quality-conditioned policy retains its largest advantage under severe local corruption.

Table S9 records the complete threshold sweep. It distinguishes the unconstrained operating point from the ratio-matched value used for the main masking comparison; the latter is retained to keep the competing policies comparable.

Table S9: Effect of quality-threshold selection on PPG reconstruction with reliability-conditioned masking.

θ	Mask ratio	RMSE↓	PRD (%)↓	SNR (dB)↑	R^2 ↑
0.9	75.7%	0.67±0.07	15.5±0.70	16.2±0.48	0.88±0.03
0.8	**45.3%**	**0.61±0.08**	**14.5±0.68**	**16.8±0.43**	**0.90±0.02**
0.7	38.5%	0.63±0.10	14.9±0.89	16.5±0.50	0.89±0.02
0.6	35.3%	0.67±0.15	17.2±0.90	15.9±0.50	0.87±0.01
0.5	29.1%	0.75±0.20	15.5±0.69	15.2±0.52	0.86±0.02

Table S10 makes the accuracy–efficiency trade-off explicit. The 0.5 s setting is not the most accurate configuration in isolation; it is the selected operating point because it retains near-best reconstruction while reducing the relative calculation and memory cost of the finest setting.

Table S10: Reconstruction and relative efficiency across mask granularities.

Granularity	Mask	RMSE↓	PRD (%)↓	SNR (dB)↑	Compute	Memory
0.125 s	44.8%	0.58±0.09	14.0±0.55	17.4±0.64	1.00×	1.00×
0.25 s	44.5%	0.60±0.09	14.6±0.60	17.0±0.59	0.60×	0.40×
0.5 s	**45.3%**	**0.61±0.08**	**14.5±0.68**	**16.8±0.43**	**0.42×**	**0.28×**
1.0 s	46.8%	0.75±0.15	17.2±0.77	15.1±0.50	0.31×	0.24×
2.0 s	48.9%	0.97±0.22	20.2±1.05	14.5±0.40	0.18×	0.11×

Compute and memory are relative to the 0.125 s setting.

The corruption controls in Table S11 complement Fig. S6 by reporting the paired values rather than only the relative trend. Reliability-conditioned masking improves RMSE at both tested severities for Gaussian noise, motion artifacts, and baseline drift, and also under the mixed corruption condition.

Table S11: Reconstruction RMSE under tested noise conditions.

Corruption	Intensity	Random	Reliability-conditioned	Improvement (%)
Gaussian noise	SNR 20 dB	0.61±0.22	0.52±0.12	14.8
Gaussian noise	SNR 10 dB	0.97±0.35	0.78±0.26	19.6
Motion artifacts	Low	0.69±0.25	0.61±0.19	11.6
Motion artifacts	High	1.13±0.41	0.87±0.20	23.0
Baseline drift	Low	0.48±0.10	0.43±0.09	10.4
Baseline drift	High	0.86±0.19	0.67±0.21	22.1
Mixed noise	–	0.95±0.23	0.74±0.31	22.1

The extended reconstruction analyses cover feature preservation, bidirectional cross-modal reconstruction, and the pretraining–fusion decomposition. Each result is reported with its corresponding control and evaluation unit.

The feature-level analysis shows that reliability-conditioned masked prediction preserves systolic and diastolic peaks while reducing heart-rate error and retaining waveform integrity. These readouts complement RMSE because they test whether the predicted waveform retains structures used by physiological downstream tasks.

The cross-modal table evaluates whether synchronized context from one source supports reconstruction in the other. ECG and PPG both benefit from the paired input, with gains in RMSE and PRD accompanied by higher spectral SNR. The table therefore tests bidirectional support within the evaluated synchronization setting.

Table S15 expands the objective ablation beyond the compact main-paper summary. It uses the same matched mask policy as the main masking comparison ($\theta = 0.7$, 38.5% masked patches),

Table S12: Preservation of physiological features under reconstruction.

Feature	Raw	Random	Contiguous	QRAFT	PI (%)
Systolic peak (%)	100.0	91.7±0.78	86.5±1.00	**95.2±0.28**	3.8
Diastolic peak (%)	100.0	89.5±1.03	82.8±1.19	**94.1±0.37**	5.1
HR error (bpm)	0.00	2.42±0.08	3.87±0.10	**1.35±0.05**	44.2
Integrity (0–1)	1.00	0.76±0.08	0.69±0.18	**0.89±0.05**	17.1

NOTE. PI is the percentage improvement of QRAFT over the best non-QRAFT mask. Raw is the unmasked reference and is excluded from PI.

Table S13: Cross-modal reconstruction on MIMIC-III within QRAFT.

Method	ECG reconstruction			PPG reconstruction		
	RMSE↓	PRD↓	SNR↑	RMSE↓	PRD↓	SNR↑
ECG only	0.76	20.7	22.4	–	–	–
PPG only	–	–	–	0.61	14.5	16.8
ECG + PPG	**0.64**	**16.9**	**23.9**	**0.52**	**12.6**	**19.5**
Relative / absolute gain	15.8%	18.4%	+1.5 dB	14.8%	13.1%	+2.7 dB

but summarizes the component ablation over five runs, whereas the main-paper Table 2 reports the masking-policy comparison over ten seeds. The matching RMSE means are therefore not a second estimate of the same table row; the PRD and SNR values reflect the separate five-run ablation aggregate. The reconstruction block shows that the temporal term is necessary for waveform fidelity, while the downstream block shows the complementary value of the contrastive and spectral terms under fine-tuning.

The auxiliary coefficient sweep over $\lambda_1, \lambda_2 \in \{0.05, 0.1, 0.2, 0.5, 1.0\}$ gives PPG reconstruction RMSEs from 0.61 to 0.70, with the selected point at $\lambda_1 = 0.2$ and $\lambda_2 = 0.1$; the reported setting is therefore not a single isolated optimum.

The factorial table isolates the two sources of improvement under a fixed downstream protocol. Changing the fuser while holding pretraining fixed improves the random-mask encoder, and changing pretraining while holding the cross-attention fuser fixed also helps. The best result occurs when both reliability-conditioned target selection and the QRAFT fusion rule are used together.

These diagnostics make the masking decision inspectable beyond the aggregate reconstruction scores. Fig. S3 compares where the policies place targets, Fig. S4 varies the quality threshold and temporal granularity, and Fig. S5 examines the corresponding spectral and reconstruction behavior. Read together, the figures connect the selected operating point to target placement, computational resolution, and the structure of the predicted signal.

Table S14: Objective and pretraining–fusion decomposition.

Loss variant	Recon. RMSE	SBP MAE
Reconstruction only	0.72	4.68
Reconstruction + frequency	0.66	4.45
Reconstruction + contrastive	0.68	4.27
Frequency + contrastive	0.75	4.51
Full objective	**0.63**	**4.09**
Pretraining / fuser	Cross-attention	QRAFT fuser
Random-mask pretraining	3.78	3.56
QRAFT pretraining	3.65	**3.36**

The factorial block separates the contribution of pretraining from that of the fusion rule under a fixed downstream protocol.

Table S15: Component-wise TSR-Loss ablation with the matched mask policy ($\theta = 0.7$, 38.5% mask ratio) and the corresponding MIMIC-III SBP fine-tuning protocol. Values are descriptive mean±std across five runs.

PPG reconstruction				
Variant	RMSE↓	PRD↓	SNR↑	R^2 ↑
Reconstruction only	0.72±0.10	17.1±1.02	15.4±0.55	0.84
Reconstruction + frequency	0.66±0.09	15.6±0.92	16.0±0.48	0.86
Reconstruction + contrastive	0.68±0.10	16.0±0.95	15.8±0.50	0.85
Frequency + contrastive	0.75±0.12	17.8±1.10	14.9±0.55	0.83
Full TSR-Loss	**0.63±0.08**	**14.5±0.68**	**16.8±0.43**	**0.90**

MIMIC-III SBP fine-tuning		
Variant	MAE↓	R^2 ↑
Reconstruction only	4.68±0.20	0.89
Reconstruction + frequency	4.45±0.18	0.90
Reconstruction + contrastive	4.27±0.17	0.91
Frequency + contrastive	4.51±0.19	0.90
Full TSR-Loss	**4.09±0.17**	**0.93**

Table S16: Pretraining and fusion-rule decomposition on MIMIC-III SBP fine-tuning.

Pretraining / fusion	Cross-Att	QRAFT-FUSE	Δ_{FUSE}
Random-mask PT	3.78±0.13	3.56±0.12	−0.22 (−5.8%)
QRAFT PT	3.65±0.12	**3.36±0.10**	−0.29 (−7.9%)
Δ_{PT}	−0.13 (−3.4%)	−0.20 (−5.6%)	–

NOTE. Each entry is SBP MAE in mmHg. The four cells share encoder capacity, decoder, batch schedule, and labeled fine-tuning data.

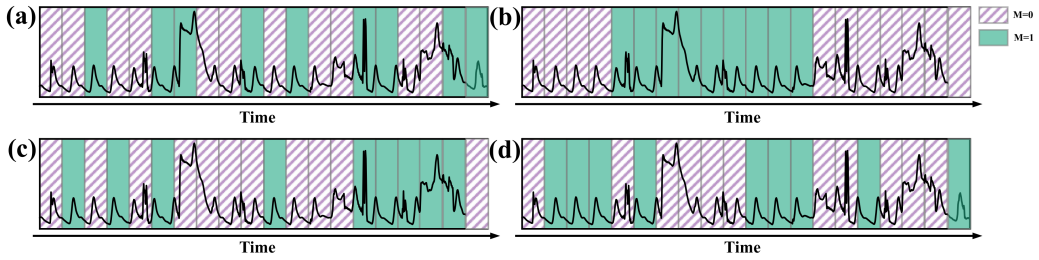


Figure S3: Mask placement on a shared PPG segment for the compared policies.

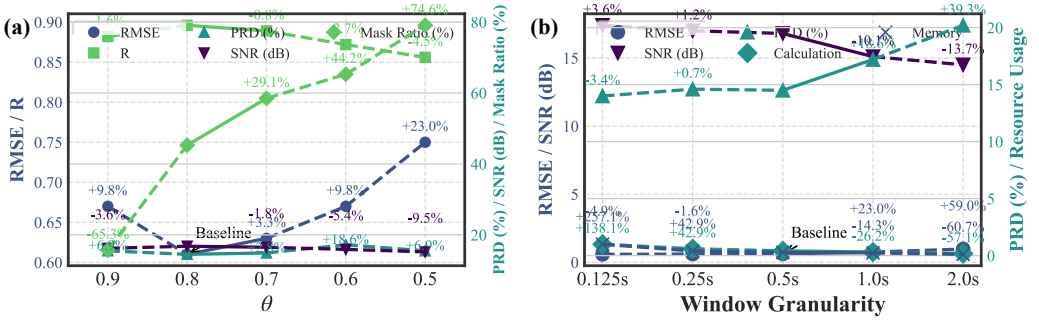


Figure S4: Threshold and mask-granularity sensitivity.

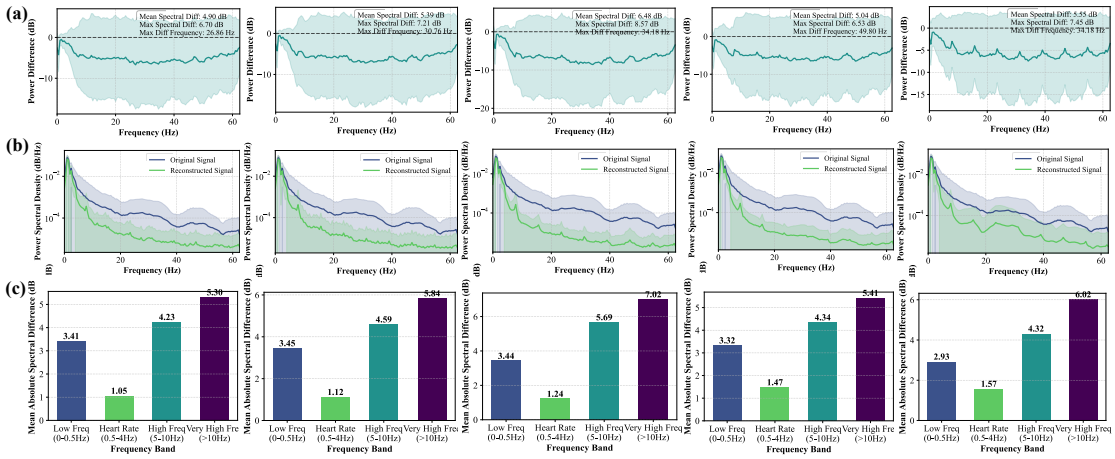


Figure S5: Spectral and reconstruction analysis across mask granularities.

Figure S5 gives complementary frequency- and reconstruction-domain views of the masking choices. The spectral panels show how the residual structure changes across frequency bands, while the reconstruction panels track prediction error as mask granularity varies. Together, these views indicate that the matched masking comparison is sensitive to both waveform detail and spectral shape; the time-domain RMSE alone does not describe this behavior completely.

Across Gaussian noise, motion artifacts, baseline drift, and mixed corruption, reliability-conditioned masking improves reconstruction over random masking at every tested severity. The relative reduction is 10.4–14.8% under mild corruption and 19.6–23.0% under severe corruption, with the largest gains under motion and baseline disturbances.

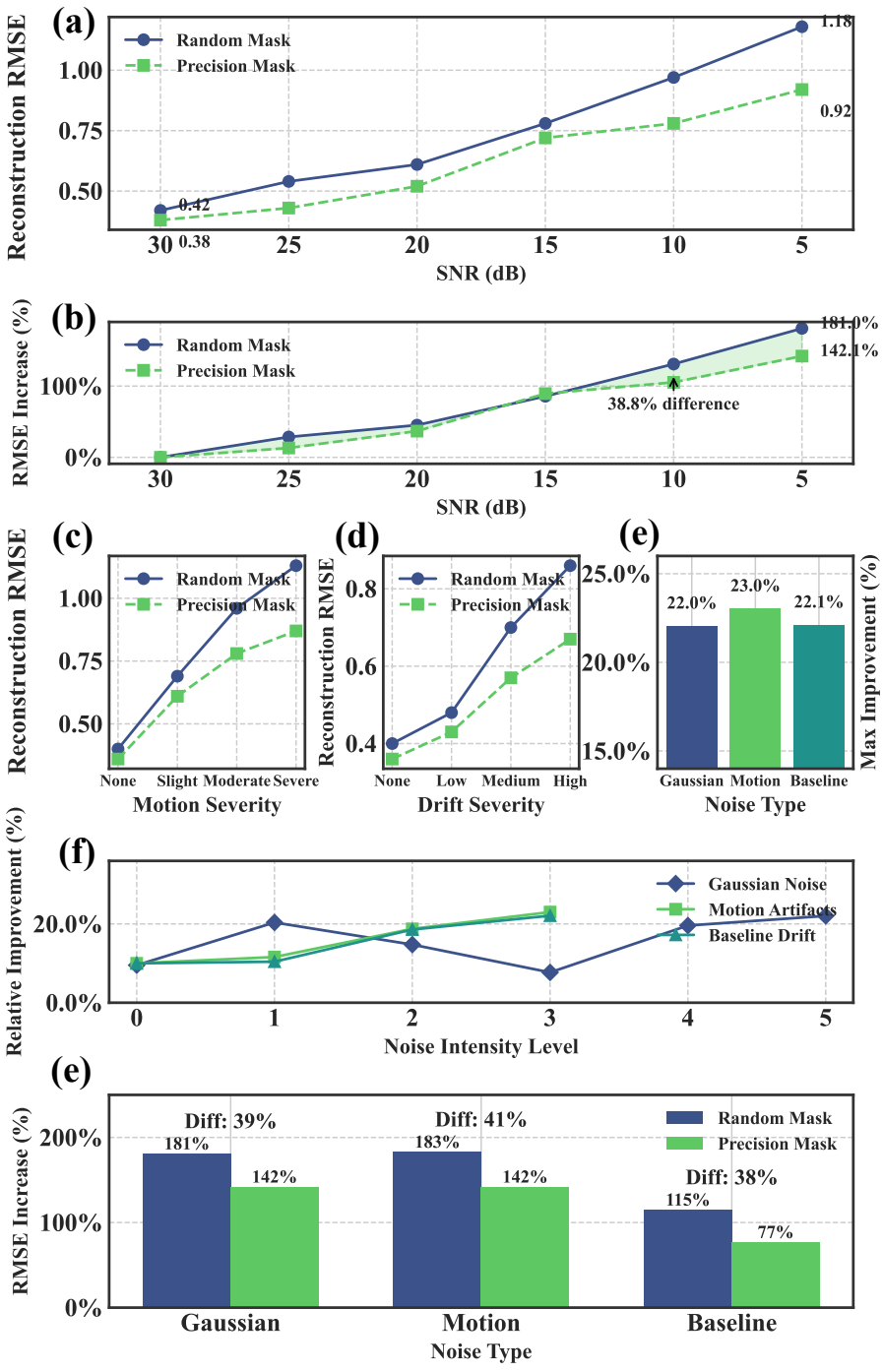


Figure S6: Noise robustness under signal corruption.

D. BLOOD-PRESSURE TRANSFER DIAGNOSTICS

Table S20 reports descriptive BP-transfer diagnostics for MIMIC-III, UCI, and IHC. The IHC cohort is treated as a feasibility evaluation. Fig. S7 provides representative regression and Bland–Altman plots. The SBP limits of agreement are approximately ± 9.82 mmHg; these plots are descriptive and do not establish compliance with AAMI, BHS, or IEEE device-validation standards.

The quality-stratified analysis below provides detailed support for the main-text fusion result. The relative gain increases as either source becomes unreliable and is largest in the L/L regime. Computed on synchronized MIMIC-III segments, the table provides regime-level evidence for complementarity.

Table S17: Quality-stratified SBP MAE on MIMIC-III.

Regime	Share	ECG	PPG	Fusion	Gain
H/H	41.2%	4.42	4.18	3.95	5.5%
H/L	19.6%	4.58	6.92	4.02	12.2%
L/H	22.3%	6.74	4.36	3.88	11.0%
L/L	16.9%	7.31	7.15	5.82	18.6%
Overall	100%	5.41	4.99	**4.09**	**18.0%**

The two tables below provide detailed BP controls. They separate the effect of explicit quality perception from the effect of reconstruction and show whether the ranking is stable across demographic and pressure strata.

Table S18: Impact of quality assessment, masking, and reconstruction on BP estimation.

Method	MAE↓	ME±SD↓	R^2 ↑
	SBP		
Coarse-grained	5.93	-1.15 ± 7.63	0.84
No quality perception	7.21	-2.43 ± 9.42	0.73
Without reconstruction	6.37	-1.82 ± 8.25	0.80
QRAFT	**4.09**	-0.57 ± 5.20	**0.93**
	MAP		
Coarse-grained	4.32	-0.95 ± 6.18	0.88
No quality perception	5.56	-1.82 ± 7.75	0.81
Without reconstruction	4.78	-1.35 ± 6.83	0.85
QRAFT	**3.12**	-0.42 ± 4.35	**0.93**
	DBP		
Coarse-grained	3.85	0.78 ± 5.27	0.87
No quality perception	4.76	1.42 ± 6.35	0.79
Without reconstruction	4.13	0.95 ± 5.74	0.84
QRAFT	**2.63**	-0.32 ± 3.95	**0.94**

Table S19: Descriptive BP MAE across evaluated population and pressure strata.

Group	Comparison I	Comparison II	Comparison III	QRAFT
		Age group		
18–40	5.72±0.63	6.95±0.70	6.12±0.56	3.96±0.33
41–60	5.89±0.60	7.13±0.77	6.28±0.59	4.03±0.42
>60	6.21±0.54	7.57±0.80	6.73±0.66	4.31±0.30
		Gender group		
Male	5.85±0.66	7.09±0.80	6.25±0.45	4.01±0.38
Female	6.02±0.70	7.34±0.82	6.49±0.44	4.17±0.40
		BP group		
$< 90/60$	6.15±0.55	7.48±0.78	6.67±0.50	4.24±0.39
90–140/60–90	5.74±0.38	6.98±0.60	6.15±0.49	3.92±0.24
$> 140/90$	6.05±0.56	7.31±0.70	6.41±0.45	4.12±0.26

Table S20: Descriptive BP-transfer diagnostics.

Dataset	SBP MAE	DBP MAE	MAP MAE
MIMIC-III	4.09	2.63	3.12
UCI	3.52	2.60	3.15
IHC	4.55	2.78	3.31

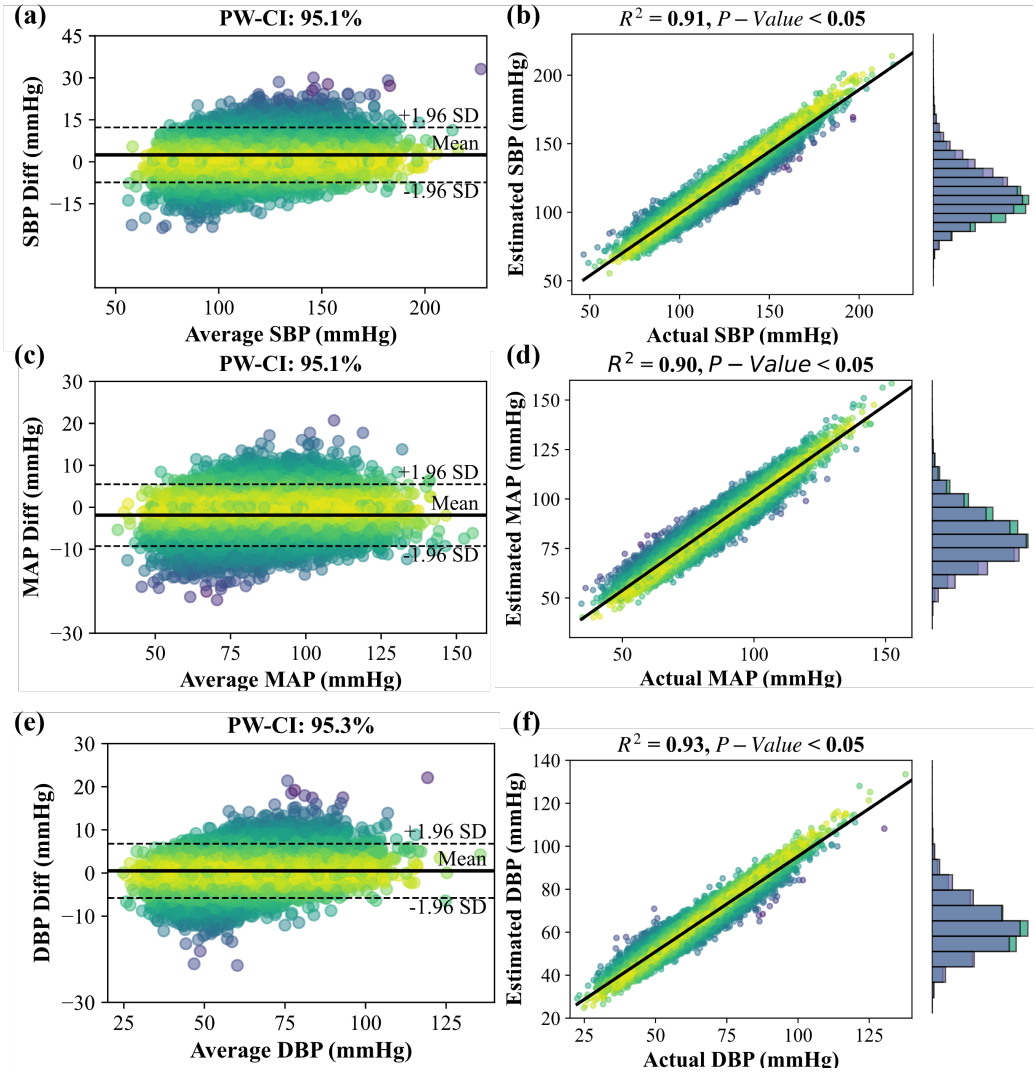


Figure S7: Representative regression and Bland–Altman diagnostics.

Table S21: Fusion-rule comparison and hard-versus-soft gating ablation.

Fusion rule	SBP	DBP	MAP
PPG-only	$4.09 \pm .15$	$2.63 \pm .07$	$3.12 \pm .10$
Early	$3.95 \pm .17$	$2.58 \pm .05$	$3.04 \pm .09$
Late	$3.83 \pm .13$	$2.52 \pm .05$	$2.96 \pm .07$
Cross-attention	$3.65 \pm .12$	$2.41 \pm .04$	$2.84 \pm .06$
Soft-QW	$3.57 \pm .11$	$2.35 \pm .04$	$2.78 \pm .06$
QRAFT	$\mathbf{3.36 \pm .10}^{\dagger}$	$\mathbf{2.22 \pm .04}^{\dagger}$	$\mathbf{2.63 \pm .05}^{\dagger}$
Gate variant	Recon. RMSE	SBP	R^2
Soft	0.68	4.31	0.91
Thresholded-soft	0.65	4.17	0.92
Hard	**0.63**	**4.09**	**0.93**

† denotes $p < 0.05$ versus Soft-QW using an exact two-sided paired Wilcoxon signed-rank test across ten seeds.

The complete gating comparison is given in Table S22. Hard gating remains the best setting on both reconstruction and downstream BP metrics, while thresholded-soft gating is intermediate.

Table S22: Complete hard-versus-soft gating ablation.

Gating variant	RMSE↓	SNR (dB)↑	SBP MAE↓	R^2 ↑
Soft ($w_i = 1 - Q_i$)	0.68±0.09	15.9±0.46	4.31±0.18	0.91
Thresholded-soft	0.65±0.08	16.4±0.44	4.17±0.16	0.92
Hard (QRAFT)	**0.63±0.08**	**16.8±0.43**	**4.09±0.17**	**0.93**

E. Auxiliary Transfer and Resource Analyses

When fine-grained quality scores are unavailable, downstream classification uses the QRAFT encoder with random-mask pretraining. This evaluates architecture transfer rather than the complete quality-conditioned mechanism. The main paper reports the corresponding classification transfer figure, while Fig. S8 summarizes the low-resource behavior as the pretraining and annotation fractions vary. The complete pretraining pool gives SBP MAE 4.09 mmHg, while performance improves most rapidly at low-to-moderate pretraining scale.

Tables S23–S26 report the complete classification results. Linear evaluation and fine-tuning are separated into distinct tables so that the values remain readable. Keeping these results in the supplement avoids conflating encoder transfer under random masking with the full quality-conditioned method.

Table S23: Auxiliary classification transfer accuracy under linear evaluation.

Method	Epilepsy	HAR	ECG-SQA	Sleep-EDF	PPG-SQA
SimCLR	78.02±3.54	82.18±2.87	76.36±3.42	67.25±4.73	85.15±2.95
TS-TCC	89.27±1.62	89.54±1.38	86.42±1.47	75.19±2.84	92.31±1.35
MoCo v3	86.42±2.18	87.39±1.96	83.17±2.05	73.28±3.17	90.35±1.82
MAE	80.24±2.47	83.35±2.35	78.53±2.81	69.47±3.86	87.26±2.53
QRAFT	**92.13±0.87**	**91.25±0.93**	**88.37±0.76**	**77.43±1.72**	**94.28±0.95**

Table S24: Auxiliary classification transfer accuracy under fine-tuning.

Method	Epilepsy	HAR	ECG-SQA	Sleep-EDF	PPG-SQA
SimCLR	85.37±2.76	84.24±2.45	86.43±2.63	74.32±3.84	89.27±2.18
TS-TCC	95.32±1.28	91.25±1.42	94.17±1.35	81.28±2.16	96.43±1.21
MoCo v3	93.24±1.75	89.47±1.93	92.36±1.78	79.43±2.65	94.35±1.43
MAE	87.43±2.34	85.68±2.17	88.25±2.42	76.18±3.27	91.36±1.93
QRAFT	**97.25±0.82**	**92.37±0.89**	**95.18±0.92**	**83.24±1.64**	**98.17±0.78**

Table S25: Auxiliary classification transfer F1 under linear evaluation.

Method	Epilepsy	HAR	ECG-SQA	Sleep-EDF	PPG-SQA
SimCLR	72.34±3.68	80.19±3.25	79.27±3.47	54.18±4.85	82.35±3.17
TS-TCC	83.27±1.87	89.24±1.63	88.31±1.76	63.18±2.93	91.24±1.42
MoCo v3	79.43±2.34	86.17±2.12	85.28±2.25	60.29±3.35	88.19±1.98
MAE	74.27±3.12	82.34±2.76	81.18±2.91	56.29±4.13	84.37±2.75
QRAFT	**86.35±0.92**	**91.18±0.85**	**90.27±0.88**	**65.19±1.98**	**93.24±0.97**

Table S26: Auxiliary classification transfer F1 under fine-tuning.

Method	Epilepsy	HAR	ECG-SQA	Sleep-EDF	PPG-SQA
SimCLR	82.26±2.83	83.15±2.64	83.42±2.73	57.34±3.92	87.19±2.34
TS-TCC	93.32±1.35	91.19±1.53	91.28±1.47	66.27±2.32	95.32±1.26
MoCo v3	90.37±1.85	89.26±1.82	89.42±1.93	63.25±2.86	93.18±1.57
MAE	84.32±2.46	85.24±2.32	85.19±2.54	59.28±3.45	89.26±2.08
QRAFT	**96.28±0.76**	**92.24±0.83**	**93.17±0.87**	**68.15±1.85**	**97.23±0.82**

QRAFT leads both linear evaluation and fine-tuning on all five auxiliary tasks for accuracy and F1. The consistent ordering across the two evaluation modes indicates that the transferred encoder features remain useful after task-specific adaptation, while the random-mask protocol keeps this result separate from the quality-conditioned masking and fusion claims.

Table S27 reports the complete low-resource sweep. The largest marginal gains occur between 1% and 25% pretraining data, after which the curves approach the full-pool result. This supports the resource analysis without changing the main fusion comparison.

Table S27: Effect of pretraining scale on downstream BP estimation.

Pretraining	SBP		DBP		MAP	
ratio	MAE↓	R^2 ↑	MAE↓	R^2 ↑	MAE↓	R^2 ↑
1%	8.83	0.70	5.76	0.81	6.31	0.78
5%	6.97	0.82	4.51	0.85	4.83	0.88
10%	5.82	0.86	3.70	0.88	4.00	0.90
25%	4.96	0.87	3.02	0.91	3.64	0.91
50%	4.37	0.91	2.78	0.94	3.31	0.93
75%	4.10	0.93	2.60	0.94	3.15	0.93
100%	4.09	0.93	2.63	0.94	3.12	0.93

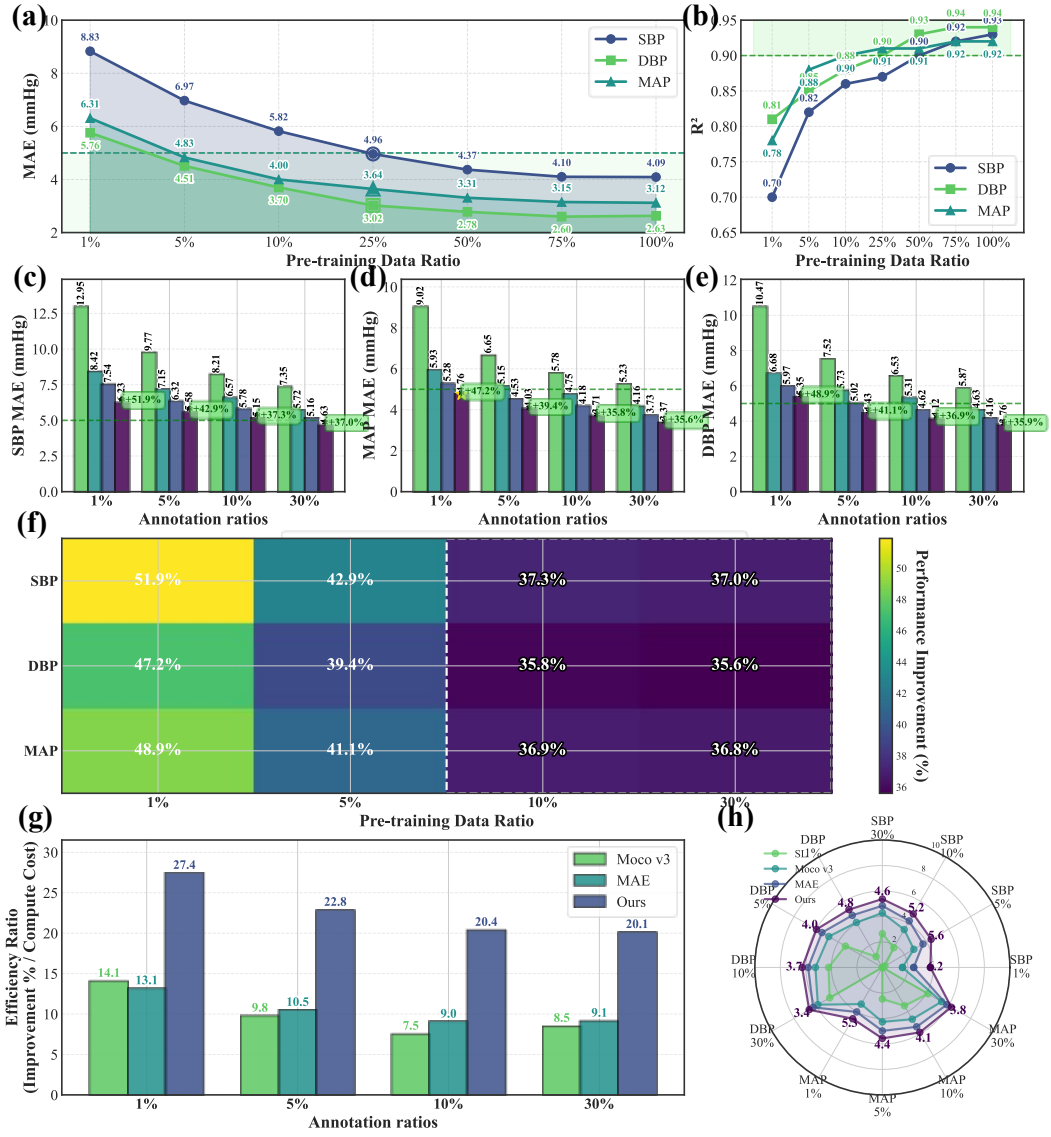


Figure S8: Low-resource pretraining and annotation analysis.

F. Robustness and Sensitivity

The matched comparison fixes the mask ratio at 38.5% and uses $\theta = 0.7$ so that every policy withholds the same amount of input. When the ratio is allowed to vary, the best reconstruction point occurs at $\theta = 0.8$ (45.3% masked, RMSE 0.61), compared with 0.63 at the matched operating point. Mask granularity gives the expected accuracy–cost trade-off: 0.125 s patches reach 0.58 RMSE at the highest cost, whereas the 0.5 s setting reaches 0.61 RMSE with $0.42\times$ the computation and $0.28\times$ the memory of the finest setting. The 0.5 s setting is therefore used in the main experiments.

Temperature scaling reduces ECE from 0.046 to 0.031 for PPG and from 0.058 to 0.042 for ECG without changing binary accuracy. Perturbing annotation agreement over $\kappa \in [0.6, 0.9]$ yields reconstruction RMSE in the 0.60–0.71 range, and a distilled estimator with 92.3% QA accuracy still outperforms random masking (0.75 versus 0.82 RMSE). Performance changes gradually as QA quality varies.

G. Prototype Description

The host-side prototype provides an implementation feasibility demonstration. It supports synchronized ECG and PPG acquisition, visualization, QA, reconstruction, BP estimation, and report generation. In the reported 30-minute streaming test, average latency was below 80 ms and the 95th-percentile latency was 112 ms.

The acquisition hardware used in the prototype is summarized in Table S28. The device combines ECG acquisition, optical PPG sensing, inertial monitoring, low-power Bluetooth streaming, and battery management. Table S29 lists the MCU specifications relevant to the reported host-side feasibility test.

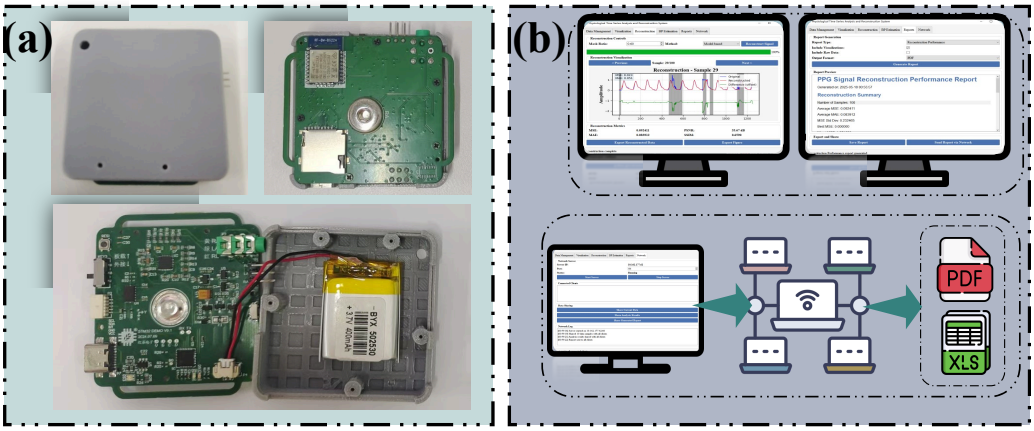


Figure S9: Host-side prototype for synchronized acquisition and analysis.

Table S28: Components of the multimodal synchronous acquisition device.

ID	Device	Manufacturer	Function
1	STM32L442KC	STMicroelectronics	MCU
2	AD8232	Analog Devices Inc.	ECG ADC
3	MAX30102	Maxim Integrated	PPG sensor
4	RF-BM-BG22A	RF-star	Low-power Bluetooth
5	TP4057	UMW	Lithium-battery charger
6	ADXL345BCCZ	Analog Devices Inc.	Accelerometer

Table S29: Technical parameters of the STM32L442KC MCU.

Index	Specification
Core	ARM Cortex-M4
Program memory	256 kB
Bus width	32 bits
ADC	2×12 bit
Maximum clock frequency	80 MHz
Number of inputs/outputs	26 I/O
Supply voltage	1.71–3.60 V
Streaming duration	30 min
Mean processing latency	< 80 ms
95th-percentile latency	112 ms

The streaming test measures processing latency and report generation under synchronized acquisition, complementing the representation-learning evaluation on the reported datasets.

H. Dataset Distribution Context

Figure S10 summarizes the SBP and DBP distributions of the pretraining and downstream cohorts, providing context for cross-dataset transfer. The cohorts differ in pressure range and density, so the transfer results reflect adaptation across related but non-identical distributions rather than interpolation within a single pooled population. The figure is descriptive and is included to make this distributional context explicit.

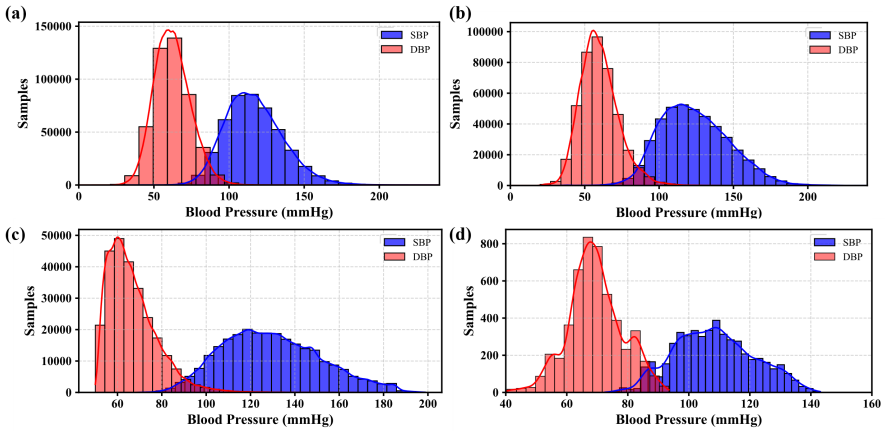


Figure S10: Descriptive SBP and DBP distributions for the evaluated BP datasets.