

EgoHRV: Continuous Heart Rate Variability Estimation from Egocentric Systems for Autonomic Response and Skill Assessment

Berken Utku Demirel[✉] and Christian Holz[✉]

Department of Computer Science, ETH Zürich, Zürich, Switzerland

<https://siplab.org/projects/EgoHRV>

Abstract. Egocentric vision systems capture human behavior from visible cues, but overlook physiological indicators of autonomic states such as stress, engagement, and attention. Heart rate variability (HRV) is a widely used noninvasive marker of autonomic regulation under stress. HRV reflects small timing differences between successive heartbeats and has so far been out of reach for egocentric platforms, where motion and noise in gaze video mask exactly this fine-grained timing. We propose *EgoHRV*, a method that estimates HRV as well as heart rate (HR) from the gaze cameras that are already integrated into egocentric headsets. Our pipeline combines a 3D backbone with a novel low-high decomposition module that extracts the blood volume pulse (BVP) signal from gaze video. Our cross-domain pretraining aligns the frequency-domain representations of contact-based and camera-derived signals. This alignment gives EgoHRV the temporal precision to recover HRV from the subtle fluctuations in gaze video. EgoHRV achieves state-of-the-art accuracy for HR and HRV estimation from egocentric video, and its uncertainty-aware design improves downstream behavioral modeling. Integrating our HRV estimates and confidence measures into EgoExo4D’s proficiency estimator raises accuracy by 17.8%. Beyond skill, continuous HRV estimation also opens egocentric systems to stress- and arousal-aware estimation tasks. Code: <https://github.com/eth-siplab/EgoHRV>

Keywords: Egocentric vision · heart rate · heart rate variability · stress

1 Introduction

Egocentric vision systems, such as Magic Leap [1] or Project Aria [20], can aid in understanding human behavior and interaction from a first-person perspective [26]. These wearable devices capture both the surrounding environment and the wearer’s actions by providing rich multimodal data [26, 27] for studying activity and social interaction [77]. Recent work has analyzed egocentric video to reveal behavioral patterns in daily tasks that static third-person systems often miss [13, 14, 27]. As their sensing capabilities expand to eye tracking and inertial cues, egocentric platforms can now enable research on context-aware behavior modeling and human sensing more broadly [35, 55, 90].

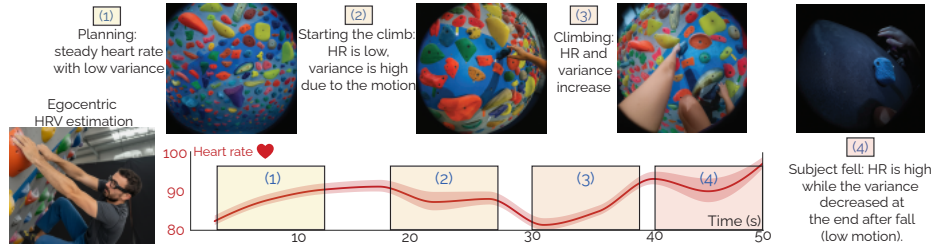


Fig. 1: EgoHRV continuously estimates heart rate as well as heart rate variability with uncertainty from gaze cameras in egocentric systems to enrich understanding of human behavior, cognitive effort, stress, and skill assessment.

Physiological indicators play a central role in modeling human behavior [32, 33, 76]. While a person’s motion or gaze can be directly observed by vision models [58], physiological signals are subtle or invisible, yet they reveal latent human autonomic [67] and affective [49] processes. A common measure of physiological state is heart rate (HR), which reflects overall autonomic activity [50] and offers insight into how people respond to tasks [48], environments [71], and social contexts [8]. While recent work showed that HR can be recovered from eye-gaze cameras on egocentric systems (egoPPG [5]), estimates have coarse temporal resolution (60s averages) and cannot capture a person’s short-term responses.

More than mere *rate*, heart rate *variability* (HRV) captures autonomic regulation and is thus a rich noninvasive marker of stress [9, 31, 37, 63, 80], cognitive effort [45], and emotional reactivity [70, 86]. Unlike HR, HRV captures the relation between sympathetic and parasympathetic branches of the autonomic nervous system [50, 70, 79]. This modulation thus helps link moment-to-moment physiological changes to perception, decision-making, and human behavior [23, 36, 53]. HRV also reveals mental workload during complex tasks, attention in learning or driving, and stress adaptation in natural environments [16, 40, 83]. In situated interactions, moments of fear, frustration, time pressure, and safety-critical decisions produce autonomic responses that manifest in HRV decreases alongside increased electrodermal activity [25, 47, 74]. Continuous tracking of HRV on egocentric systems can thus move research in this domain beyond the current focus on action and activity tracking, because the same signals capture indicators of stress, engagement, or fatigue in real time and supports downstream tasks such as proficiency assessment and stress-aware interaction.

However, accurately decoding HRV from video is challenging because it requires beat timing at a millisecond level, as even small timing errors substantially distort HRV estimates. This makes it highly sensitive to motion and noise common in mobile scenarios, in which egocentric systems are used. Additional factors such as head movement and posture changes further degrade signal quality [64, 75]. As we show in this paper, methods designed for HR estimation from video fail to generalize to estimating HRV because of the key requirement to preserve temporal structure rather than just the overall trend [57, 66].

In this paper, we introduce a method for estimating continuous HR and HRV as a basis for stress- and skill-aware modeling in egocentric systems. We demonstrate its benefits for behavioral understanding in egocentric vision through the proficiency estimation benchmark task. Since HRV is an established noninvasive proxy for autonomic stress responses [37, 80], our continuous estimates can support stress- and arousal-aware tasks on egocentric systems without additional body-worn sensors. Taken together, we contribute:

1. an uncertainty-aware HR & HRV estimator from egocentric gaze video via a 3D backbone with low-high decomposition, setting state-of-the-art results,
2. a large-scale cross-domain pretraining strategy that leverages contact-based blood volume pulse (BVP) datasets with ground truth and uses frequency-domain alignment to transfer physiologically meaningful representations to our egocentric vision approach, and
3. a multimodal PhysFusion module that fuses HRV with video representations to improve proficiency estimation on EgoExo4D by 17.8% and supply a continuous HRV signal to downstream stress- and arousal-aware applications.

2 Related Work

Egocentric vision. Egocentric vision research has rapidly expanded with the advent of wearable mixed-reality devices, e.g., Magic Leap [1] and Project Aria [20]. These platforms enable continuous first-person perception for understanding user actions, surroundings, and interactions [27, 91]. Much of the work related to these platforms has focused on visual understanding tasks, including action recognition [30, 78, 85], anticipation [24, 88], and hand-object interaction [44, 91], as well as 3D reconstruction, and mapping [28, 89].

Large-scale datasets, Ego4D [26] and EgoExo4D [27], have accelerated the research by providing thousands of hours of first-person recordings across diverse contexts and users. These datasets enabled progress on social interaction modeling and multimodal perception tasks that integrate gaze, speech, and motion [52]. However, even with these developments, physiological sensing and autonomic state estimation remain absent from existing egocentric benchmarks, leaving an opportunity to connect external behavior with internal physiological state.

Physiological sensing in vision. Wearable sensors have enabled continuous monitoring of cardiovascular and affective states in daily life [54, 60, 64]. In parallel, camera-based methods estimate BVP from subtle skin-color changes, commonly referred to as remote photoplethysmography (rPPG), caused by cardiac activity [84]. Traditional rPPG methods rely on chrominance or blind-source separation techniques [29], while more recent approaches adopt spatiotemporal networks [34, 92, 93] to improve robustness under varying illumination and skin tone. Beyond heart rate, camera-based approaches also decode further autonomic indicators, such as sympathetic arousal, which can be recovered from peripheral blood-flow changes in video [6, 7].

Despite much progress, most rPPG datasets and models assume fixed cameras and stationary subjects [7, 41, 56, 82], which limits their applicability to wearable scenarios. In egocentric systems, cameras are mounted on the head and undergo significant motion, making direct transfer of conventional rPPG methods unreliable. Recent work, egoPPG [5], addressed this limitation by estimating HR from eye-tracking videos and inertial cues. Although egoPPG demonstrated the feasibility of extracting coarse HR trends, its estimates averaged over long windows and achieved errors around 8–9 bpm, which is far below the temporal fidelity and accuracy of contact sensors [4].

Heart rate variability estimation. HRV provides a more sensitive measure of autonomic function than HR by capturing short-term fluctuations in cardiac rhythm [70]. HRV estimation from non-contact sensors remains a major challenge. Accurate HRV computation requires millisecond-level timing precision in inter-beat intervals, which is difficult to achieve from video affected by noise, motion, and low sampling rates [57, 66, 75]. In egocentric systems, additional challenges arise from variable head motion, changing illumination, and limited skin visibility, all of which disrupt consistent waveform morphology. PulseFormer, an egocentric method for HR, estimates a single HR value per 60 s [5], discarding the dynamics that characterize real-world behavior. In contrast, EgoHRV estimates short-term mean inter-beat intervals and derives HRV at this higher temporal resolution.

HRV as a proxy for perceived stress. Stress activates the autonomic nervous system, which modulates the cardiac dynamics that HRV indexes [12]. HRV widely serves as an established peripheral indicator of adaptive responses to challenge [80, 81] and is a standard input feature for biosignal-based stress classification [25, 74]. It drops under perceived stress [9, 37], including in situated and immersive settings [47]. HRV thus serves as a stress-related autonomic marker rather than a direct sympathetic readout [39, 72, 79].

3 Problem: HRV is a highly sensitive metric

HRV estimation from egocentric video is challenging due to algorithmic and data limitations. We highlight two key issues that motivate our approach.

3.1 Error Accumulation

Existing approaches for physiological estimation from vision, including models for rPPG and egoPPG tasks [5, 34, 43, 46, 92, 94], reconstruct the BVP indirectly by predicting frame-to-frame differences of skin reflectance and regressing them toward the temporal derivatives of the reference PPG signal. During inference, these predicted differences are cumulatively summed to recover the waveform. While this formulation improves stability and reduces spurious correlations with the background, it introduces a drift. Small frame-wise prediction errors accumulate over time, leading to deviations in the integrated signal. This issue is further amplified in egocentric settings, where head motion and dynamic

illumination can cause larger variations than in stationary rPPG. As a result, even minor prediction noise can distort the reconstructed waveform. Such cumulative bias is detrimental for HRV estimation, which depends on millisecond-level accuracy rather than waveform trends.

Let $\hat{d}_t$ denote the model’s predicted BVP difference at frame t , and d_t the ground-truth derivative. The reconstructed BVP is obtained by integration.

$$\hat{p}_t = \sum_{i=1}^t \hat{d}_i, \quad p_t = \sum_{i=1}^t d_i. \quad (1)$$

The instantaneous reconstruction error is given in Equation 2.

$$\epsilon_t = \hat{p}_t - p_t = \sum_{i=1}^t (\hat{d}_i - d_i), \quad (2)$$

and even if the per-frame prediction noise is zero-mean with variance σ^2 , its cumulative variance increases with time:

$$\text{Var}[\epsilon_t] = t \sigma^2. \quad (3)$$

As t grows, small local prediction errors compound, progressively distorting waveform morphology. This drift propagates directly into inter-beat intervals, producing large HRV estimation errors even when frame-wise losses remain small. Given HRV’s extreme sensitivity to timing, derivative-based reconstruction schemes become unreliable for dynamic, motion-heavy egocentric data.

3.2 Limited use of large-scale physiological datasets

Although egocentric datasets with synchronized ground-truth physiological recordings are limited, large-scale public BVP datasets already exist from mobile settings [51, 64]. These datasets capture a wide range of heart rate and waveform under varying illumination, providing valuable priors. Yet, current methods have not leveraged these broader data sources for pretraining. Learning BVP representations from such datasets and then adapting them for visual inputs from cameras can significantly improve the performance.

4 Method

Our method has two aims. First, we estimate HR and HRV directly from egocentric video. Second, we leverage these physiological estimates to improve downstream proficiency prediction in large-scale egocentric datasets. The following sections describe each component in detail.

4.1 HR & HRV Estimation

HR and HRV are both determined by the sequence of inter-beat intervals (IBIs). HR is the reciprocal of the mean IBI over a window, while HRV is the variability

of consecutive IBIs over the same window. Estimating IBIs once therefore gives both metrics without separate heads or losses. We design our pipeline to (a) recover a clean BVP waveform from video (Section 4.1) and (b) predict IBIs from that waveform using cross-domain supervision (Section 4.1); HR and HRV are then computed from the predicted IBIs.

3D backbone with low-high decomposition We adopt a 3D convolutional architecture [5, 92] that processes temporal segments of $T=128$ frames (4.3 s) at a spatial resolution of 48×128 pixels. Each input clip $\mathbf{x}$ and its output representation $\mathbf{y}$ are defined as in Equation 4.

$$\mathbf{x} \in \mathbb{R}^{T \times H \times W}, \quad \mathbf{y} = f_{3D}(\mathbf{x}) \in \mathbb{R}^{1 \times T}, \quad (4)$$

where (H, W) denote the spatial dimensions, and T is the time (frames). The representation $\mathbf{y}$ captures frame-wise reflectance differences that encode temporal variations in skin appearance. Our 3D backbone model predicts temporal differences rather than absolute intensity to emphasize dynamic changes similar to previous work [5, 34].

However, raw differencing behaves as a high-pass filter, amplifying motion noise. We aim to mitigate this effect with a learnable low-high decomposition that explicitly separates high frequency noise from pulse. We define the low-frequency component and the complementary high-frequency component as in Equation 5.

$$\mathbf{y}_{lo} = \mathbf{y} * h, \quad h = \frac{1}{k} \mathbf{1}_k, \quad k=11, \quad \mathbf{y}_{hi} = \mathbf{y} - \mathbf{y}_{lo}, \quad (5)$$

Two lightweight convolutional heads refine these signals:

$$\tilde{\mathbf{y}}_{lo} = f_{lo}(\mathbf{y}_{lo}), \quad \tilde{\mathbf{y}}_{hi} = f_{hi}(\mathbf{y}_{hi}), \quad (6)$$

and their outputs are fused through a gated residual connection with a learnable scalar (g) controlling the correction.

$$\hat{\mathbf{y}} = \mathbf{y} + \tanh(g) [(\tilde{\mathbf{y}}_{lo} + \tilde{\mathbf{y}}_{hi}) - \mathbf{y}] \quad (7)$$

This operation can be interpreted as an adaptive filter that preserves temporal variation while suppressing broadband noise before the integration operation. This component adds ~ 500 parameters (less than 0.1% of total parameters) into the model yet improves the performance. Finally, the BVP waveform $\mathbf{p}_{BVP}$ is reconstructed by integrating the temporal differences $\hat{\mathbf{y}}$ over time as in Equation 8.

$$\mathbf{p}_{BVP} = \sum_{i=1}^t \hat{\mathbf{y}}_i, \quad t = 1, \dots, T \quad (8)$$

This integration step recovers the cumulative reflectance variation corresponding to blood flow and serves as the input for the cross-domain pretraining stage.

Cross-domain pretraining The visual backbone provides a preliminary estimate of the blood volume pulse signal $\mathbf{p}$, which still contains residual noise. To improve temporal precision, we leverage large-scale physiological datasets [51, 64]

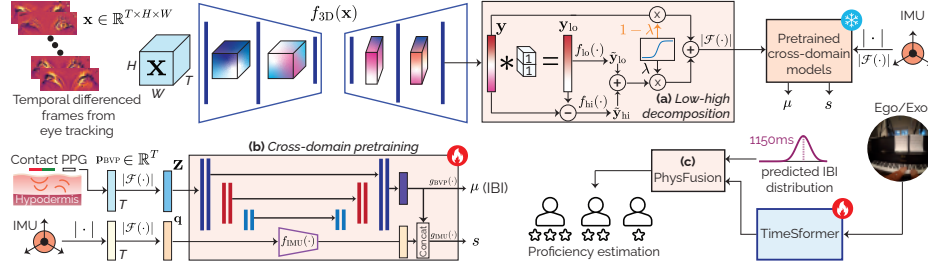


Fig. 2: EgoHRV estimates HR and HRV from egocentric gaze videos. Our method introduces three key components: (1) A low–high decomposition module integrated into the 3D backbone to suppress high-frequency noise and preserve waveform shape. (2) Cross-domain pretraining of blood volume pulse signals using large-scale physiological datasets for learning frequency-domain representations to transfer into BVP signals obtained from cameras. (3) *PhysFusion*, a multimodal fusion head that combines TimeFormer video embeddings with temporally weighted HRV for proficiency estimation in the EgoExo4D benchmark [27].

containing synchronized contact BVP and electrocardiogram (ECG) recordings. From each ECG trace, we extract reference IBIs using Pan–Tompkins [59] by detecting successive cardiac cycles. Using ECG-derived IBIs as supervision, we train a model to predict IBIs directly from camera BVP segments.

A key challenge in this cross-domain setting is the phase mismatch between contact and vision-based BVP signals caused by differences in sensor location, optical path, and propagation delay [10, 69]. To mitigate this, we transform each BVP segment into the frequency domain using the discrete Fourier transform (DFT) $\mathcal{F}(\cdot)$ as in Equation 9.

$$\mathbf{F} = \mathcal{F}(\mathbf{p}_{\text{BVP}}), \quad \mathbf{z} = |\mathbf{F}| \in \mathbb{R}^K, \quad (9)$$

and use the magnitude spectrum, $\mathbf{z}$, as input to the encoder. This representation preserves the frequency structure while discarding phase that varies across domains, which has been shown to be effective for cross-domain BVP signals [4].

The cross-domain BVP encoder, $f_{\text{BVP}}(\cdot)$, uses a three-block U-Net architecture [65], which has been shown to be effective for this task [17], to map $\mathbf{z}$ to a probabilistic IBI distribution. To further improve robustness under motion, we integrate inertial features through multimodal fusion. For each IMU segment, we first compute the magnitude of the tri-axial acceleration signal and transform it into the frequency domain using the DFT.

The resulting Fourier magnitude spectrum, which captures the dominant motion frequencies and their magnitude, is then processed by a lightweight two-layer convolutional encoder $f_{\text{IMU}}(\mathbf{q})$. After decoding the BVP spectrum, we apply global average pooling, which is concatenated with the IMU features and passed through a small MLP head predicting the log-variance s of the distribution.

$$\mu = g_{\text{BVP}}(f_{\text{BVP}}(\mathbf{z})) \quad s = g_{\text{IMU}}(f_{\text{BVP}}(\mathbf{z}), f_{\text{IMU}}(\mathbf{q})) \quad (10)$$

Detailed architectural configurations (kernel sizes, activation layers, and projection blocks) for cross-domain models are provided in the Appendix. The model is trained with a negative log-likelihood (NLL) loss as in Equation 11.

$$\mathcal{L}_{\text{NLL}} = \frac{1}{N} \sum_i \left[\frac{(\text{IBI}_i - \mu_i)^2}{2e^{s_i}} + \frac{s_i}{2} \right], \quad (11)$$

where (μ_i, s_i) denote the predicted mean and log-variance for each window. A training window spans 128 frames, aligned with reference ECG segments, supervised at the window level using the mean IBI. The mean μ provides the estimated mean IBI value, while the variance e^s captures predictive uncertainty. This multimodal fusion allows the model to adjust its uncertainty based on motion intensity, reflecting lower confidence during motion-heavy segments.

At inference, the pretrained encoder is transferred to the visual domain, where it operates on egocentric camera-derived BVP signals to obtain IBI estimates.

4.2 PhysFusion for Proficiency Estimation

We introduce *PhysFusion*, a multimodal fusion module that jointly encodes weighted HRVs and video representations for the downstream task of proficiency prediction in egocentric applications.

Given a clip $\mathbf{x}$ from eye tracking cameras, we predict user proficiency by fusing video representations with physiological features. The video is fed to a TimeSformer [3] to output a class token $\mathbf{v} \in \mathbb{R}^d$ for proficiency levels (4-class: novice, early expert, intermediate expert, late expert). From the pretrained cross-domain model, we obtain per-window IBI features (μ_t, s_t) . We convert μ_t [ms] to HR $h_t = 60,000/\mu_t$ and form the physiological feature vector $\mathbf{c}_t = [h_t, s_t]^\top \in \mathbb{R}^2$ for each 128 frames.

We build the physiological feature vector for each 128 frames (4.3 s) in the egocentric eye tracking video $\mathbf{x}$. Each step is projected with a linear layer and temporally encoded with a transformer as in Equation 12.

$$\{\mathbf{T}_t\}_{t=1}^T = \mathcal{T}_\theta(\{\mathbf{u}_t\}_{t=1}^T), \quad \mathbf{u}_t = \mathbf{W}_p \mathbf{c}_t + \mathbf{b}_p \in \mathbb{R}^{d_p}, \quad (12)$$

where $\mathbf{W}_p, \mathbf{b}_p$ are learnable linear layer, and $\mathcal{T}_\theta$ is a 2-layer, 4-head Transformer encoder (multi-head self-attention with pre-norm residual blocks). We employ a lightweight encoder to capture temporal dependencies in HRV that are lost under the simple averaging used in prior work [5]. We aggregate the sequence with learned attention as in Equation 13.

$$\mathbf{p} = \sum_{t=1}^T \alpha_t \mathbf{T}_t \in \mathbb{R}^{d_p} \quad \alpha_t = \frac{\exp(\mathbf{w}^\top \mathbf{T}_t)}{\sum_k \exp(\mathbf{w}^\top \mathbf{T}_k)} \quad (13)$$

We designed the projection for heteroscedastic weighting to reduce the contribution of HRVs with larger variance.

Finally, we concatenate the TimeSformer representations with the processed physiological features and apply a linear classification head as in Equation 14.

$$\mathbf{z} = [\mathbf{v}; \mathbf{p}] \in \mathbb{R}^{d+d_p}, \quad \hat{y} = \mathbf{W}_h \mathbf{z} + \mathbf{b}_h. \quad (14)$$

The TimeSformer and PhysFusion modules are trained jointly, while the cross-domain IBI estimator remains frozen. Gradients flow through both the TimeSformer and PhysFusion modules, while the cross-domain IBI estimator (frozen) produces (μ_t, s_t) during inference. Loss is chosen to match the target (classification) and applied to $\hat{y}$. All projection layers use GELU activations. We provide architectural details in Appendix A.

5 Experimental setup

We evaluate our method on egoPPG-DB [5], which provides eye-gaze videos, contact-based PPG, and electrocardiograms (ECG) for ground-truth beat timing.

5.1 Datasets

HR & HRV. We used egoPPG-DB [5] as a dataset with ground-truth values. egoPPG-DB consists of total 13 hours of recordings from 25 participants performing activities such as kitchen tasks, dancing, and cycling. Each participant wore Aria glasses equipped with inward-facing eye-tracking cameras (30 fps, 320×240 px) and an onboard IMU. Ground-truth signals were collected using a custom contact PPG sensor located at the nasal bridge and validated against an ECG chest strap. The diverse activities cause widely ranging motion intensities and HR values (44–164 bpm). This diversity provides a realistic test bed for evaluating HR and HRV estimation under naturalistic egocentric conditions.

Proficiency estimation. To assess the behavioral relevance of the obtained physiological cues, we evaluate our estimated HRV features on the *proficiency estimation* benchmark from the EgoExo4D [27]. EgoExo4D contains over 5,000 videos from 740 participants performing skilled human activities such as dancing and sports. The task aims to classify proficiency (novice, early expert, intermediate expert, late expert) based on egocentric and exocentric video.

Following the previous protocols [5], we integrate the IBIs estimated by our method into the existing video classification pipeline. Specifically, we obtain IBIs for video clips and concatenate them with the latent features from the TimeSformer [3] before the final classification head.

5.2 Training setup

HR & HRV. We use the official dataset splits introduced for egoPPG [5] and follow a five-fold cross-validation. Each model was trained on four folds and evaluated on the remaining fold. Video frames were preprocessed using spatial normalization and temporal differencing to enhance pulsatile signals. We trained the 3D backbone and low-high decomposition module using mean squared error (MSE) loss between predicted and differentiated reference signals. Optimization employed the Adam [38] (learning rate 9×10^{-4} , batch size 4) for 100 epochs. All experiments were performed on an NVIDIA RTX 4090 GPU.

Proficiency estimation. We implement the TimeSformer with the same configuration as for the EgoExo4D [27] with a clip size of 16 frames and a sampling rate of 16, trained for 15 epochs. We use all videos of the EgoExo4D, for which the proficiency labels are available (using the official benchmark training/validation sets) and which have at least 16 frames at a sampling rate of 16, resulting in 2044 videos. From the official training set, we use 10% as the held-out validation set for testing, following the exact egoPPG implementation, including upsampling the eyetracking video frame rate to 30 fps through linear interpolation between frames [5]. We use our predicted IBIs and their log-variance for the corresponding videos. We first train the HR/HRV estimator on egoPPG-DB and then freeze it to extract IBI features (μ_t) and log-variance (s_t) for EgoExo4D clips. We integrate these features using the *PhysFusion* with the TimeSformer’s backbone before feeding it into the classification head.

5.3 Evaluation Metrics

We report results for both HR and HRV. For HR estimation, we compute the mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and Pearson correlation (r) between predicted and reference HR sequences for each 128-frame window.

For HRV evaluation, we used the IBIs estimated by our model and those extracted from the ground-truth ECG signals using the Pan–Tompkins algorithm [59], and computed standard metrics consistent with HR evaluation. Additionally, we report common time-domain HRV indices [63], such as the standard deviation of IBIs (SDNN) and the root mean square of successive differences (RMSSD). All HRV measures are compared against ECG-derived reference values.

For proficiency estimation, we report top-1 classification accuracy following the official EgoExo4D protocol [27].

Why HRV and not just HR? We report HR and HRV every 128 frames (4.3 s), in addition to the 60 s evaluation used in prior work. Although HRV is conventionally computed from beat-to-beat intervals, window-level evaluation is also used in recent biomedical HRV estimation studies in real-world situations [19].

How much information does a 60 s mean lose? Let H denote 4-second HR in bpm. Across egoPPG-DB [5], the mean HR is $\mathbb{E}[H] = 82.8$ bpm with $\text{Std}(H) = 12.0$ bpm ($\text{Var}(H) = 144$). Within 60 s windows, the average standard deviation is $\bar{\sigma}_{60} = 7.6$ bpm, with $\text{min} = 1.1$ and $\text{max} = 53.2$ bpm. These statistics are computed from the reference ECG recorded alongside the egocentric videos from EgoPPG [5] for accurate ground-truth timing. This indicates that intra-minute fluctuations are highly variable across contexts. Such variability is expected, as the dataset includes dynamic activities such as dancing, biking, and walking, which induce rapid HR changes even within a minute. This is consistent with our prior findings that cardiac dynamics carry nonlinear temporal structure that minute-level aggregates obscure [18].

How much variance is discarded? To quantify the variance lost through 60 s averaging, as done in prior work [5], we compute how much information is discarded using the law of total variance (Equations 15 and 16), highlighting the limitation of such coarse temporal aggregation.

$$\text{Var}(H) = \mathbb{E}[\text{Var}(H \mid \text{minute})] + \text{Var}(\mathbb{E}[H \mid \text{minute}]) \quad (15)$$

Thus, minute-level means retain only $\approx 59.8\%$ of the total variance, discarding about 40% of the information.

$$\frac{\text{Var}(\mathbb{E}[H \mid \text{minute}])}{\text{Var}(H)} = \frac{144 - 57.9}{144} \approx 59.8\% \quad (16)$$

Implications for egocentric systems. Egocentric video captures rapid motions, illumination changes, and viewpoint shifts [27] that can modulate HR in little time (5–15 s). Prior egocentric work averages HR over 60 s windows, which smooths out these fluctuations and discards up to 7–50 bpm of local error. EgoHRV instead evaluates HR and HRV continuously at the timescale of these variations and preserves the rapid dynamics that characterize egocentric settings.

This temporal resolution matters in practice. rPPG methods assume a stationary camera and subject and therefore lose the within-minute dynamics that egocentric applications must capture for users who move freely during daily life.

6 Results

6.1 HR & HRV estimation

Table 1 summarizes the results for HR and HRV estimation. For HR, we follow prior work and report errors in beats per minute. For HRV, we first evaluate the estimated IBIs in milliseconds before reporting the derived time-domain indices. To provide a comprehensive evaluation, we report results for both short-term (~ 4 s) and long-term (60 s) windows.

Table 1: Comparison of heart rate and heart rate variability estimation on egoPPG-DB. Metrics are reported at two different time resolutions for HR and IBI: 4 s and 60 s. SDNN and RMSSD are computed from estimated IBIs over the evaluation windows.

Method	HR								HRV (IBI) [ms]								HRV Time-Domain [ms]			
	MAE		RMSE		MAPE (%)		r		MAE		RMSE		MAPE (%)		r		SDNN		RMSSD	
	4 s	60 s	4 s	60 s	4 s	60 s	4 s	60 s	4 s	60 s	4 s	60 s	4 s	60 s	4 s	60 s	MAE	MAPE	MAE	MAPE
Baseline eyes	27.30	14.60	33.68	18.88	33.71	18.37	0.12	0.20	218.84	237.61	280.12	263.90	44.37	28.37	0.16	0.22	84.92	100.39	84.70	153.42
Baseline skin	26.12	12.40	32.60	15.54	30.16	15.29	0.20	0.50	192.68	223.64	263.23	266.69	42.02	25.98	0.18	0.23	79.71	96.74	81.86	152.35
PhysNet [92]	20.96	11.28	29.88	14.09	26.81	14.89	0.36	0.67	181.97	167.72	235.75	205.88	23.10	21.03	0.38	0.68	28.26	72.71	45.73	131.24
PhysFormer [93]	16.70	13.09	21.45	15.41	19.68	16.60	0.32	0.63	130.44	105.58	178.02	145.48	17.71	15.03	0.34	0.67	29.54	73.14	47.61	129.56
FactorizePhys [34]	14.51	13.27	19.80	16.50	16.85	17.48	0.27	0.69	200.72	210.27	275.51	250.13	27.77	26.27	0.31	0.62	28.65	72.85	45.84	129.35
PulseFormer [5]	22.10	8.10	29.25	11.08	27.96	10.10	0.38	0.83	147.55	123.86	203.00	160.86	19.20	15.67	0.43	0.75	33.71	85.49	53.90	146.45
Ours	12.72	7.13	18.27	10.76	14.47	8.73	0.42	0.84	105.91	87.75	149.78	119.80	13.56	12.81	0.54	0.77	10.89	27.05	19.14	42.55
Gain (%)	12.34	11.98	7.73	2.89	14.12	13.56	10.53	1.20	18.81	16.89	15.86	17.65	23.43	14.77	25.58	2.67	61.46	62.80	58.15	67.10

Table 2: Proficiency estimation results on the EgoExo4D with three random seeds. Accuracy (%) is reported for each scenario and overall for all clips. Following PulseFormer (PF) [5], we reproduce their coarse-grained HR (60s averaged) with secondary statistics such as mean and standard deviation, while our HRV features are computed every 128 frames and fused with TimeSformer representations using our method (+HRV).

Scenario	Majority	Ego			Exo			Ego + Exo		
		Baseline	+HR (PF)	+HRV (ours)	Baseline	+HR (PF)	+HRV (ours)	Baseline	+HR (PF)	+HRV (ours)
Basketball	38.00	45.45±0.62	47.47±0.58	48.48 ±0.55	48.48±0.60	48.98 ±0.57	47.67±0.62	49.49±0.59	50.10±0.55	52.52 ±0.53
Cooking	0.00	20.00±0.70	20.00±0.68	25.00 ±0.65	35.00±0.63	33.75±0.66	36.00 ±0.62	25.00±0.69	35.50 ±0.61	35.00±0.60
Dancing	24.59	43.44±0.64	44.26±0.60	47.54 ±0.58	42.62±0.66	42.62±0.65	47.54 ±0.60	50.82±0.62	55.73±0.59	60.00 ±0.57
Music	57.89	78.94±0.48	77.50±0.50	79.78 ±0.46	57.89±0.55	62.50 ±0.53	52.10±0.60	57.89±0.54	60.13±0.52	60.52 ±0.50
Bouldering	15.29	24.50±0.78	30.37±0.72	39.89 ±0.70	8.61±0.60	14.54±0.58	16.99 ±0.56	14.64±0.74	18.75±0.70	25.49 ±0.66
Soccer ^a	62.50	50.00±8.62	60.00±6.35	67.50 ±7.78	81.25 ±7.11	75.00±8.26	68.75±7.59	75.75 ±7.44	62.50±8.52	70.80±7.67
Overall	27.80	39.69±0.41	44.30±0.38	46.75 ±0.36	34.75±0.43	37.00±0.40	37.77 ±0.39	37.80±0.42	41.18±0.39	43.52 ±0.37

^a Underrepresented scenario (70 out of ~2k samples), contributing to higher variance.

6.2 HRV time-domain metrics

Table 1 additionally lists SDNN and RMSSD, the most widely adopted measures in practical HRV analysis [73] that quantify overall and short-term variability in beat intervals, respectively. SDNN reflects overall variability in IBIs, providing a measure of total autonomic activity, while RMSSD emphasizes short-term fluctuations linked to parasympathetic modulation [50]. We focus on time-domain metrics because frequency-domain HRV measures (e.g., LF/HF power ratios) require much longer continuous recordings (typically several hours up to 24 h [15, 50]), which current egocentric datasets do not capture. We report frequency-domain HRV results for completeness in Appendix E.1 (Table 9), although these metrics are less reliable for short, activity-rich egocentric clips.

Both time and frequency-domain HRV correlate to stress [42], workload [16], and fatigue [21], making them especially relevant for modeling behavioral state in egocentric settings.

Our method improves the estimation of these indices by up to 67% compared to baselines while these improvements increase when we investigate activities as in Appendix E.4. This shows our method’s capability of recovering physiologically meaningful dynamics from short, motion-rich recordings.

6.3 Proficiency estimation on EgoExo4D

Table 2 presents the contribution of our HRV estimation to the proficiency prediction benchmark on EgoExo4D. Integrating the predicted IBIs and their log variances into TimeSformer improves classification accuracy. Our method achieves the highest accuracy in all the activity categories and delivers the best overall performance. When combining egocentric videos with our HRV features, the overall accuracy reaches 46.75%, representing a 17.8% relative improvement over egocentric video alone. These results demonstrate that physiological cues offer complementary behavioral information beyond visual appearance.

Table 3: Ablation study on HR and HRV estimation. Each component is added cumulatively; “−” indicates the exclusion of a design from cross-domain pretraining. The backbone is a 3D CNN [92] with spatial attention [87], applied to eye-tracking video [5].

Configuration	HR (bpm)				HRV (IBI, ms)				HRV Time Indices (ms)					
	MAE	RMSE	MAPE (%)	r	MAE	RMSE	MAPE (%)	r	SDNN			RMSSD		
									MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
Baseline (3D Backbone)	21.76	29.25	27.96	0.38	152.61	209.97	19.55	0.42	31.27	35.73	63.75	48.07	54.20	93.25
+ Low-High decomposition	14.80	20.46	18.10	0.40	117.72	167.99	15.93	0.44	26.96	27.66	35.35	34.72	38.94	55.61
+ Cross-domain pretraining (ours)	12.72	18.27	14.47	0.42	105.91	149.78	13.56	0.54	10.89	16.12	27.05	19.14	26.13	42.55
− Fourier magnitude	14.87	20.51	16.72	0.31	130.30	173.57	19.23	0.33	12.38	18.91	31.28	20.48	27.83	48.37
− NLL objective	15.34	20.09	18.81	0.08	133.47	179.33	18.14	0.10	28.36	33.08	47.59	29.89	36.81	46.15
− U-Net (use 3-layer MLP)	13.40	19.12	15.74	0.08	111.34	160.82	16.62	0.51	11.91	17.29	28.16	20.42	26.93	44.30

Table 4: Ablation on proficiency estimation using EgoExo4D. Top-1 accuracy (%) per activity and overall. Columns use compact labels due to space constraints: Ba=Basketball, Co=Cooking, Da=Dancing, Mu=Music, Bo=Bouldering, So=Soccer.

Configuration	Ba	Co	Da	Mu	Bo	So	Overall
Baseline	45.45	20.00	43.44	78.94	24.50	50.00	39.69
+ HR (60 s mean)	47.47	20.00	44.26	77.50	30.37	60.00	44.30
+ HRV (deterministic, 4 s)	46.50	25.00	46.72	75.31	37.25	31.25	45.31
+ HRV (uncertainty-aware, ours)	48.48	25.00	47.54	79.78	39.89	67.50	46.75

6.4 Ablation experiments

HR & HRV estimation ablation. Table 3 presents the results. The baseline model (3D backbone) predicts temporal differences using standard differencing. Introducing the low-high decomposition substantially reduces HR error. Finally, incorporating cross-domain pretraining yields the largest overall gain, demonstrating the benefit of transferring representations learned from contact BVP-ECG data. The full model achieves the best HR and HRV accuracy across all metrics.

Proficiency estimation ablation. To assess the impact of HRV estimation, we analyze how different HRV representations affect proficiency prediction accuracy on the EgoExo4D benchmark. Table 4 reports results for three variants. The first variant (+ HR) employs 60 s averaged HR values. The second uses HRV features without log variances and the third incorporates probabilistic HRV features trained with the NLL objective to model uncertainty with our method (the uncertainty evaluations with a calibration plot are given in Appendix Section E.3).

Introducing probabilistic modeling improves accuracy. Using HRV features yields the largest performance gain, indicating that inter-beat intervals carry more discriminative information about user proficiency than HR trends. These results demonstrate that the physiological representations learned through cross-domain pretraining enhance downstream reasoning about human skill and behavior.

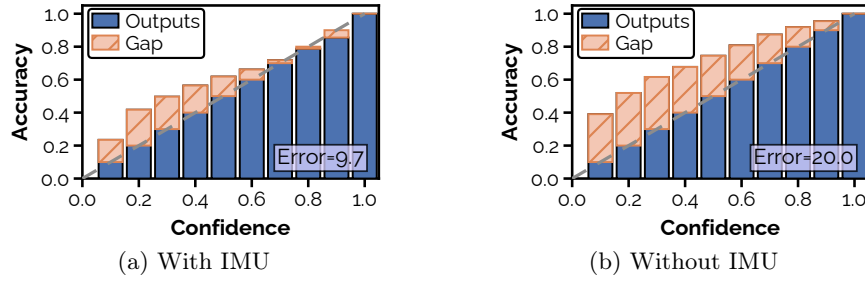


Fig. 3: Reliability diagrams for uncertainty estimation with and without IMU input.

6.5 Uncertainty modeling

We evaluate the contribution of predictive uncertainty and IMU separately. First, removing the uncertainty features from PhysFusion does not change the HRV point estimates, but reduces proficiency accuracy, showing that the predicted uncertainty provides additional information for downstream fusion. Second, we evaluate the role of IMU in uncertainty calibration. Removing the IMU increases the expected calibration error (ECE) from 9.7% to 20.0% (Fig. 3). The IMU provides a motion cue that is independent of visual appearance and skin visibility.

6.6 Efficiency

Our HR/HRV pipeline runs online on a single GPU and stays faster than real time on CPU. The model has 1 M parameters and requires 40.80 GFLOPs per 128-frame (4.3s) window. On an NVIDIA RTX 4090, inference takes 5.75 ms per window with 74 MB peak memory, a real-time factor of $748\times$. A 10-thread CPU still reaches $9.6\times$ real time. Full timings are reported in the appendix (Table 8).

7 Discussion

Performance & implications. Our results show that our method accurately estimates HR and HRV from gaze videos on egocentric vision systems. Our method outperforms prior HR estimation approaches. Our results set the state of the art across both HR and HRV, reducing error in HRV indices, SDNN and RMSSD, by about 65% relative to the best baselines on egoPPG-DB.

Despite this step forward in accuracy, there is room for improvement in HRV indices that depend on the standard deviation of IBIs. Because these measures are sensitive to even a *single* wrongly estimated IBI, local errors can disproportionately affect overall variability. Future work could leverage the predicted log-variance to adaptively downweight uncertain intervals to stabilize HRV metric computation, similar to our proposed *PhysFusion* mechanism.

Relevance of uncertainty-aware HRV for human behavior modeling. Integrating HRV into the EgoExo4D proficiency benchmark resulted in accuracy

gains, with the largest improvements observed in physically or cognitively demanding tasks such as bouldering. Our ablation study shows that our integration of uncertainty estimation accounts for the largest proportional improvement. Further analysis in Appendix E.3 shows that the predicted uncertainty is calibrated, correlates with estimation error, and enables unreliable windows to be down-weighted or excluded. These results suggest that IBIs provide complementary information about effort, attention, and skill assessment.

Towards stress- and arousal-aware egocentric systems. HRV is a well-established noninvasive proxy for autonomic stress responses [9, 37, 80]. The same continuous HRV signal that drives our proficiency gains can thus feed downstream stress- and arousal-aware applications, such as fatigue and workload monitoring in safety-critical activities or affect-sensitive interaction in Mixed Reality.

Limitations. First, the dataset we built on (egoPPG-DB) lacks outdoor recordings, such that our evaluation does not cover weather variability or broader activities. Second, we treat HRV as a stress-related autonomic marker rather than a direct stress label, in line with standard HRV interpretation guidance [39, 72, 79]. We leave a dedicated stress-benchmark evaluation to future work. Finally, it remains to be tested how well our learned HRV representations transfer to additional downstream tasks (e.g., workload estimation).

8 Conclusion

We have presented a method for estimating heart rate and heart rate variability from gaze videos on egocentric vision systems. HRV extends the capacity of egocentric vision models for behavioral understanding much beyond the heart rate tracking in previous work. Using public datasets, our model accurately estimates physiological metrics even under substantial head motion and dynamic illumination changes. We have also demonstrated that our HRV estimates improve downstream user proficiency prediction on the EgoExo4D benchmark by a relative 17.8%, which underscores the behavioral relevance of heart rate variability for modeling skill and better understanding human behavior more generally. Since HRV is an established autonomic proxy for perceived stress, our estimates therefore enables stress- and arousal-aware first-person applications without additional body-worn sensors. Our results position HRV as a core physiological signal for egocentric systems, linking observable behavior with internal physiological state and advancing the foundation for behavioral AI. Our uncertainty-aware method also enhances reliability by quantifying confidence in physiological estimates, which affords egocentric systems more trustworthy behavioral inference in real-world settings. Integrating richer physiological representations into multimodal models in future work could enable deeper behavioral inference to jointly capture intent, attention, and adaptation in real-world settings.

References

1. Magic Leap | Groundbreaking augmented reality solutions,
2. Alqaraawi, A., Alwosheel, A., Alasaad, A.: Heart rate variability estimation in photoplethysmography signals using bayesian learning approach. *Healthcare Technology Letters* **3**(2), 136–142 (2016). ,
3. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: *Proceedings of the International Conference on Machine Learning (ICML)* (July 2021)
4. Bieri, V., Streli, P., Demirel, B.U., Holz, C.: Belieppg: Uncertainty-aware heart rate estimation from ppg signals via belief propagation. In: *Conference on Uncertainty in Artificial Intelligence (UAI)*. PMLR (2023)
5. Braun, B., Armani, R., Meier, M., Moebus, M., Holz, C.: egoPPG: Heart rate estimation from eye-tracking cameras in egocentric systems to benefit downstream vision tasks. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 5579–5590 (2025)
6. Braun, B., McDuff, D., Baltrusaitis, T., Holz, C.: Video-based sympathetic arousal assessment via peripheral blood flow estimation. *Biomedical Optics Express* **14**(12), 6607–6628 (2023)
7. Braun, B., McDuff, D., Baltrusaitis, T., Streli, P., Moebus, M., Holz, C.: Sympcam: Remote optical measurement of sympathetic arousal. In: *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. pp. 1–8. IEEE (2024)
8. Cacioppo, J.T., Tassinary, L.G.: Inferring psychological significance from physiological signals. *American Psychologist* **45**(1), 16–28 (1990).
9. Castaldo, R., Melillo, P., Bracale, U., Caserta, M., Triassi, M., Pecchia, L.: Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis. *Biomedical Signal Processing and Control* **18**, 370–377 (2015).
10. Charlton, P.H., Marozas, V., Mejía-Mejía, E., Kyriacou, P.A., Mant, J.: Determinants of photoplethysmography signal quality at the wrist. *PLOS Digital Health* **4**, 1–24 (06 2025). ,
11. Choi, A., et al: Photoplethysmography sampling frequency: pilot assessment of how low can we go to analyze pulse rate variability with reliability? *Physiological Measurement*
12. Chrousos, G.P.: Stress and disorders of the stress system. *Nature Reviews Endocrinology* **5**(7), 374–381 (2009).
13. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: *European Conference on Computer Vision (ECCV)* (2018)
14. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **43**(11), 4125–4141 (2021).
15. De Maria, B., Parati, M., Dalla Vecchia, L.A., La Rovere, M.T.: Day and night heart rate variability using 24-h ECG recordings: a systematic review with meta-analysis using a gender lens. *Clinical Autonomic Research* **33**(6) (Dec 2023). ,
16. Delliaux, S., Delaforge, A., Deharo, J.C., Chaumet, G.: Mental workload alters heart rate variability, lowering non-linear dynamics. *Frontiers in Physiology* **Volume 10 - 2019** (2019). ,

17. Demirel, B.U., Holz, C.: An unsupervised approach for periodic source detection in time series. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)
18. Demirel, B.U., Holz, C.: Temporal cardiovascular dynamics for improved PPG-based heart rate estimation. *IEEE Journal of Biomedical and Health Informatics* (2025)
19. Demirel, B.U., Holz, C.: Continuous heart rate variability estimation from ppg via state-space modeling. *IEEE Transactions on Biomedical Engineering* pp. 1–8 (2026).
20. Engel, J., et al: Project aria: A new tool for egocentric multi-modal ai research (2023),
21. Escorihuela, R.M., Capdevila, L., Castro, J.R., Zaragozà, M.C., Maurel, S., Alegre, J., Castro-Marrero, J.: Reduced heart rate variability predicts fatigue severity in individuals with chronic fatigue syndrome/myalgic encephalomyelitis. *Journal of Translational Medicine* **18**(1) (Jan 2020). ,
22. Fernandes, G., Wei, B., Romano, C., Ulusel, D., Dambanemuya, H.K., Gao, Y., Ghaffari, R., Rogers, J., Alshurafa, N.: Healthsense: Unobtrusive continuous stress monitoring using a novel dual ecg-ppg patch. In: 2024 IEEE 20th International Conference on Body Sensor Networks (BSN). pp. 1–4 (2024).
23. Forte, G., Morelli, M., Grässler, B., Casagrande, M.: Decision making and heart rate variability: A systematic review. *Applied Cognitive Psychology* **36**(1), 100–110 (2022). ,
24. Furnari, A., Battiato, S., Farinella, G.M.: Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In: International Workshop on Egocentric Perception, Interaction and Computing (EPIC) in conjunction with ECCV (2018)
25. Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., Tsiknakis, M.: Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing* **13**(1), 440–460 (2022).
26. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., González, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolář, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Ruiz, P., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbeláez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18995–19012 (June 2022)
27. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khiredkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu,

- X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanov, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Damen, D., Engel, J., Farinella, G.M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 19383–19400 (June 2024)
28. Gu, Q., Lv, Z., Frost, D., Green, S., Straub, J., Sweeney, C.: EgoLifter: Open-world 3d segmentation for egocentric perception. *arXiv preprint arXiv:2403.18118* (2024)
29. de Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering* **60**(10), 2878–2886 (2013).
30. Hatano, M., Hachiuma, R., Fujii, R., Saito, H.: Multimodal cross-domain few-shot learning for egocentric action recognition. In: *European Conference on Computer Vision (ECCV)* (2024)
31. Immanuel, S., Teferra, M.N., Baumert, M., Bidargaddi, N.: Heart rate variability for evaluating psychological stress changes in healthy adults: A scoping review. *Neuropsychobiology* **82**(4), 187–202 (2023). ,
32. Jamnot, M., Braun, B., Strel, P., Wampfler, R., Holz, C.: egoEMOTION: Egocentric vision and physiological signals for emotion and personality recognition in real-world tasks. *Advances in Neural Information Processing Systems* **38** (2026)
33. Jerri, S., Murugappan, M., Nagarajan, R., Wan, K.: Physiological signals based human emotion recognition: a review. In: *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*. pp. 410–415 (2011).
34. Joshi, J., Agaian, S.S., Cho, Y.: Factorizephys: Matrix factorization for multidimensional attention in remote physiological sensing. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) *Advances in Neural Information Processing Systems*. vol. 37, pp. 96607–96639. Curran Associates, Inc. (2024),
35. Julien, C., Roman, G.C.: Egocentric context-aware programming in ad hoc mobile environments. In: *Proceedings of the 10th ACM SIGSOFT Symposium on Foundations of Software Engineering*. p. 21–30. SIGSOFT '02/FSE-10, Association for Computing Machinery, New York, NY, USA (2002). ,
36. Katahira, K., Fujimura, T., Matsuda, Y.T., Okanoya, K., Okada, M.: Individual differences in heart rate variability are associated with the avoidance of negative emotional events. *Biological Psychology* **103**, 322–331 (2014). ,
37. Kim, H.G., Cheon, E.J., Bai, D.S., Lee, Y.H., Koo, B.H.: Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation* **15**(3), 235–245 (2018).
38. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
39. Laborde, S., Mosley, E., Thayer, J.F.: Heart rate variability and cardiac vagal tone in psychophysiological research—recommendations for experiment planning, data analysis, and data reporting. *Frontiers in Psychology* **8**, 213 (2017).
40. Lee, C., Shin, M., Eniyandunmo, D., Anwar, A., Kim, E., Kim, K., Yoo, J.K., Lee, C.: Predicting driver’s mental workload using physiological signals: A functional data analysis approach. *Applied Ergonomics* **118**, 104274 (2024). ,

41. Li, X., Alikhani, I., Shi, J., Seppanen, T., Junttila, J., Majamaa-Voltti, K., Tulppo, M., Zhao, G.: The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 242–249 (2018).
42. Lischke, A., Jacksteit, R., Mau-Moeller, A., Pahnke, R., Hamm, A.O., Weippert, M.: Heart rate variability is associated with psychosocial stress in distinct social domains. *Journal of Psychosomatic Research* **106**, 56–61 (2018). ,
43. Liu, X., Narayanswamy, G., Paruchuri, A., Zhang, X., Tang, J., Zhang, Y., Wang, Y., Sengupta, S., Patel, S., McDuff, D.: rppg-toolbox: Deep remote ppg toolbox. arXiv preprint arXiv:2210.00716 (2022)
44. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21013–21022 (June 2022)
45. Luft, C.D.B., Takase, E., Darby, D.: Heart rate variability and cognitive function: Effects of physical effort. *Biological Psychology* **82**(2), 186–191 (2009). ,
46. Luo, C., Xie, Y., Yu, Z.: Phymamba: Efficient remote physiological measurement with slowfast temporal difference mamba. In: Chinese Conference on Biometric Recognition (CCBR) (2024)
47. Luong, T., Holz, C.: Characterizing physiological responses to fear, frustration, and insight in virtual reality. *IEEE Transactions on Visualization and Computer Graphics* **28**(11), 3917–3927 (2022).
48. Luque-Casado, A., Perales, J.C., Cárdenas, D., Sanabria, D.: Heart rate variability and cognitive processing: The autonomic response to task demands. *Biological Psychology* **113**, 83–90 (2016). ,
49. Maaoui, C., Pruski, A.: Emotion recognition through physiological signals for human-machine communication. In: Kordic, V. (ed.) *Cutting Edge Robotics 2010*, chap. 20. IntechOpen, London (2010). ,
50. Malik, M., Hnatkova, K., Huikuri, H.V., Lombardi, F., Schmidt, G., Zabel, M.: Crosstalk proposal: Heart rate variability is a valid measure of cardiac autonomic responsiveness. *The Journal of Physiology* **597**(10), 2595–2598 (2019). ,
51. Meier, M., Demirel, B.U., Holz, C.: WildPPG: A real-world PPG dataset of long continuous recordings. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024)
52. Min, K., Corso, J.J.: Integrating human gaze into attention for egocentric activity recognition. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1068–1077 (2021).
53. Miyatsu, T., Smith, B.M., Koutnik, A.P., Pirolli, P., Broderick, T.J.: Resting-state heart rate variability after stressful events as a measure of stress tolerance among elite performers. *Frontiers in Physiology* **Volume 13 - 2022** (2023). ,
54. Moebus, M., Hauptmann, L., Kopp, N., Demirel, B., Braun, B., Holz, C.: Nightbeat: Heart rate estimation from a wrist-worn accelerometer during sleep. *IEEE Journal of Biomedical and Health Informatics* (2024)
55. Nagarajan, T., Ramakrishnan, S.K., Desai, R., Hillis, J., Grauman, K.: Egoenv: Human-centric environment representations from egocentric video. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*. vol. 36, pp. 60130–60143. Curran Associates, Inc. (2023),

56. Niu, X., Han, H., Shan, S., Chen, X.: Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In: Asian Conference on Computer Vision (2018),
57. Paalasmaa, J., Toivonen, H., Partinen, M.: Adaptive heartbeat modeling for beat-to-beat heart rate measurement in ballistocardiograms. *IEEE Journal of Biomedical and Health Informatics* (2015)
58. Palmer, C.J., Caruana, N., Clifford, C.W.G., Seymour, K.J.: Perceptual integration of head and eye cues to gaze direction in schizophrenia. *Royal Society Open Science* **5**(12), 180885 (2018). ,
59. Pan, J., Tompkins, W.J.: A real-time qrs detection algorithm. *IEEE Transactions on Biomedical Engineering* **BME-32**(3), 230–236 (1985).
60. Park, J., Seok, H.S., Kim, S.S., Shin, H.: Photoplethysmogram analysis and applications: An integrative review. *Frontiers in Physiology* **Volume 12 - 2021** (2022).
61. Peláez-Coca, M.D., Hernando, A., Lázaro, J., Gil, E.: Impact of the ppg sampling rate in the pulse rate variability indices evaluating several fiducial points in different pulse waveforms. *IEEE Journal of Biomedical and Health Informatics* **26**(2), 539–549 (2022).
62. Radomski, A., Arai, M., Oshima, K., Oneda, N., Ueno, A., Teichmann, D.: Reflective ppg sensor measures heart rate through clothing during driving. *IEEE Sensors Journal* **25**(3), 4157–4166 (2025).
63. Rajendra Acharya, U., Paul Joseph, K., Kannathal, N., Lim, C.M., Suri, J.S.: Heart rate variability: a review. *Medical and Biological Engineering and Computing* **44**(12), 1031–1051 (Dec 2006). ,
64. Reiss, A., Indlekofer, I., Schmidt, P., Van Laerhoven, K.: Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors* (2019). ,
65. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing (2015)
66. Sahroni, A., Hassya, I.A., Rifaldi, R., Jannah, N.U., Irawan, A.F., Rahayu, A.W.: Hrv assessment using finger-tip photoplethysmography (pulserate) as compared to ecg on healthy subjects during different postures and fixed breathing pattern. *Procedia Computer Science* (2019)
67. Saul, J.P., Valenza, G.: Heart rate variability and the dawn of complex physiological signal analysis: methodological and clinical perspectives. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **379**(2212), 20200255 (2021). ,
68. Scalia, G., Grambow, C.A., Pernici, B., Li, Y.P., Green, W.H.: Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of Chemical Information and Modeling* **60**(6) (Jun 2020)
69. Scardulla, F., Cosoli, G., Spinsante, S., Poli, A., Iadarola, G., Pernice, R., Busacca, A., Pasta, S., Scalise, L., D’Acquisto, L.: Photoplethysmographic sensors, potential and limitations: Is it time for regulation? a comprehensive review. *Measurement* **218**, 113150 (2023). ,
70. Schroeder, E.B., Liao, D., Chambless, L.E., Prineas, R.J., Evans, G.W., Heiss, G.: Hypertension, blood pressure, and heart rate variability. *Hypertension* **42**(6), 1106–1111 (2003).
71. Scott, E.E., LoTemplio, S.B., McDonnell, A.S., McNay, G.D., Greenberg, K., McKinney, T., Uchino, B.N., Strayer, D.L.: The autonomic nervous system in its natural environment: Immersion in nature is associated with changes in heart rate and heart rate variability. *Psychophysiology* **58**(4), e13698 (2021). ,

72. Shaffer, F., Ginsberg, J.P.: An overview of heart rate variability metrics and norms. *Frontiers in Public Health* **5**, 258 (2017).
73. Shaffer, F., Ginsberg, J.P.: An overview of heart rate variability metrics and norms. *Frontiers in Public Health* **Volume 5 - 2017** (2017). ,
74. Sharma, N., Gedeon, T.: Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer Methods and Programs in Biomedicine* **108**(3), 1287–1301 (2012).
75. Sheridan, D.C., Dehart, R., Lin, A., Sabbaj, M., Baker, S.D.: Heart rate variability analysis: How much artifact can we remove? *Psychiatry Investigation* (2020)
76. Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., Yang, X.: A review of emotion recognition using physiological signals. *Sensors* **18**(7) (2018). ,
77. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Actor and observer: Joint modeling of first and third-person videos. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018)
78. Sudhakaran, S., Escalera, S., Lanz, O.: LSTA: Long Short-Term Attention for Egocentric Action Recognition. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019)
79. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology: Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation* **93**(5), 1043–1065 (1996).
80. Thayer, J.F., Åhs, F., Fredrikson, M., Sollers, J.J., Wager, T.D.: A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews* **36**(2), 747–756 (2012).
81. Thayer, J.F., Lane, R.D.: A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders* **61**(3), 201–216 (2000).
82. Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J.F., Sebe, N.: Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2396–2404 (2016).
83. Veltman, J., Gaillard, A.: Indices of mental workload in a complex task environment. *Neuropsychobiology* (02 2008). ,
84. Verkruysse, W., Svaasand, L.O., Nelson, J.S.: Remote plethysmographic imaging using ambient light. *Opt. Express* **16**(26), 21434–21445 (Dec 2008). ,
85. Wang, X., Zhu, L., Wang, H., Yang, Y.: Interactive prototype learning for egocentric action recognition. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 8148–8157 (2021).
86. Wang, Z., Zou, Y., Liu, J., Peng, W., Li, M., Zou, Z.: Heart rate variability in mental disorders: an umbrella review of meta-analyses. *Translational Psychiatry* **15**(1), 1–11 (Mar 2025). ,
87. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (September 2018)
88. Wu, Y., Zhu, L., Wang, X., Yang, Y., Wu, F.: Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing* **30**, 1143–1152 (2021).
89. Xu, X., Xue, F., Zhao, S., Pan, Y., Scherer, S., Huang, X.: Mac-ego3d: Multi-agent gaussian consensus for real-time collaborative ego-motion and photorealistic 3d reconstruction. *arXiv preprint arXiv:2412.09723* (2024)

90. Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., Ouyang, B., Lin, Z., Cominelli, M., Cai, Z., Li, B., Zhang, Y., Zhang, P., Hong, F., Widmer, J., Gringoli, F., Yang, L., Liu, Z.: Egolife: Towards egocentric life assistant. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28885–28900 (June 2025)
91. Yang, Y., Zhai, W., Wang, C., Yu, C., Cao, Y., Zha, Z.J.: Egochoir: Capturing 3d human-object interaction regions from egocentric views. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) *Advances in Neural Information Processing Systems*. vol. 37, pp. 54529–54557. Curran Associates, Inc. (2024),
92. Yu, Z., Li, X., Zhao, G.: Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In: *Proc. BMVC* (2019)
93. Yu, Z., Shen, Y., Shi, J., Zhao, H., Torr, P., Zhao, G.: Physformer: Facial video-based physiological measurement with temporal difference transformer. In: *CVPR* (2022)
94. Zou, B., Guo, Z., Hu, X., Ma, H.: Rhythmmamba: Fast remote physiological measurement with arbitrary length videos. *arXiv preprint arXiv:2404.06483* (2024)

A Architectures

In this section, we give details about the architectures that are used for our method. Since our 3D backbone is adapted from previous works, we have reported the low-high decomposition, cross-domain pretraining models and PhysFusion.

A.1 Low-high decomposition

We integrate a lightweight low-high frequency decomposition block into the 3D backbone to suppress broadband motion noise while preserving pulsatile dynamics. This module operates directly on the temporal feature sequence $\mathbf{y} \in \mathbb{R}^{B \times 1 \times T}$ and introduces fewer than 0.1% additional parameters to the overall model.

Table 5: Low-high decomposition block details.

Component	Specification
Input	$\mathbf{y} \in \mathbb{R}^{B \times 1 \times T}$ (from 3D backbone).
Low-pass	Box filter (AvgPool/Conv) with $k=11$, stride 1, pad 5: $\mathbf{y}_{lo} = \mathbf{y} * h$, $h = \frac{1}{11} \mathbf{1}_{11}$.
High-pass	$\mathbf{y}_{hi} = \mathbf{y} - \mathbf{y}_{lo}$.
Head _{lo}	Conv1D $1 \rightarrow 16$ ($k=9, p=4$) $\rightarrow$ GELU $\rightarrow$ Conv1D $16 \rightarrow 1$ ($k=1$). Output $\tilde{\mathbf{y}}_{lo} \in \mathbb{R}^{B \times 1 \times T}$.
Head _{hi}	Conv1D $1 \rightarrow 16$ ($k=5, p=2$) $\rightarrow$ GELU $\rightarrow$ Conv1D $16 \rightarrow 1$ ($k=1$). Output $\tilde{\mathbf{y}}_{hi} \in \mathbb{R}^{B \times 1 \times T}$.
Fusion gain	Learnable scalar g (initialized to 0.3).
Residual fusion	$\hat{\mathbf{y}} = \mathbf{y} + \tanh(g) [(\tilde{\mathbf{y}}_{lo} + \tilde{\mathbf{y}}_{hi}) - \mathbf{y}]$ (identity as $g \rightarrow 0$).
Output	$\hat{\mathbf{y}} \in \mathbb{R}^{B \times 1 \times T}$.
Params	~ 300 total (two heads + gate).

Notes. k : kernel size, p : padding. All convolutions use stride 1 with "same" padding. The low-pass kernel ($k=11$) spans ≈ 0.37 s at 30 Hz.

A.2 Cross-domain models

To enhance the performance of egocentric IR cameras in extracting physiological indicators, we leverage large-scale datasets from contact-based sensors that record blood volume pulse (BVP) and inertial data. The architecture used for this cross-domain pretraining is detailed in Table 6.

The model first processes the Fourier magnitude of the integrated BVP signals using a U-Net encoder-decoder to extract temporal features. In parallel, the Fourier magnitude of the inertial signal magnitude is processed by a lightweight convolutional encoder to capture motion dynamics. The BVP encoder predicts

the mean IBI, while the fused BVP and IMU representations are used to predict the log-variance.

The overall network contains approximately 230k parameters in the multi-modal configuration. During proficiency prediction, the cross-domain model is kept frozen and used only as a feature extractor.

Table 6: Architectural details of the cross-domain models used for estimating IBIs. All inputs are Fourier magnitude features. k : kernel size, p : padding.

Component	Specification
BVP encoder-decoder (f_{BVP})	
Input	$\mathbf{z} \in \mathbb{R}^{B \times 1 \times 128}$ (FFT magnitude of BVP segment).
Downsampling	Four sequential blocks: (1) Conv1D $1 \rightarrow N$ ($k=3, p=1$) $\rightarrow$ BN $\rightarrow$ ReLU; (2) Conv1D $N \rightarrow 2N$ ($k=3, p=1$) $\rightarrow$ BN $\rightarrow$ ReLU; (3-4) additional down paths with concatenated pooled inputs (kernel 3, stride 2).
Upsampling	Three symmetric decoder stages with linear interpolation (scale factor 2) followed by concatenation and Conv1D-BN-ReLU blocks ($k=3, p=1$).
Output (mean head)	Conv1D $N \rightarrow 1$ ($k=3, p=1$) $\rightarrow$ Linear ($128 \rightarrow 1$) $\Rightarrow \mu$.
Global pooling	AdaptiveAvgPool1D(1) $\rightarrow$ Linear($N \rightarrow 128$) to form BVP embedding.
IMU encoder (f_{IMU})	
Input	$\mathbf{q} \in \mathbb{R}^{B \times 1 \times 128}$ (FFT magnitude of IMU magnitude signal).
Conv blocks	Conv1D $1 \rightarrow 64$ ($k=9, p=4$) $\rightarrow$ GELU $\rightarrow$ Conv1D $64 \rightarrow 64$ ($k=9, p=4$) $\rightarrow$ GELU $\rightarrow$ AdaptiveAvgPool1D(1).
Projection	Linear($64 \rightarrow 64$) $\Rightarrow$ IMU feature.
Fusion and probabilistic output	
Fusion	Concatenate BVP (128-D) and IMU (64-D) embeddings $\rightarrow$ 192-D vector.
Log-variance head (IMU)	MLP: Linear($192 \rightarrow 64$) $\rightarrow$ GELU $\rightarrow$ Linear($64 \rightarrow 1$) $\Rightarrow s$.
Output	Predicted IBI mean μ and log-variance s ; trained with NLL loss.
Params	~ 227 k (with IMU fusion), ~ 210 k (PPG-only).

Notes. $N=32$ channels in the base layer. All convolutions use stride 1 and “same” padding. FFT input length $K=128$ corresponds to 4.3 s temporal windows at 30 Hz.

A.3 PhysFusion

We fuse video features with per-timestep physiological cues. We mainly use the IBI estimations while normalizing them into per minute values and model

Table 7: The proposed PhysFusion block details (video CLS fused with physiological stream).

Component	Specification
Inputs	Per-timestep HR h_t and log-variance $\log \sigma_t^2$; tensors $h, \log \sigma^2 \in \mathbb{R}^{B \times T}$. Stack to $[B, T, 2]$.
Projection	Linear ($2 \rightarrow 128$) applied at each time step $\Rightarrow [B, T, 128]$ ($d_{\text{phys}} = 128$).
Temporal encoder	TransformerEncoder, 2 layers; each layer: $d_{\text{model}}=128$, nhead = 4, dim_feedforward = 256, dropout = 0.1, batch_first =True.
Attention pooling	Learned scores via Linear ($128 \rightarrow 1$); masked softmax over T ; weighted sum $\Rightarrow$ physiology vector $v_{\text{phys}} \in \mathbb{R}^{B \times 128}$.
Fusion	Concatenate ViT CLS ($\in \mathbb{R}^{B \times \text{embed_dim}}$) with $v_{\text{phys}} \Rightarrow [B, \text{embed_dim}+128]$.
Head	Linear ($\text{embed_dim}+128 \rightarrow \text{num_classes}$) for prediction.

Notes. The cross-domain encoder is kept frozen during the training of PhysFusion, which learns end-to-end with the classifier.

log-variance ($\log \sigma_t^2$). At each time step we stack $[h_t, \log \sigma_t^2]$, project with a linear layer to a 128-D embedding, and encode the sequence with a 2-layer Transformer. We apply a learned attention pooling over time to get a single 128-D physiology vector. This vector is concatenated with the ViT CLS token and passed to a linear head for prediction. The details of the model are given in Table 7.

B Training details

B.1 Low-high decomposition

The low-high decomposition module is trained jointly with the 3D backbone following the same training setup as in prior work [5]. Training uses the Adam optimizer with a learning rate of 9×10^{-4} , batch size of 4, and runs for 100 epochs. The best model is chosen with the lowest MAE.

B.2 Cross-domain pretraining

We pretrain the cross-domain encoder on large-scale contact-sensor datasets using the Fourier magnitude inputs. Each training window spans $T=128$ samples, with supervision provided by ECG-derived IBIs extracted using the Pan-Tompkins algorithm.

The Dalia dataset is preprocessed with a second-order Butterworth band-pass filter (0.5–8 Hz) applied to the PPG signal. The IMU magnitude is computed and both signals are resampled to match the temporal specifications of the Aria in egoPPG dataset [5] and the 3D backbone ($T=128$ samples per window). After preprocessing, the magnitude spectra from the discrete Fourier transform are used as input to the BVP and IMU encoders.

The training of the cross-domain models is performed using the Adam optimizer [38] with a learning rate of 5×10^{-4} , weight decay of 1×10^{-7} , batch size of 128. We train for 100 epochs. We also use a learning scheduler adjusts the learning rate based on validation MAE (factor 0.5, patience 15, minimum learning rate 1×10^{-6}), and early stopping is applied using the same criterion. The model checkpoint with the lowest validation MAE is selected and reloaded for usage in egocentric setups without requiring further finetuning. After pretraining, the cross-domain encoder is frozen and used as a fixed feature extractor for proficiency prediction.

C Efficiency report

We evaluate the computational cost of the HR/HRV pipeline on a 128-frame (4.3 s) input window. The model has approximately 1 M parameters and requires 20.40 B multiply-accumulate operations (40.80 GFLOPs) per window.

Table 8 reports inference latency on GPU and CPU. On an NVIDIA RTX 4090, the model processes one window in 5.75 ms, or roughly 174 windows per second. This is about 748 times faster than the 4.3 s acquisition duration of the input window, and requires only 74 MB of peak GPU memory. On CPU, inference takes 447.7 ms with 10 threads, or 2.23 windows per second and a real-time factor of approximately $9.6\times$.

These results show that our method supports online inference on a GPU and stays faster than real time on a multi-threaded CPU. The low parameter count and memory footprint also make the model suitable for continuous egocentric sensing pipelines.

Table 8: Inference efficiency for a 128-frame (4.3 s) input window. The real-time factor is the input duration divided by inference latency.

Device	Time (ms)	Windows/s	Real-time factor	Peak mem. (MB)
RTX 4090	5.75	173.9	$747.8\times$	74
CPU (10 threads)	447.7	2.23	$9.6\times$	–

D Reproducibility

For HR estimates, we followed the official implementation and evaluation settings provided in the *egoPPG* repository (<https://github.com/eth-siplab/egoPPG>). We note that our reproduced PulseFormer results differ slightly from the values reported in the original *egoPPG* paper. The original paper reports results from a single training run, whereas the repository additionally reports the mean and standard deviation across three random seeds and notes variability across seeds and execution environments. Our reproduced values are close to the range reported by this multi-seed evaluation.

For proficiency estimation, we used the official TimeSformer implementation (<https://github.com/facebookresearch/TimeSformer>) with its default hyperparameter configuration. All additional modifications related to HRV integration and fusion are described in the main paper and supplementary Sections 4–4.2.

E Additional results

E.1 Frequency domain based HRV indices

In this section, we have provided additional results regarding frequency-domain based HRV indices derived from estimated IBIs. Table 9 presents the results for LF, HF and LF/HF ratio while comparing our method with previous techniques. As can be seen from the table, our method improves the performance up to 31.8% compared to previous works.

Table 9: Frequency-domain HRV indices computed from estimated IBIs. LF (0.04–0.15 Hz) and HF (0.15–0.40 Hz) are reported in ms^2 ; LF/HF is unitless. These metrics are included for completeness, though short egocentric clips make them less reliable.

Method	LF (ms^2)		HF (ms^2)		LF/HF Ratio	
	MAE ↓	r ↑	MAE ↓	r ↑	MAE ↓	r ↑
PhysNet [92]	310.8	0.20	278.2	0.20	0.42	0.18
PhysFormer [93]	285.5	0.22	252.4	0.22	0.38	0.20
FactorizePhys [34]	298.1	0.21	269.9	0.21	0.40	0.19
PulseFormer [5]	271.6	0.25	246.1	0.25	0.35	0.22
Ours	209.7	0.30	197.4	0.31	0.29	0.29
Gain (%)	22.8	20.0	19.8	24.0	17.1	31.8

Beyond these improvements, frequency-domain HRV metrics still rely on relatively long and stable windows, which limits their robustness in short egocentric clips with rapid motion. While our method already raises the accuracy and correlation across LF, HF, and LF/HF, there is room for progress. Future work can model temporal structure with sequence-level representations or architectures that capture transient patterns in IBIs. Such approaches may recover the fast dynamics, giving more reliable estimates and better downstream performance in egocentric settings.

60 s HRV Since we reported the performance of 4 s HRV in Table 1, we also performed additional experiments to show the advantage of higher resolution HRV in the proficiency estimation downstream task. Table 10 presents the results.

Table 10: Proficiency estimation accuracy (%) on EgoExo4D (ego-view only). We also reported the coarse-grained HRV statistics averaged over 60 s, whereas our main method uses HRV features computed every 128 frames.

Scenario	Majority Baseline		+HRV	
			(60 s, ours)	(4 s, ours)
Basketball	38.00	45.45	40.00	48.48
Cooking	0.00	20.00	20.00	25.00
Dancing	24.59	43.44	40.59	47.54
Music	57.89	78.94	77.50	79.78
Bouldering	15.29	24.50	20.30	39.89
Soccer	62.50	50.00	60.00	67.50
Overall	27.80	39.69	43.50	46.75

The 60 s variant captures slow trends but loses short-term changes tied to motion and skill execution. As in Table 10, the higher-resolution representation gives better accuracy. This supports our main claim that preserving HRV dynamics is important for proficiency estimation, where local performance cues often appear within a few seconds rather than across a full minute.

The results align with the variance analysis in the main script. Coarse aggregation removes a notable part of the signal, while our method keeps within-minute variability that correlates with behavior. As shown in Table 10, the our method improves the overall accuracy by a clear margin, reinforcing the need for higher temporal resolution in egocentric settings.

E.2 Visual examples

To further illustrate our approach, we visualize the frequency characteristics of the reconstructed BVP signals. Figure 4 shows the magnitude spectra obtained using the Fourier transform for the outputs of different models alongside the ground truth BVP from the contact nose PPG sensor. Our method exhibits the closest spectral distribution to the reference, demonstrating improved recovery of physiologically relevant frequency components.

Spectral distribution analysis To quantitatively assess how closely each reconstructed BVP signal matches the spectral profile of the reference contact PPG,

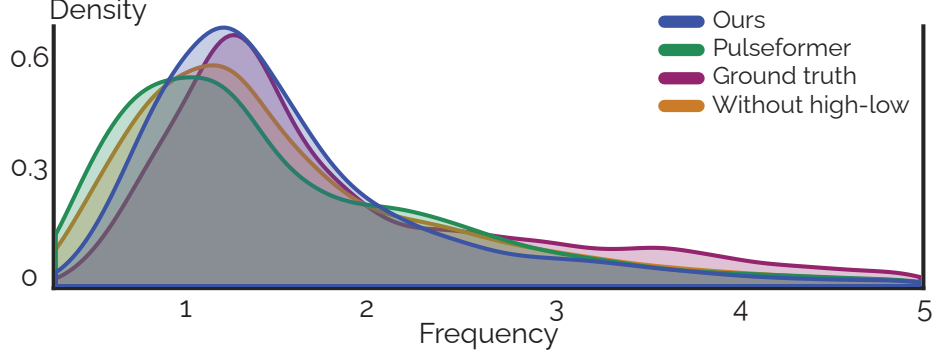


Fig. 4: Spectral similarity between model outputs and ground-truth is evaluated using kernel density estimates. Our method shows the highest overlap with the true contact PPG frequency distribution, indicating better recovery of relevant frequency components.

we evaluate the divergence between their normalized power spectral densities (PSDs). Let $P_i(f)$ denote the average PSD distribution of model i over frequency $f \in \mathcal{F}$, normalized to form a probability density.

$$p_i(f) = \frac{P_i(f)}{\sum_{f \in \mathcal{F}} P_i(f)}, \quad \sum_{f \in \mathcal{F}} p_i(f) = 1. \quad (17)$$

For two distributions $p_i(f)$ and $p_j(f)$, the *Kullback–Leibler divergence* (KL) is defined as in Equation 18.

$$D_{\text{KL}}(p_i \| p_j) = \sum_{f \in \mathcal{F}} p_i(f) \log \frac{p_i(f)}{p_j(f)}. \quad (18)$$

KL divergence is asymmetric, i.e. $D_{\text{KL}}(p_i \| p_j) \neq D_{\text{KL}}(p_j \| p_i)$, where larger $D_{\text{KL}}(p_i \| p_j)$ indicates that p_i allocates energy in frequency regions absent in p_j . To obtain a symmetric and bounded measure, we also report the *Jensen–Shannon divergence* (JS). We compute divergences between the predicted BVP distributions and the reference contact PPG (p_{ref}). Table 11 summarizes the results.

Table 11: Spectral divergence between normalized PSDs of model outputs and the reference contact PPG. Lower is better.

Comparison	$D_{\text{KL}}(p_i \ p_{\text{ref}})$	$D_{\text{KL}}(p_{\text{ref}} \ p_i)$	$D_{\text{JS}}(p_i, p_{\text{ref}})$
Ours vs. contact PPG	0.0492	0.0577	0.0131
Without High–Low vs. contact PPG	0.0919	0.0755	0.0198
PulseFormer vs. contact PPG	0.1659	0.1256	0.0337

Our model achieves the lowest divergence from the ground-truth spectrum ($D_{JS}=0.0131$), indicating the highest fidelity in recovering frequency content. Removing the low-high decomposition module increases D_{JS} by 51%, while the baseline PulseFormer model diverges by 157%. The asymmetry in D_{KL} suggests that competing models tend to overemphasize frequencies absent in the true PPG spectrum, whereas our approach yields a closer match across both low and high-frequency bands relevant to the pulse.

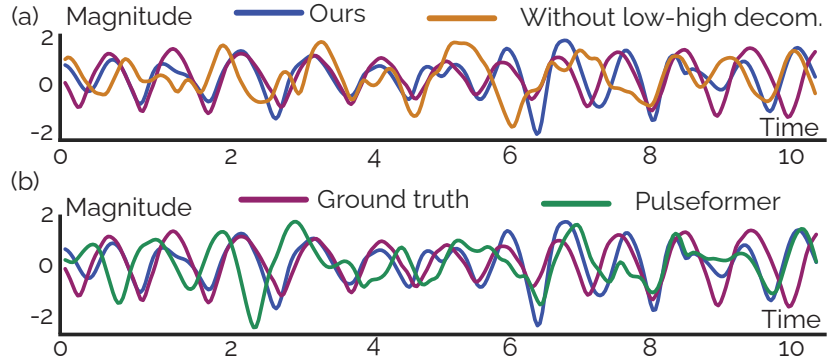


Fig. 5: Qualitative comparison of reconstructed waveforms. (a) compares our method (blue) against the ground-truth contact BVP (purple) and the ablated variant without low-high decomposition (yellow). (b) compares our method (blue) with PulseFormer [5] (green) for the same segment. Our approach aligns closely with the ground-truth waveform, whereas PulseFormer and the ablated model exhibit phase misalignment.

Time-domain analysis To further assess temporal fidelity, we visualize reconstructed pulse waveforms for a representative 10-second sequence in Figure 5. Our method maintains both phase and amplitude consistency with the ground-truth contact BVP signal, accurately capturing beat-to-beat variations. In contrast, PulseFormer [5] and the model without the low-high decomposition exhibit temporal lag and smoothed pulse contours, confirming that the proposed decomposition enhances precise recovery of cardiac cycles from egocentric video.

E.3 Evaluation of uncertainty calibration

We propagate the predictive variance using the delta method and then apply a single affine variance calibration. The reliability diagram (Fig. 6) shows empirical coverage versus nominal confidence. Uncertainty estimates are informative (Spearman’s $\rho=0.714$) and calibrated intervals are slightly conservative at low confidence (10% $\rightarrow$ 30.9%, 20% $\rightarrow$ 46.7%) but align well at high confidence (90% $\rightarrow$ 91.6%). We compute the normalized Area Under the Calibration Error Curve (AUCE=0.217) for the total deviation from perfect calibration [68]. Results show low miscalibration, and strong agreement at high confidence. We also

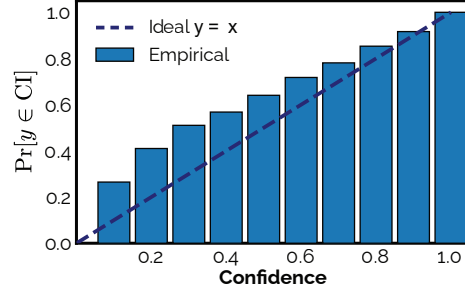


Fig. 6: Reliability diagram of empirical versus nominal coverage. The model exhibits slightly conservative uncertainty at low confidence and near-perfect calibration at high confidence.

compute the relationship between predicted uncertainty and error where higher uncertainty corresponds to higher error (Fig 7a, coverage vs IBI error). Using uncertainty for gating segments improves overall estimation reliability. In other words, keeping lower-uncertainty windows reduces errors.

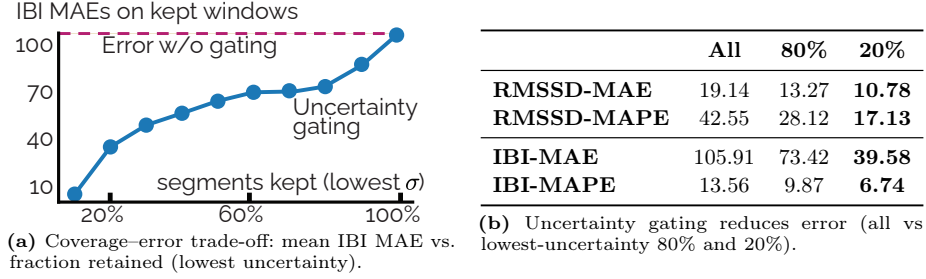


Fig. 7: Uncertainty correlates with error and supports reliability gating.

Finally, treating high-error windows as failures ($|\text{IBI error}| > 50 \text{ ms}$), uncertainty predicts these segments with $\text{AUROC} = 0.715$, confirming that uncertainty is informative for trustworthy prediction.

E.4 Breakdown of errors in egoPPG-DB

We analyze error variation across activities. Errors are lowest in low-motion segments (e.g., standing or light head movements) and increase in high-motion activities. Table 12 reports the per-session results as reported in egoPPG-DB.

Table 12 shows substantial variation across activities. In low-motion sessions (e.g., office and kitchen), errors drop markedly, and the resulting HR/HRV estimates approach ranges commonly reported for contact-based estimation in non-clinical settings [2, 11]. During high-motion activities, errors increase, but our uncertainty estimates rise accordingly. This enables identifying unreliable

Table 12: Per-activity breakdown on egoPPG-DB at 4 s resolution. We report our method and PulseFormer [5].

Activity	HR MAE		IBI MAE [ms]		SDNN MAE [ms]		RMSSD MAE [ms]	
	PF	Ours	PF	Ours	PF	Ours	PF	Ours
Kitchen	18.32	10.35	125.1	88.50	19.13	8.90	25.17	16.53
Office	16.93	8.83	115.0	80.73	17.64	7.80	30.68	14.66
Dancing	26.01	14.21	165.6	118.0	41.30	12.80	60.15	23.87
Bike	34.48	17.87	210.5	141.7	55.19	16.90	90.68	31.20
Walking	16.50	10.66	122.8	93.01	33.75	10.10	50.23	19.63
Overall	22.10	12.72	147.5	105.9	33.71	10.89	53.90	19.14

windows for down-weighting or exclusion to improve the trustworthiness of the remaining segments. We also observe a direct downstream benefit. When we remove the uncertainty from PhysFusion, the performance decreases for the proficiency classification. (Table 4).

E.5 Further results regarding the sampling rate

We recorded an additional dataset with 10 subjects at 90 fps to evaluate whether higher temporal resolution improves HR and HRV estimation. We compare the same pipeline at 30 fps and 90 fps without changing the temporal length of the windows (4 or 60 s). Results in Table 13 show that 90 fps yields only marginal gains, while 30 fps remains highly competitive for HR and sufficiently accurate for HRV indices. This supports using 30 fps as a practical operating point with lower compute and storage cost, without sacrificing performance.

Table 13: Comparison of heart rate (HR) and heart rate variability (HRV) estimation at 30 fps and 90 fps. The 90 fps setting yields slightly lower errors, while 30 fps remains sufficient for HRV indices and performs strongly for HR.

Method	HR								HRV (IBI) [ms]								HRV Time-Domain [ms]			
	MAE				RMSE				MAPE (%)				r				SDNN		RMSSD	
	4 s	60 s	4 s	60 s	4 s	60 s	4 s	60 s	4 s	60 s	4 s	60 s	4 s	60 s	4 s	60 s	MAE	MAPE	MAE	MAPE
Ours (30 fps)	8.85	5.33	12.72	8.90	10.03	5.78	0.66	0.91	81.73	62.16	102.93	86.48	8.05	6.33	0.71	0.85	7.90	22.15	15.02	33.57
Ours (90 fps)	8.72	5.12	12.45	8.58	9.41	5.41	0.68	0.92	78.60	60.90	99.10	84.20	7.85	6.15	0.74	0.86	7.55	21.10	14.20	30.10

In this fps configuration, we observe error rates in the same range as those reported for EgoPPG in the office scenario, indicating that our measurements and evaluation protocol are consistent and that 30 fps is sufficient for reliable HR and HRV estimation in this setting.

F Expanded discussion and related work

Why eye-video? Contact PPG requires consistent skin coupling and pressure [69], which can vary with eyewear fit, slippage during motion, and user comfort over long wear [22, 62]. These factors introduce practical failure modes and add mechanical overhead. Since many egocentric platforms already include inward-facing cameras for interaction [1, 20], our goal is to leverage existing sensing to provide a *no-additional-contact-hardware* option for physiological estimation. We do not argue that eye-video should replace contact PPG when the sensors are available; rather, we target settings where adding or relying on contact sensing is impractical (long-term wear or studies prioritizing minimal instrumentation).

How can 30 fps be feasible for HRV? HRV is timing-sensitive, so it is reasonable to question whether 30 Hz eye video can support HRV estimation. Our method estimates HRV from a BVP signal, where prior work has reported stable HRV estimates at low sampling rates (e.g., 25 Hz) by computing HRV on an interpolated beat-to-beat representation [61]. Empirically, downsampling BVP to 25 Hz has been shown to change SDNN and RMSSD only marginally (on the order of a few milliseconds) [11], suggesting that sampling rate alone is not the dominant failure mode when the waveform is clean.

In our setting, the main challenge is not the frame rate but the corruption in the recovered BVP. We address this by (i) improving BVP fidelity from 30 Hz video with low-high decomposition, and (ii) refining temporal structure before HRV computation via interpolation in a learned frequency-domain module (U-Net), rather than relying on raw frame-level peak timing. Finally, we model uncertainty. Accordingly, we position our method as a *non-clinical*, wearable HRV estimation for behavior modeling, with calibrated confidence to indicate when the estimates should be trusted.

Are 4 s estimates sufficient? Our approach predicts the mean IBI over non-overlapping 4 s windows, whereas classical HRV pipelines compute IBIs from individual R-peaks. In our setting, peak-to-peak IBIs are sensitive to noise and missed/spurious peaks, which can dominate the error. To assess whether 4 s windowing itself introduces a limiting bias, we ran an oracle analysis on the ECG. We computed window-level mean IBIs from the ground-truth ECG using the same 4 s windows and compared them to beat-level IBIs. The resulting discrepancy was only 5–6 ms, which is substantially smaller than the current prediction error. This suggests that the 4 s aggregation is not the main bottleneck.

Role of IMU IMU is used as an auxiliary motion context cue, not as a physiological sensor. Its placement is fixed in the device coordinate system. We therefore avoid relying on absolute IMU scale and treat IMU features as weak conditioning. Our method limits sensitivity to IMU orientation by using its magnitude while still benefiting from motion cues when available on the egocentric platform.

Generalization and dataset limitations. Our evaluation is constrained by publicly available datasets [5] with synchronized ground truth and egocentric eye video. This limits coverage of conditions such as outdoor illumination and extreme motions. We view this as a key next step where broader data collection and condition-wise analysis (motion intensity, lighting proxies, speaking activity) can better characterize operating regimes and guide more robust modeling.