

Change Point–Aware Evaluation and Re-Calibration of PPG-Based Blood Pressure Estimation

Yunwon Tae^{1,2,*}

Minje Park^{1,*}

Gyunho Rho^{1,3}

Dongjoon Yoo⁴

Sunghoon Joo^{1,†}

YTAE1@UCI.EDU

MINJE.PARK@VUNO.CO

GYUNHO.RHO@SBRI.CO.KR

DJINYOO@GMAIL.COM

SUNGHOOON.JOO@VUNO.CO

¹ VUNO Inc.

² University of California, Irvine

³ Research Institute for Future Medicine, Samsung Medical Center

⁴ Department of Critical Care Medicine and Emergency Medicine, Inha University Hospital

*Equal contribution. †Corresponding author.

Abstract

Non-invasive continuous blood pressure (BP) monitoring using photoplethysmography (PPG) is a promising alternative to cuff-based measurements. However, existing PPG-based BP estimation studies predominantly rely on aggregated performance metrics (e.g., mean absolute error) computed over entire evaluation intervals, which can obscure model failures during rapid BP fluctuations and limit clinical relevance. In this work, we propose a fluctuation-aware evaluation framework for PPG-based BP estimation based on time-series change point detection. Instead of heuristic BP thresholding (e.g., $\Delta BP > 10\text{mmHg}$), we identify BP change points by capturing abrupt distributional shifts in BP trajectories and evaluate estimation performance specifically during these fluctuation periods. Our analysis shows that several state-of-the-art models exhibit substantial performance degradation around BP change points, and that periodic test-time calibration is insufficient to handle such dynamic BP variations. To address this limitation, we introduce a targeted re-calibration framework triggered by detected BP change points, improving robustness without modifying model architectures. To the best of our knowledge, this is the first systematic evaluation of PPG-based BP estimation from a BP change point perspective, highlighting the importance of fluctuation-aware evaluation and calibration for real-world continuous BP monitoring.

1. Introduction

Continuous and non-invasive blood pressure (BP) monitoring plays a critical role in early detection of hemodynamic instability, guiding treatment decisions, and improving patient outcomes across both inpatient and ambulatory settings (Shen et al., 2025; Rastegar et al., 2020; Juri et al., 2018). With the growing demand for cuffless monitoring, photoplethysmography (PPG) has emerged as one of the most accessible and scalable sensing modalities, enabling continuous BP estimation through wearable devices and bedside monitoring systems (Mukkamala et al., 2025). Recent advances in deep learning have further accelerated this progress, with numerous PPG-based BP estimation models demonstrating promising

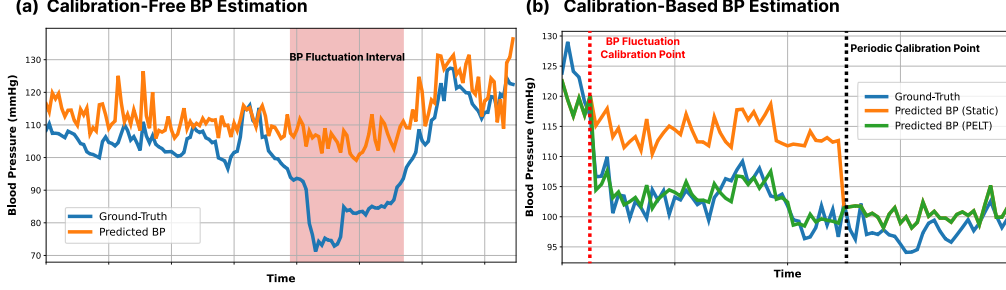


Figure 1: Limitations of existing PPG-based blood pressure (BP) estimation approaches under BP fluctuations. (a) Calibration-free methods track the overall BP trend but fail to capture rapid fluctuations (red region), resulting in large errors during clinically critical periods. (b) Calibration-based methods with periodic re-calibration (orange) struggle when BP shifts occur between calibration points. Targeted re-calibration at BP fluctuation points (green) enables timely adaptation and maintains accurate tracking throughout.

performance on public benchmarks and controlled experimental environments (Saha et al., 2025; Mohammadi et al., 2025; Ma et al., 2024).

From a methodological perspective, existing PPG-based BP estimation studies can largely be categorized into two research directions: *calibration-free* approaches and *calibration-based* approaches. Calibration-free methods aim to directly estimate BP from PPG signals without requiring subject-specific adjustment, focusing on learning generalized mappings between waveform morphology and BP values (Moulaeifard et al., 2025; Samimi and Dajani, 2023; Slapničar et al., 2019). In contrast, calibration-based methods leverage intermittent ground-truth BP measurements to correct model predictions, often improving subject-specific accuracy through periodic re-calibration (Joung et al., 2023; Jo et al., 2025).

Despite their methodological differences, both calibration-free and calibration-based approaches share a fundamental limitation: *they do not explicitly account for BP fluctuations*. BP is inherently dynamic and non-stationary, often exhibiting abrupt transitions driven by physiological stress, changes in posture, the effect of medication, or acute pathological events (Mukkamala et al., 2025; Yang et al., 2025; Mukkamala et al., 2021).

Figure 1 provides an intuitive illustration of these challenges. In Figure 1(a), the calibration-free estimation appears to be accurate when assessed throughout the whole BP intervals; however, substantial prediction errors emerge during the BP fluctuation intervals. In Figure 1(b), periodic calibration partially reduces long-term drift, yet fails to track sudden BP drops occurring between scheduled calibration points. Notably, when additional calibration is performed during fluctuation intervals, estimation trajectories more closely follow the ground-truth BP trend. These observations highlight that BP fluctuation awareness is essential for both evaluation and calibration design.

Motivated by this gap, our work systematically investigates the role of BP fluctuations in PPG-based BP estimation from both calibration-free and calibration-based perspectives. First, under the calibration-free setting, we explicitly evaluate model performance within fluctuation intervals and analyze whether simple strategies such as fluctuation-aware balanced sampling can improve robustness in these clinically critical regions. Second, under

the calibration-based setting, we examine whether aligning calibration timing with BP fluctuation events—beyond fixed periodic schedules—can further enhance estimation accuracy.

A key challenge in enabling such analyses lies in defining BP fluctuations in a principled manner. Rather than relying on heuristic threshold rules (e.g., $\Delta\text{BP} > 10\text{mmHg}$) (Wang et al., 2024; Hong et al., 2024), we formulate BP instability through time-series change point detection. Specifically, we employ the Pruned Exact Linear Time (PELT) algorithm (Killick et al., 2012) to identify distributional transitions in continuous BP signals, providing a data-driven definition of fluctuation intervals. Using these detected change points, we conduct fluctuation-aware evaluation in calibration-free settings and design targeted re-calibration strategies in calibration-based settings. Our findings reveal that calibration effectiveness depends not only on frequency but critically on timing relative to physiological transitions.

However, directly applying change point detection in real-world deployment introduces practical constraints. Conventional algorithms such as PELT operate in an offline setting, requiring access to ground-truth BP signals and future observations, making real-time fluctuation detection infeasible. To bridge this gap, we further propose a framework that leverages PPG inputs to detect BP fluctuation events in an online manner. This enables real-time identification of re-calibration timing without relying on contemporaneous invasive BP measurements.

In summary, this work makes the following contributions: (1) We introduce a fluctuation-aware evaluation paradigm for PPG-based BP estimation, analyzing model behavior specifically within intervals exhibiting abrupt blood pressure changes. (2) We demonstrate that targeted re-calibration at BP change points significantly improves estimation accuracy over periodic re-calibration alone. (3) We propose an online change point detection framework using only PPG signals, enabling targeted re-calibration without explicit BP monitoring.

To the best of our knowledge, this is the first study to investigate PPG-based BP estimation through the lens of BP change point dynamics, highlighting the necessity of fluctuation-aware evaluation and calibration for clinically reliable continuous BP monitoring.

Generalizable Insights about Machine Learning in the Context of Healthcare

This paper offers the following generalizable insights for machine learning in healthcare:

- Standard aggregate evaluation metrics (e.g., mean absolute error over entire recordings) can systematically mask model failures during clinically critical physiological transitions; fluctuation-aware evaluation protocols are essential for a reliable assessment of real-world performance.
- The timing of calibration matters at least as much as its frequency: targeted re-calibration at change points detected from PPG alone outperforms more frequent but temporally uninformed periodic re-calibration strategies.
- Online change point detection driven by self-supervised contrastive PPG representations provides a practical, architecture-agnostic trigger for adaptive calibration in continuous monitoring pipelines.

2. Related Work

2.1. Calibration-Free PPG to BP Estimation

Calibration-free PPG to BP estimation has traditionally relied on Pulse Wave Analysis (PWA) and Pulse Transit Time (PTT), where handcrafted morphological features or timing intervals are mapped to BP using conventional machine learning models (Kachuee et al., 2015). More recently, deep learning architectures have been introduced to learn representations directly from raw PPG signals, demonstrating competitive estimation performance on benchmark datasets (Moulaeifard et al., 2025; Samimi and Dajani, 2023; Slapničar et al., 2019). Despite these advances, prior studies primarily report aggregate metrics computed over entire BP values. Such evaluation protocols may obscure model behavior during abrupt blood pressure fluctuations, where estimation reliability is clinically most critical. This highlights the need for fluctuation-aware analysis to better understand model robustness.

2.2. Calibration-Based PPG to BP Estimation

Calibration-based PPG to BP estimation approaches incorporate subject-specific references to personalize BP predictions (Joung et al., 2023; Jo et al., 2025). These methods typically leverage periodic cuff-based measurements to adjust model outputs and compensate for inter-individual variability, improving estimation accuracy in real-world deployment settings. However, existing calibration paradigms commonly rely on fixed periodic schedules, implicitly assuming stable physiological conditions between calibration points (Vybornova et al., 2021). Given the inherently dynamic nature of blood pressure, such strategies may fail to provide timely correction during abrupt fluctuation intervals. Determining when calibration should be performed, therefore, remains underexplored, which we investigate in this work.

2.3. BP Change Point Detection

While explicit change point detection has rarely been explored in the context of BP estimation, several prior studies have addressed the concept of BP change from a different perspective. They primarily focus on tracking relative fluctuations rather than estimating absolute values (Wang et al., 2024; Hong et al., 2024), typically defining BP change as an event where the magnitude of variation over a specific time window exceeds a fixed threshold (e.g., $\Delta BP > 10\text{mmHg}$). However, their primary objective is to bypass the difficulty of regression, rather than to enable fluctuation-aware evaluation and determine optimal timing for re-calibration. Furthermore, relying on fixed absolute thresholds fails to account for inter-subject variability and neglects the temporal structure of BP sequences. In contrast, we approach identifying these transitions through time-series change point detection, enabling a statistically grounded criterion for re-calibration.

2.4. Change Point Detection Algorithms in a Time-Series Domain

Change Point Detection (CPD) is the process of partitioning a time-series sequence into distinct, statistically homogeneous segments by identifying abrupt shifts in the underlying distribution. CPD algorithms are categorized into offline (Killick et al., 2012; Gharghabi et al., 2018) and online (Adams and MacKay, 2007; Gharghabi et al., 2018) settings. Offline

CPD retrospectively segments the entire sequence using full historical data. In contrast, online CPD detects changes in real-time using only historical data, making it suitable for streaming applications.

For re-calibration in continuous BP monitoring, online CPD is necessary to trigger re-calibration in real-time when BP changes occur. However, in the deployment setting, the reference BP cannot be obtained, making existing online BP change point detection methods inapplicable. These methods fundamentally rely on ground-truth BP measurements to identify distributional shifts. Therefore, we need a CPD method that works with PPGs alone, without requiring BP values as input.

3. Problem Formulation

3.1. Calibration-Free PPG to BP Estimation

The standard PPG to BP estimation is defined as a regression task that predicts systolic (SBP) and diastolic (DBP) blood pressure using a fixed-length PPG segment. Let $\mathbf{x} \in \mathbb{R}^L$ denote the normalized PPG input of length L , and $y \in \mathbb{R}^+$ denote the corresponding scalar blood pressure value. The estimation model is formulated as a function $\hat{y} = f(\mathbf{x}; \theta)$, parameterized by θ . The objective is to learn the optimal parameters θ^* that minimize the discrepancy between the prediction $\hat{y}$ and the ground-truth y , typically measured by the Mean Squared Error (MSE) loss.

3.2. Blood Pressure Fluctuations and Change Points

Blood pressure is inherently non-stationary, exhibiting rapid fluctuations due to physiological stress or postural changes. Standard evaluation protocols, which aggregate errors over the entire recording, often mask significant performance degradation during these critical transitions. Models with low aggregated MAE can still produce clinically unacceptable errors during critical hemodynamic shifts.

To address this gap, we explicitly define **BP change points** as time steps where the underlying BP distribution undergoes a significant shift. Formally, given a BP sequence $y_{1:T}$, a change point τ is defined where the BP distribution $p(y_t)$ changes ($p(y_t | t < \tau) \neq p(y_t | t \geq \tau)$). In this work, we derive ground-truth BP change points using the **PELT** algorithm (Killick et al., 2012) on the reference BP sequence.

Crucially, we define **Unstable BP Intervals** as the segments, derived from detected change points, exhibiting severe hypertensive or hypotensive levels (SBP > 180 mmHg or mean arterial pressure < 65 mmHg) (Bress et al., 2024; Evans et al., 2021). To assess robustness under these clinically critical conditions, we introduce a dual-evaluation protocol measuring accuracy on both (1) all samples for baseline performance and (2) the samples within unstable intervals.

3.3. Calibration-Based PPG to BP Estimation

Beyond the temporal fluctuations described above, BP estimation faces the challenge of significant inter-subject variability. Calibration-based estimation addresses this by incorporating subject-specific reference data. This task extends the input space to include a calibration set $\mathcal{D}_{cal} = \{(\mathbf{x}_{cal}, y_{cal})\}$, acquired from the subject prior to the inference time

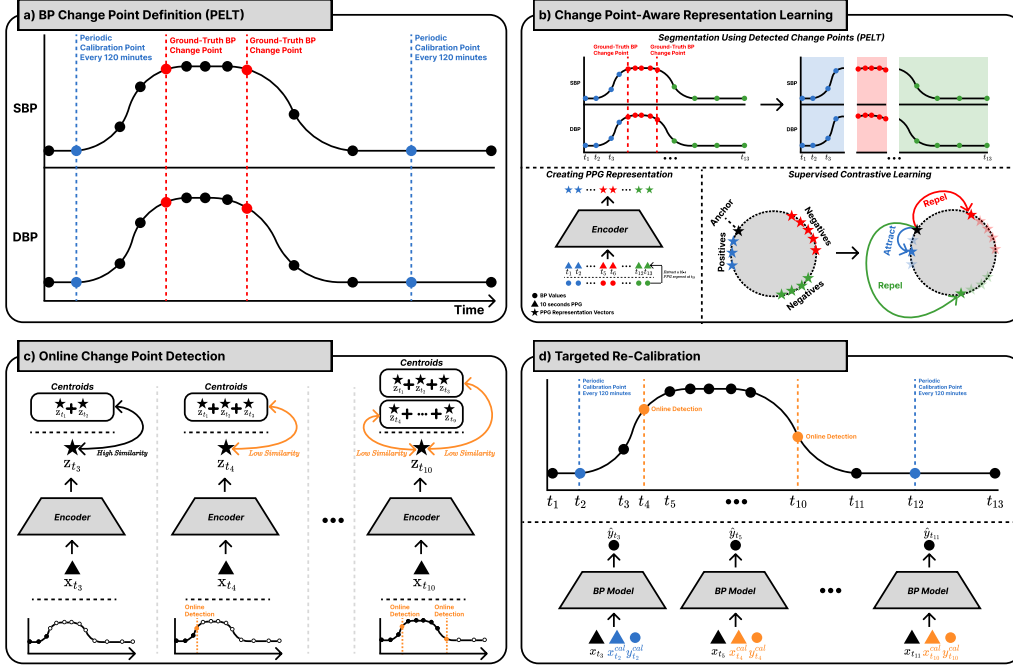


Figure 2: Overview of the proposed targeted re-calibration framework. (a) Ground-truth blood pressure (BP) change points are obtained from the reference BP sequence. (b) Supervised contrastive learning trains a PPG encoder to attract features within the same segment (defined by change points) and repel those from different segments. (c) Online change point detection via sequential clustering: PPG features are sequentially compared against cluster centroids; low similarity to the current cluster triggers change point detection. (d) Detected change points supplement periodic calibration, enabling timely re-calibration upon physiological shifts.

t . The goal is to estimate the target blood pressure y_t given the current input x_t and the calibration set $\mathcal{D}_{cal}$. We categorize existing approaches into two frameworks based on how they utilize $\mathcal{D}_{cal}$.

Input-Level Calibration This framework utilizes the calibration data as an auxiliary input to the model, formulated as $\hat{y}_t = f(x_t, x_{cal}, y_{cal}; \theta)$. For example, **PPG2BP-Net** (Joung et al., 2023) computes differential features between x_t and x_{cal} , and combines them with calibration BP values y_{cal} to regress the target value. In this setting, the parameters θ remain fixed during the inference phase.

Parameter-Level Calibration This framework, exemplified by **Test-Time Calibration** (Jo et al., 2025), formulates calibration as an online adaptation process. Instead of fixing the model parameters, the calibration set $\mathcal{D}_{cal}$ is used to dynamically update the parameters from an initial state θ to a personalized state θ' . The inference is then performed as $\hat{y}_t = f(x_t; \theta')$, allowing the model to adapt its internal representations to the subject’s physiological state.

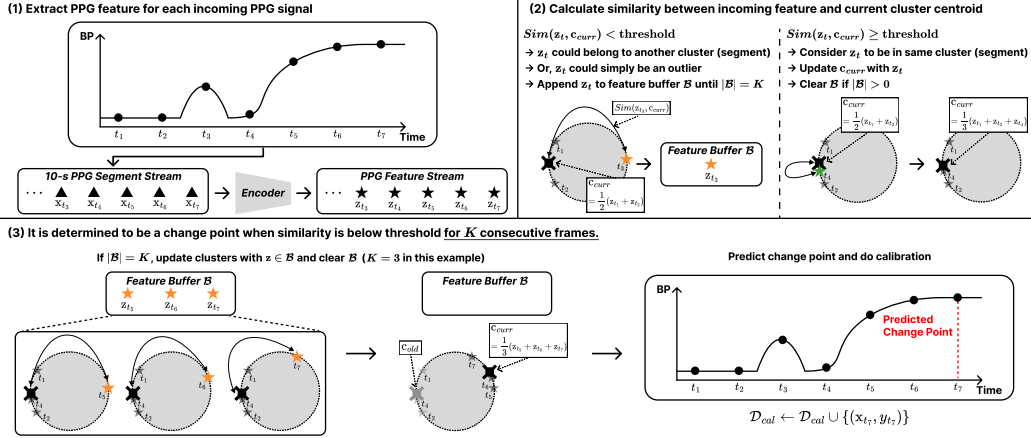


Figure 3: An illustration of a sequential clustering algorithm for PPG-based blood pressure change point detection and re-calibration. (1) The PPG feature (z_t) is extracted sequentially from the incoming PPG signal (x_t). (2) Similarity between incoming features and the current cluster centroid (c_{curr}) determines whether the physiological state changes. (3) The change point is detected when similarity remains below the threshold for K consecutive frames, triggering cluster update and re-calibration.

3.4. Targeted Re-Calibration

Existing calibration methods rely on periodic re-calibration at fixed intervals. However, when BP fluctuates between calibration points, the model operates on outdated references, leading to degraded accuracy.

We propose **Targeted Re-Calibration**, which supplements periodic calibration with additional re-calibration triggered by detected BP change points mentioned in Section 3.2. This approach ensures timely adaptation when physiological shifts occur while maintaining regular updates otherwise.

The challenge is that detecting BP change points typically requires continuous BP monitoring. In Section 4, we address this by proposing an online change point detection framework that operates solely on PPG signals, enabling targeted re-calibration without explicit BP measurements.

4. Targeted Re-Calibration with Online Blood Pressure Change Point Detection

To enable targeted re-calibration, we propose an online CPD framework based solely on PPG signals. The overall process is illustrated in Figure 2. Our approach employs contrastive learning (Chen et al., 2020) to learn discriminative PPG features, then applies sequential clustering to predict BP change points in real-time, eliminating the need for explicit BP tracking. The predicted change points serve as immediate triggers for re-calibration.

4.1. Supervised Contrastive Learning for Segment-Aware PPG Representation

To detect BP change points effectively without direct BP measurements, we require a PPG representation space where distinct physiological states are separable. We achieve this by training a PPG feature encoder using **Supervised Contrastive Learning (SCL)** (Khosla et al., 2020), guided by reference BP change points.

Let $\mathbf{x}_{1:T} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)$ denote the sequence of input PPG signals. We define an encoder $h(\cdot; \phi)$ that maps each PPG signal $\mathbf{x}_t$ to a normalized feature embedding $\mathbf{z}_t = h(\mathbf{x}_t; \phi) \in \mathbb{R}^d$, where d denotes the feature dimension. Assume we are given a set of ground-truth BP change points $\mathcal{T} = \{\tau_1, \tau_2, \dots, \tau_M\}$. For each time step t that falls in the segment $[\tau_m, \tau_{m+1})$, we define the set of positive indices $P(t) = \{p \mid \tau_m \leq p < \tau_{m+1}, p \neq t\}$, meaning all other time steps that belong to the same segment. The encoder is then optimized using the supervised contrastive loss:

$$\ell_{SCL}(t) = \frac{-1}{|P(t)|} \sum_{p \in P(t)} \log \frac{\exp(\mathbf{z}_t \cdot \mathbf{z}_p / \gamma)}{\sum_{a \in A(t)} \exp(\mathbf{z}_t \cdot \mathbf{z}_a / \gamma)},$$

where $A(t) = \{a \mid 1 \leq a \leq T, a \neq t\}$ and γ is a scalar temperature parameter. By minimizing this objective, the encoder learns to cluster PPG samples from the same segment while pushing apart those from different segments, ensuring that BP fluctuations result in distinct shifts in the feature space. Once trained, this encoder is deployed for online CPD.

4.2. Online Change Point Detection via Sequential Clustering

To identify BP change points in real-time, we employ a **Sequential Clustering** algorithm on the feature stream $\mathbf{z}_{1:T}$. The algorithm maintains a dynamic set of cluster centroids $\mathcal{C}$, tracking an **active centroid** $\mathbf{c}_{curr}$ that represents the current physiological state.

The core logic operates by sequentially checking the cosine similarity between the incoming feature $\mathbf{z}_t$ and the active centroid ($\text{Sim}(\mathbf{z}_t, \mathbf{c}_{curr})$). If $\mathbf{z}_t$ remains sufficiently close to the active centroid with a threshold η ($\text{Sim}(\mathbf{z}_t, \mathbf{c}_{curr}) \geq \eta$), we update $\mathbf{c}_{curr}$ via incremental averaging:

$$\mathbf{c}_{curr} \leftarrow \frac{N_{curr} \cdot \mathbf{c}_{curr} + \mathbf{z}_t}{N_{curr} + 1}, N_{curr} \leftarrow N_{curr} + 1 \quad (1)$$

where N_{curr} is the number of samples accumulated in the current state.

However, when a significant deviation occurs ($\text{Sim}(\mathbf{z}_t, \mathbf{c}_{curr}) < \eta$), the algorithm searches for the centroid with the highest similarity to $\mathbf{z}_t$ among all previously stored centroids, denoted $\mathbf{c}^*$. If a match is found ($\text{Sim}(\mathbf{z}_t, \mathbf{c}^*) \geq \eta$), the algorithm transitions to that existing state, updating the active centroid accordingly:

$$\mathbf{c}_{curr} \leftarrow \frac{N^* \cdot \mathbf{c}^* + \mathbf{z}_t}{N^* + 1}, N_{curr} \leftarrow N^* + 1 \quad (2)$$

where N^* is the number of samples accumulated in the matched state $\mathbf{c}^*$. Otherwise, a new centroid is initialized with $\mathbf{z}_t$:

$$\mathbf{c}_{curr} \leftarrow \mathbf{z}_t, N_{curr} \leftarrow 1, \mathcal{C} \leftarrow \mathcal{C} \cup \{\mathbf{z}_t\}. \quad (3)$$

Any such transition flags a change point and triggers re-calibration. The overview is illustrated in Figure 3.

Table 1: PulseDB Dataset Statistics for MIMIC-III and VitalDB.

	MIMIC-III		VitalDB	
	Train	Test	Train	Test
N. Subjects/Cases	1,213/2,095	135/236	1,293/1,347	144/146
N. 10s Segments	436,680	54,000	465,480	57,600
N. Unstable Segments	53,518	6,085	54,507	7,132
SBP (mean $\pm$ SD)	121.94 $\pm$ 22.58	122.47 $\pm$ 21.83	115.48 $\pm$ 18.93	115.43 $\pm$ 18.72
DBP (mean $\pm$ SD)	60.73 $\pm$ 13.14	60.91 $\pm$ 12.58	62.92 $\pm$ 12.08	63.02 $\pm$ 11.87

4.3. Practical Refinements for Robust Change Point Detection

The sequential clustering algorithm described in Section 4.2 assumes a fixed threshold η and triggers a state transition immediately upon detecting a deviation. In practice, this naive approach is prone to false positives: even a single outlier can exceed the threshold, and the PPG features tend to be unstable right after a genuine state change. Raising the threshold mitigates this but risks missing subtle shifts. To address these issues, we incorporate two refinements.

Duration-Dependent Thresholding The fixed threshold η in Section 4.2 is replaced with a duration-dependent variant $\eta = \eta_{base} \cdot (1 - \exp(-\lambda n))$, where n denotes the duration of the current segment, measured as the number of frames elapsed since the last detected change point. Here, η_{base} denotes the target threshold and λ controls how quickly the threshold reaches it. The threshold starts low right after a state transition to prevent false change points when the features are still unstable, and gradually tightens to η_{base} as the state stabilizes.

Delayed Prediction To filter out short-term outliers, we enforce a persistence constraint. When a deviation occurs, z_t is held in a feature buffer $\mathcal{B}$ rather than triggering an immediate transition. If no deviation is detected, $\mathcal{B}$ is reset. A change point is confirmed only if the deviation persists for K consecutive frames (i.e., $|\mathcal{B}| = K$). Upon confirmation, clusters are updated with the buffered features following Equations (2) and (3).

We present a more detailed procedure in Algorithm 1, which can be found in Appendix A.3.

5. Cohort

5.1. Dataset

We conduct our experiments using the publicly available PulseDB dataset (Wang et al., 2023), a large-scale benchmark for cuff-less blood pressure (BP) estimation. PulseDB integrates waveform data from the MIMIC-III (Johnson et al., 2016) and VitalDB (Lee et al., 2022) datasets, comprising electrocardiogram (ECG), photoplethysmogram (PPG), and arterial blood pressure (ABP) signals.

The dataset consists of 10-second waveform segments sampled at 125 Hz. To ensure high signal reliability, PulseDB applies strict quality control criteria: segments without

complete cardiac cycles are excluded, PPG segments with negative skewness are removed, and samples with a correlation lower than 0.9 between PPG and ABP are discarded. As a result of this filtering, the retained segments exhibit irregular temporal spacing across subjects. Ground-truth systolic and diastolic blood pressure (SBP and DBP) values are derived directly from the ABP waveform.

We use split dataset provided from the PulseDB dataset, which is subject-independent training and test sets for both MIMIC-III and VitalDB sources. Table 1 summarizes the dataset statistics, including the number of subjects and cases, total number of 10-second segments, number of segments of unstable BP interval, and the distribution of SBP and DBP values across splits.

5.2. Preprocessing

All PPG signals are processed using a band-pass filter to suppress baseline drift and high-frequency noise. For experiments that leverage the Pulse-PPG pretraining model (Saha et al., 2025), the filtered signals are resampled to 50 Hz to ensure compatibility with the pretraining configuration. In all other experimental settings, the original sampling rate of 125 Hz provided by the PulseDB dataset is retained. Before being fed into the models, each PPG segment is standardized using z-normalization, such that the input has zero mean and unit variance. This preprocessing pipeline is applied consistently across all experiments.

5.3. Baselines

We consider both calibration-free and calibration-based BP estimation baselines to comprehensively evaluate the proposed framework. We additionally include multiple BP change point definitions in the calibration-based setting.

Calibration-Free BP Estimation We evaluate four calibration-free baselines: *Inception1D* (Ismail Fawaz et al., 2020; Moulaeifard et al., 2025), *PaPaGei-S* (Pillai et al., 2025), *Pulse-PPG* (Saha et al., 2025), and *AnyPPG* (Nie et al., 2025). Except for *Inception1D*, which is trained from scratch, all models are initialized from pretrained checkpoints. All models are fine-tuned on PulseDB using the same training protocol for fair comparison. We adopt pretrained models to leverage robust and generalizable PPG representations, which have been shown to be effective in handling signal variability and distribution shifts. Such properties are particularly important for evaluating model performance under BP fluctuation scenarios. Furthermore, representation learning-based models are well suited for our setting, as the proposed framework relies solely on PPG signals to detect online BP change points without explicit BP measurements. A discriminative and stable PPG representation is therefore critical for reliable online change point detection.

Calibration-Based BP Estimation For calibration-based baselines, we consider two representative calibration methods. First, we evaluate an input-level calibration approach using **PPG2BP-Net** (Joung et al., 2023), where the model parameters are fixed during inference and calibration data are provided as additional inputs. Second, we adopt **Test-Time Calibration (TTC)** (Jo et al., 2025), a parameter-level calibration method that dynamically updates model parameters at inference time using calibration samples. **TTC**

enables personalized adaptation by optimizing model parameters based on subject-specific calibration data and serves as a strong baseline for dynamic BP estimation.

Different BP Change Point Definitions Periodic calibration is fixed at a 120-minute duration across all experiments, reflecting a clinically reasonable setting for ICU and operating room data (MIMIC and VitalDB), with additional ablations using shorter intervals reported in Appendix B. In calibration-based experiments, we evaluate multiple BP change point definitions to disentangle the effect of re-calibration timing from re-calibration frequency. In addition to change points detected by *PELT*, we consider a ΔBP criterion, where a change point is triggered when the absolute BP difference exceeds 10 mmHg, following prior work (Wang et al., 2024). We also include a *Random* baseline that matches the number of change points detected by *PELT* while randomly sampling their temporal locations. This design controls for re-calibration frequency and isolates the importance of where re-calibration is applied.

5.4. Training Details

For calibration-based baselines, including **PPG2BP-Net** and **TTC**, we follow the original model architectures and hyperparameter settings. For **TTC**, although the original formulation leverages both ECG and PPG signals, we modify the input to use PPG only for consistency with the cuffless BP estimation setting. Calibration-free baselines are implemented following their original publications. Minor deviations in hyperparameter settings are introduced only for the BP estimation fine-tuning stage and the online CPD module; details are provided in Appendix A.2.

6. Results on Real Data

6.1. Fluctuation-Aware Evaluation

Table 2 presents a fluctuation-aware evaluation of blood pressure (BP) estimation models, comparing performance over the entire recording (Whole) and over unstable BP intervals.

Across all models and training strategies, BP estimation errors consistently increase during unstable intervals compared to whole-interval evaluation. Specifically, we observe an average degradation of approximately 2–4 mmHg MAE for systolic BP (SBP) and 1–2 mmHg for diastolic BP (DBP). This trend highlights that standard evaluation over the full time-series substantially underestimates model errors during clinically important BP fluctuation periods.

Notably, models trained with the interval-balanced strategy (i.e., balanced sampling strategy) demonstrate improved robustness during unstable BP intervals compared to their vanilla counterparts. While whole-interval performance remains comparable, interval-balanced training consistently reduces SBP estimation error in unstable regions (e.g., from 17.04 to 16.40 mmHg for Inception1D and from 17.84 to 16.41 mmHg for Pulse-PPG). This suggests that explicitly balancing training samples across stable and unstable BP segments encourages models to better capture abrupt hemodynamic changes.

Overall, these results emphasize that accurate BP estimation during unstable intervals remains significantly more challenging than during stable periods, even for state-of-the-art PPG-based models. More importantly, they demonstrate the necessity of fluctuation-aware

Table 2: Comparison of BP estimation performance on whole time-series segments and clinically unstable BP intervals. Mean absolute error in mmHg with 95% confidence intervals are reported. Baseline (Median) represents the result of predicting the training set median for all samples.

Models	Whole		Unstable BP Intervals	
	SBP ↓	DBP ↓	SBP ↓	DBP ↓
Baseline (Median)	16.48 ± 0.04	9.75 ± 0.02	21.10 ± 0.07	11.85 ± 0.04
Vanilla				
Inception1D	13.79 ± 0.03	8.36 ± 0.02	17.04 ± 0.07	8.86 ± 0.04
PaPaGei-S	13.68 ± 0.03	8.66 ± 0.02	15.81 ± 0.07	9.12 ± 0.04
Pulse-PPG	13.72 ± 0.03	8.37 ± 0.02	17.84 ± 0.07	9.38 ± 0.04
AnyPPG	14.79 ± 0.04	8.41 ± 0.02	17.40 ± 0.07	10.03 ± 0.04
Interval-balanced				
Inception1D	13.67 ± 0.03	8.17 ± 0.02	16.40 ± 0.07	10.08 ± 0.04
PaPaGei-S	13.82 ± 0.03	8.70 ± 0.02	15.34 ± 0.07	8.38 ± 0.04
Pulse-PPG	14.00 ± 0.03	8.41 ± 0.02	16.41 ± 0.07	9.48 ± 0.04
AnyPPG	13.92 ± 0.03	8.41 ± 0.02	17.32 ± 0.07	10.30 ± 0.04

evaluation and training strategies to avoid overly optimistic performance assessments and to better reflect real-world clinical deployment scenarios.

6.2. Importance of Re-Calibration

Calibration is a widely adopted strategy in PPG-based blood pressure (BP) estimation to mitigate inter-subject variability. However, most prior work (Vybornova et al., 2021; Nachman et al., 2025) relies on fixed periodic calibration schemes (e.g., calibrating every week or month), implicitly assuming that BP dynamics evolve smoothly over time. Such approaches overlook a critical question: *when* calibration is actually needed.

Table 3 investigates the impact of different re-calibration strategies under a fixed base calibration interval of 120 minutes. While periodic calibration provides a reasonable baseline, performance improves substantially when additional calibration points are introduced in a targeted manner. In particular, change point-aware re-calibration using *PELT* yields the largest error reduction across both **TTC** and **PPG2BP-Net**, achieving the lowest SBP and DBP MAE.

Importantly, the results reveal that improved performance is not simply a consequence of increased calibration frequency. Although the ΔBP strategy introduces significantly more calibration points (17.36 points per case), its performance remains inferior to the *PELT*-based approach, which uses nearly half as many calibration points (9.82 points per case). Similarly, random re-calibration, despite having a comparable number of calibration points, fails to consistently match the performance of the change point-driven strategy.

Table 3: Impact of targeted re-calibration using ground-truth (offline) blood pressure change points. Mean absolute error in mmHg with 95% confidence intervals is reported. The bold represents the significant difference ($p < 0.05$) against other baseline methods. Points / Case denotes the average number of re-calibration points per case (mean $\pm$ std). The plus sign (“+”) keeps the 120-minute schedule as the base and adds the change points of strategies such as ΔBP , *Random*, and *PELT*. *Random* uses the per-case point count of *PELT* with uniformly sampled locations. Additional results for the unstable subgroup are provided in Appendix F.1.

Calibration Points	TTC		PPG2BP-Net		Points / Case
	SBP $\downarrow$	DBP $\downarrow$	SBP $\downarrow$	DBP $\downarrow$	
Every 120 minutes	11.28 $\pm$ 0.06	6.22 $\pm$ 0.03	9.33 $\pm$ 0.05	5.21 $\pm$ 0.03	3.30 $\pm$ 1.52
+ ΔBP	9.89 $\pm$ 0.05	5.34 $\pm$ 0.03	8.05 $\pm$ 0.05	4.53 $\pm$ 0.03	17.36 $\pm$ 22.57
+ <i>Random</i>	9.27 $\pm$ 0.05	5.08 $\pm$ 0.03	7.44 $\pm$ 0.04	4.01 $\pm$ 0.03	9.91 $\pm$ 4.02
+ <i>PELT</i>	8.97 $\pm$ 0.05	4.94 $\pm$ 0.03	5.84 $\pm$ 0.04	3.85 $\pm$ 0.02	9.82 $\pm$ 4.02

These findings highlight that indiscriminate or overly frequent calibration is neither efficient nor optimal. Instead, selectively triggering re-calibration during physiologically unstable periods enables models to adapt more effectively to abrupt BP shifts while minimizing unnecessary calibration overhead. This underscores the importance of *fluctuation-aware* re-calibration strategies for reliable continuous BP monitoring in real-world settings.

6.3. Online Change Point Detection

Following the formulation described in Sections 4.1 and 4.2, all models are trained to identify BP change points in an online manner using 10-second PPG segments as input.

As shown in Table 4, we evaluate CPD performance using Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), weighted accuracy (Mosley, 2013), and Intersection of Union (IoU) (Van den Burg and Williams, 2020), reflecting the substantial class imbalance inherent to online BP change point detection, where fluctuation events are sparse. Among the evaluated models, Inception1D achieves the highest weighted accuracy (0.6457) and IoU (0.6149), while Pulse-PPG slightly outperforms others in terms of AUROC (0.7308). Overall, the evaluated models exhibit comparable detection performance across metrics.

Importantly, the objective of this experiment is not to establish a new state-of-the-art CPD method, but to assess the practical feasibility of integrating online BP change point detection as a trigger for targeted re-calibration. The results demonstrate that existing CPD models, when operated in an online setting, can reliably capture BP fluctuation intervals with sufficient accuracy to support downstream re-calibration strategies. This suggests that online CPD can serve as a viable and modular component within BP monitoring pipelines, and that future advances in CPD methodology may further enhance the effectiveness of fluctuation-aware re-calibration.

Table 4: Online Change Point Detection Performance.

Models	AUROC	AUPRC	Weighted Accuracy	IoU
Inception1D	0.7275	0.3079	0.6457	0.6149
PaPaGei-S	0.7223	0.3070	0.6412	0.6062
Pulse-PPG	0.7308	0.3066	0.6112	0.6099
AnyPPG	0.7286	0.3075	0.6176	0.6106

Table 5: Performance of targeted re-calibration using the proposed online change point detection. Mean absolute error in mmHg with 95% confidence intervals is reported. The bold represents the significant difference ($p < 0.05$) against other baseline methods. Points / Case denotes the average number of re-calibration points per case (mean $\pm$ std). The plus sign (“+”) keeps the 120-minute schedule as the base and adds the change points of strategies such as *Random* and *Online CPD*. *Random* uses the per-case point count of *Online CPD* with uniformly sampled locations. Both *Classification* and *Online CPD* use *Inception1D* as the CPD backbone. Additional results for the unstable subgroup are provided in Appendix F.3.

Calibration Points	TTC		PPG2BP-Net		Points / Case
	SBP $\downarrow$	DBP $\downarrow$	SBP $\downarrow$	DBP $\downarrow$	
Every 120 minutes	11.28 $\pm$ 0.06	6.22 $\pm$ 0.03	9.33 $\pm$ 0.05	5.21 $\pm$ 0.03	3.30 $\pm$ 1.52
+ <i>Random</i>	9.52 $\pm$ 0.05	5.20 $\pm$ 0.03	7.60 $\pm$ 0.04	4.07 $\pm$ 0.03	8.77 $\pm$ 3.34
+ <i>Online CPD</i>	9.24 $\pm$ 0.05	5.12 $\pm$ 0.03	7.27 $\pm$ 0.04	3.87 $\pm$ 0.02	8.77 $\pm$ 3.33

6.4. Effectiveness of Online CPD-Based BP Re-Calibration

Offline change point-based re-calibration has already demonstrated the importance of calibration timing over fixed or random strategies in Table 3. Here, we focus on whether this benefit can be retained in an online setting, where BP change points must be detected sequentially without access to future information.

As shown in Table 5, *Online CPD*-based re-calibration consistently outperforms periodic and *Random* strategies across both **TTC** and **PPG2BP-Net**. Despite relying solely on online change point detection, the proposed method achieves comparable or better accuracy while using a similar number of calibration points per case.

These results suggest that BP change points can be effectively identified and exploited in real time, enabling practical online re-calibration for continuous BP monitoring.

7. Discussion

Summary of findings In this work, we demonstrate that standard aggregate evaluation masks significant model failures during hemodynamic fluctuations. By introducing a fluctuation-aware protocol, we reveal consistent performance degradation in calibration-free models during these critical intervals—an average increase of 2–4 mmHg MAE for SBP

and 1–2 mmHg for DBP—that would otherwise go unnoticed under conventional whole-recording metrics. This finding suggests that fluctuation-stratified reporting should serve as a standard complement to whole-interval evaluation, particularly when assessing model suitability for patients with highly dynamic BP profiles. For calibration-based approaches, we establish that the timing of re-calibration is at least as important as its frequency: change point-triggered re-calibration using *PELT* outperforms both periodic-only and ΔBP strategies while using fewer calibration points, indicating that re-calibration design should prioritize alignment with physiological transitions rather than simply increasing update rates. Finally, by training a contrastive PPG encoder to detect these transitions online without access to ground-truth BP, we show that the benefit of targeted re-calibration can be retained in practical deployment, enabling timely re-calibration in ambulatory or step-down settings where invasive BP references are unavailable.

Why change point timing helps. The re-calibration triggers we compare differ in what each treats as a blood pressure fluctuation. The ΔBP criterion is a local, univariate rule that fires whenever the systolic change within a short window exceeds a fixed threshold, using SBP alone with no model of the underlying distribution, so it reacts to transient excursions and noise and yields redundant change points. *PELT* instead detects distributional shifts in the joint SBP and DBP sequence, placing change points only where the signal statistics genuinely change, which is why its sparser triggers align more closely with true hemodynamic transitions.

Linking detection quality to re-calibration gain. The re-calibration gains in Table 5 depend on how well the online module detects true change points, not merely on triggering additional updates. To make this relationship explicit, we compare our formulation against a naive binary classifier trained on the same *PELT*-derived labels (Table 11). Detection quality and downstream accuracy move together: raising AUROC from 0.5332 to 0.7275 lowers SBP MAE from 8.31 to 7.27 mmHg under the same **PPG2BP-Net** estimator, and the same ordering holds for the PaPaGei-S backbone. The consistency of this pattern across a from-scratch and a pretrained encoder indicates that the gain originates from the detection formulation rather than the choice of backbone.

What transitions the detector captures. The detected change points concentrate on a specific class of physiological event. Cross-referencing the unstable intervals with VitalDB vasopressor records (Table 14), the large majority of unstable segments (85.4%) occur under vasopressor administration, indicating that *PELT* identifies pharmacologically driven hemodynamic shifts rather than arbitrary statistical fluctuations. These are also the events during which a stale calibration reference is most likely to mislead the estimator, which supports change point triggering as a clinically meaningful criterion.

Limitations The current framework relies on PulseDB data drawn from ICU and operating room settings; generalizability to ambulatory or wearable monitoring contexts remains to be validated. The online CPD module requires an initial training phase using ground-truth BP change points, which may limit applicability in fully unsupervised deployment scenarios. The *PELT* penalty hyperparameter was fixed across all subjects; subject-adaptive tuning may further improve change point localization. Finally, the targeted re-calibration strategy

still assumes the availability of an occasional cuff measurement at detected change points, and further work is needed to explore fully cuffless adaptation mechanisms.

References

- Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.
- Adam P Bress, Timothy S Anderson, John M Flack, Lama Ghazi, Michael E Hall, Cheryl L Laffer, Carolyn H Still, Sandra J Taler, Kori S Zachrisson, Tara I Chang, et al. The management of elevated blood pressure in the acute care setting: a scientific statement from the american heart association. *Hypertension*, 81(8):e94–e106, 2024.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.
- Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M Cooper-smith, Craig French, Flávia R Machado, Lauralyn McIntyre, Marlies Ostermann, Hallie C Prescott, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Critical care medicine*, 49(11):e1063–e1143, 2021.
- Shaghayegh Gharghabi, Chin-Chia Michael Yeh, Yifei Ding, Wei Ding, Paul Hibbing, Samuel LaMunion, Andrew Kaplan, Scott E Crouter, and Eamonn Keogh. Domain agnostic online semantic segmentation for multi-dimensional time series. *Data mining and knowledge discovery*, 33(1):96, 2018.
- Jingyuan Hong, Manasi Nandi, Weiwei Jin, and Jordi Alastruey. Using photoplethysmography to detect real-time blood pressure changes with a calibration-free deep learning model. *arXiv preprint arXiv:2407.03274*, 2024.
- Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.
- Yong-Yeon Jo, Byeong Tak Lee, Jeong-Ho Hong, Hak Seung Lee, Joon-myung Kwon, and Beom Joon Kim. Test-time calibration: A framework for personalized test-time adaptation in real-world biosignals. In *Conference on Health, Inference, and Learning*, pages 381–394. PMLR, 2025.
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016. doi: <https://doi.org/10.1038/sdata.2016.35>.

- Jingon Joung, Chul-Woo Jung, Hyung-Chul Lee, Moon-Jung Chae, Hae-Sung Kim, Jonghun Park, Won-Yong Shin, Changhyun Kim, Minhyung Lee, and Changwoo Choi. Continuous cuffless blood pressure monitoring using photoplethysmography-based ppg2bp-net for high intrasubject blood pressure variations. *Scientific reports*, 13(1):8605, 2023.
- Takashi Juri, Koichi Suehiro, Aya Kimura, Akira Mukai, Katsuaki Tanaka, Tokuhiko Yamada, Takashi Mori, and Kiyonobu Nishikawa. Impact of continuous non-invasive blood pressure monitoring on hemodynamic fluctuation during general anesthesia: a randomized controlled study. *Journal of Clinical Monitoring and Computing*, 32(6):1005–1013, 2018.
- Mohamad Kachuee, Mohammad Mahdi Kiani, Hoda Mohammadzade, and Mahdi Shabany. Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time. In *2015 IEEE international symposium on circuits and systems (ISCAS)*, pages 1006–1009. IEEE, 2015.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.
- Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.
- Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279, 2022.
- Chenbin Ma, Peng Zhang, Haonan Zhang, Zeyu Liu, Fan Song, Yufang He, and Guanglei Zhang. Stp: Self-supervised transfer learning based on transformer for noninvasive blood pressure estimation using photoplethysmography. *Expert Systems with Applications*, 249:123809, 2024.
- Hanieh Mohammadi, Bahram Tarvirdizadeh, Khalil Alipour, and Mohammad Ghamari. Cuff-less blood pressure monitoring via ppg signals using a hybrid cnn-bilstm deep learning model with attention mechanism. *Scientific Reports*, 15(1):22229, 2025.
- Lawrence Mosley. A balanced approach to the multi-class imbalance problem doctoral dissertation. *Iowa State University of Science and Technology, USA*, 6, 2013.
- Mohammad Moulaeifard, Peter H Charlton, and Nils Strodthoff. Generalizable deep learning for photoplethysmography-based blood pressure estimation—a benchmarking study. *arXiv preprint arXiv:2502.19167*, 2025.
- Ramakrishna Mukkamala, Mohammad Yavarimanesh, Keerthana Natarajan, Jin-Oh Hahn, Konstantinos G Kyriakoulis, Alberto P Avolio, and George S Stergiou. Evaluation of the accuracy of cuffless blood pressure measurement devices: challenges and proposals. *Hypertension*, 78(5):1161–1167, 2021.

- Ramakrishna Mukkamala, Sanjeev G Shroff, Konstantinos G Kyriakoulis, Alberto P Avolio, and George S Stergiou. Cuffless blood pressure measurement: Where do we actually stand? *Hypertension*, 82(6):957–970, 2025.
- Dean Nachman, Arik Eisenkraft, Eldad Rahamim, Mahsati Ibrahimli, Asen Asenov, Nir Goldstein, Yotam Kolben, Segev Huly, Arik Ben Ishay, Meir Fons, et al. Assessing cardiac flow measurements using a noninvasive photoplethysmography-based device compared to invasive pulmonary artery catheter. *JACC: Advances*, 4(9):102093, 2025.
- Guangkun Nie, Gongzheng Tang, Yujie Xiao, Jun Li, Shun Huang, Deyun Zhang, Qinghao Zhao, and Shenda Hong. Anyppg: An ecg-guided ppg foundation model trained on over 100,000 hours of recordings for holistic health profiling. *arXiv preprint arXiv:2511.01747*, 2025.
- Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=kYwTmlq6Vn>.
- Solmaz Rastegar, Hamid GholamHosseini, and Andrew Lowe. Non-invasive continuous blood pressure monitoring systems: current and proposed technology issues and challenges. *Physical and Engineering Sciences in Medicine*, 43(1):11–28, 2020.
- Mithun Saha, Maxwell A Xu, Wanting Mao, Sameer Neupane, James M Rehg, and Santosh Kumar. Pulse-ppg: An open-source field-trained ppg foundation model for wearable applications across lab and field settings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–35, 2025.
- Hamed Samimi and Hilmi R Dajani. A ppg-based calibration-free cuffless blood pressure estimation method using cardiovascular dynamics. *Sensors*, 23(8):4145, 2023.
- Zhan Shen, Jian Li, Hao Hu, Chentao Du, Xiaorong Ding, Tingrui Pan, and Xinge Yu. A review of non-invasive continuous blood pressure measurement: From flexible sensing to intelligent modeling. *AI Sensors*, 1(2):8, 2025.
- Gašper Slapničar, Nejc Mlakar, and Mitja Luštrek. Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network. *Sensors*, 19(15):3420, 2019.
- Gerrit JJ Van den Burg and Christopher KI Williams. An evaluation of change point detection algorithms. *arXiv preprint arXiv:2003.06222*, 2020.
- Anna Vybornova, Erietta Polychronopoulou, Arlene Wurzner-Ghajarzadeh, Sibylle Fallet, Josep Sola, and Gregoire Wuerzner. Blood pressure from the optical aktiia bracelet: a 1-month validation study using an extended iso81060-2 protocol adapted for a cuffless wrist device. *Blood pressure monitoring*, 26(4):305–311, 2021.
- Weinan Wang, Pedram Mohseni, Kevin L Kilgore, and Laleh Najafizadeh. Pulsedb: A large, cleaned dataset based on mimic-iii and vitaldb for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4:1090854, 2023.

Weinan Wang, Pedram Mohseni, Kevin L Kilgore, and Laleh Najafizadeh. Bp-net: Monitoring “changes” in blood pressure using ppg with self-contrastive masking. *IEEE Journal of Biomedical and Health Informatics*, 28(12):7103–7115, 2024.

Eugene Yang, Aletta E Schutte, George Stergiou, Fernando Stuardo Wyss, Yvonne Commodore-Mensah, Augustine Odili, Ian Kronish, Hae-Young Lee, and Daichi Shimbo. Cuffless blood pressure measurement devices—international perspectives on accuracy and clinical use: A narrative review. *JAMA cardiology*, 2025.

Appendix A. Further Implementation Details

A.1. Blood Pressure Change Point Detection with PELT

We use the Pruned Exact Linear Time (PELT) algorithm to detect ground-truth BP change points from reference SBP and DBP sequences. In practice, we use the PELT implementation provided by the `ruptures`¹ library. Table 6 summarizes the hyperparameter settings.

Table 6: Hyperparameters used for ground-truth BP change point detection using the PELT implementation in `ruptures`.

Parameter	Value
model	rbf
min_size	1
jump	1
pen	5

A.2. Calibration-Free PPG to BP Estimation and Online Change Point Detection

Table 7 summarizes the hyperparameters for PPG preprocessing, encoder training, and sequential clustering. The encoder training configuration is shared for both calibration-free PPG to BP estimation and supervised contrastive learning for CPD. For encoder backbones, we use Inception1D², PaPaGei-S³, Pulse-PPG⁴, and AnyPPG⁵, and adopt their original implementations for architecture-specific details.

A.3. Pseudo-Code for Online Change Point Detection via Sequential Clustering

Algorithm 1 summarizes the online change point detection procedure used in Section 4.2, where change points are triggered by persistent deviations from the current cluster centroid under a duration-dependent similarity threshold. Further details can be found at the official implementation: <https://github.com/bakqui/ppg-bpcp>.

Appendix B. Effect of Periodic Calibration Frequency

To analyze the effect of calibration frequency, we conduct an additional ablation study where the BP change point detector (PELT) is fixed, while the periodic calibration interval is varied among 120, 60, 30, and 10 minutes. Figure 4 reports the resulting MAE for both systolic BP (SBP) and diastolic BP (DBP).

1. <https://centre-borelli.github.io/ruptures-docs>
2. <https://github.com/AI4HealthUOL/ppg-ood-generalization>
3. <https://github.com/Nokia-Bell-Labs/papagei-foundation-model>
4. <https://github.com/maxxu05/pulseppg>
5. <https://github.com/Ngk03/AnyPPG>

Algorithm 1: Sequential clustering for online change point detection

Input:

Feature stream $z_{1:T}$
 Threshold params η_{base}, λ
 Persistence K

Output: List of change points $\mathcal{T}$ **Initialize:**

$\mathcal{T} \leftarrow [1], \mathcal{C} \leftarrow [z_1]$
 $c_{curr} \leftarrow z_1, n \leftarrow 1$
 $k \leftarrow 0, \mathcal{B} \leftarrow []$

for $t = 2$ **to** T **do**

$\eta_t \leftarrow \eta_{base} \cdot (1 - \exp(-\lambda n))$

if $Sim(z_t, c_{curr}) \geq \eta_t$ **then**

$k \leftarrow 0, \mathcal{B} \leftarrow []$
 $n \leftarrow n + 1$
 Update c_{curr} with z_t (Equation (1))

else

$k \leftarrow k + 1$

 Append (z_t, η_t) to $\mathcal{B}$

if $k = K$ **then**

 /* Change point detected. Update the current state. */

 Append t to $\mathcal{T}$

 /* Update centroids with stored features */

for $(z, \eta) \in \mathcal{B}$ **do**

$c^* \leftarrow \operatorname{argmax}_{c \in \mathcal{C}} Sim(z, c)$

if $Sim(z, c^*) \geq \eta$ **then**

 Update c_{curr} with c^* and z (Equation (2))

else

 Initialize new c_{curr} with z (Equation (3))

end

end

$k \leftarrow 0, \mathcal{B} \leftarrow []$

$n \leftarrow 0$

end

end

end

return $\mathcal{T}$

Table 7: Hyperparameters for calibration-free BP estimation and online change point detection

Parameter	Value
PPG Preprocessing	
Filter type	Third-order Butterworth
Filter bandwidth	0.5–8.0 Hz
Normalization	Z-normalization
PPG Encoder Training	
Epoch	50
Batch size	512
Optimizer	AdamW
Learning rate	0.001
Weight decay	0.001
Sequential Clustering	
η_{base}	0.99
λ	0.05
K	5

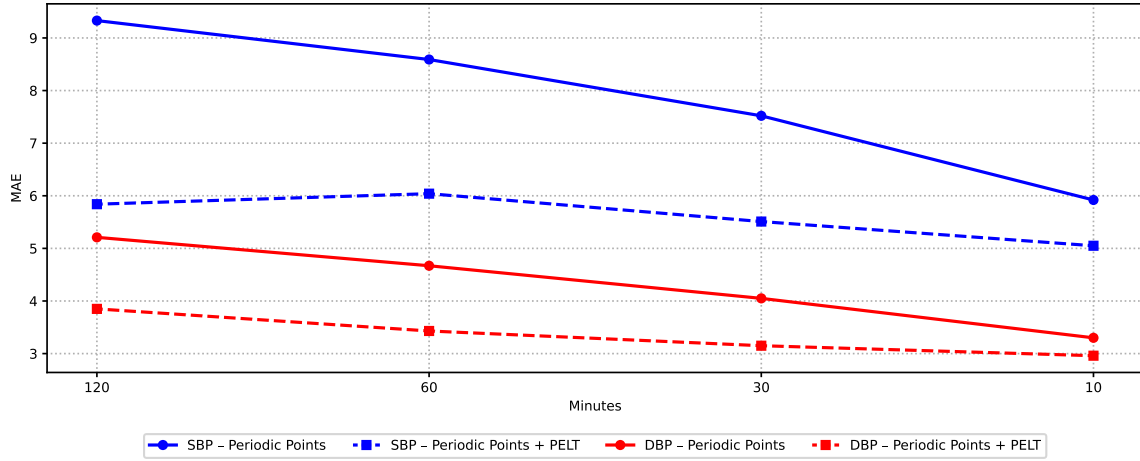


Figure 4: Effect of periodic calibration interval on BP estimation error, with and without PELT-based change point triggering.

As the calibration interval becomes shorter, periodic calibration alone consistently improves estimation accuracy, indicating that more frequent re-calibration helps mitigate performance degradation caused by physiological drift. However, incorporating PELT-based

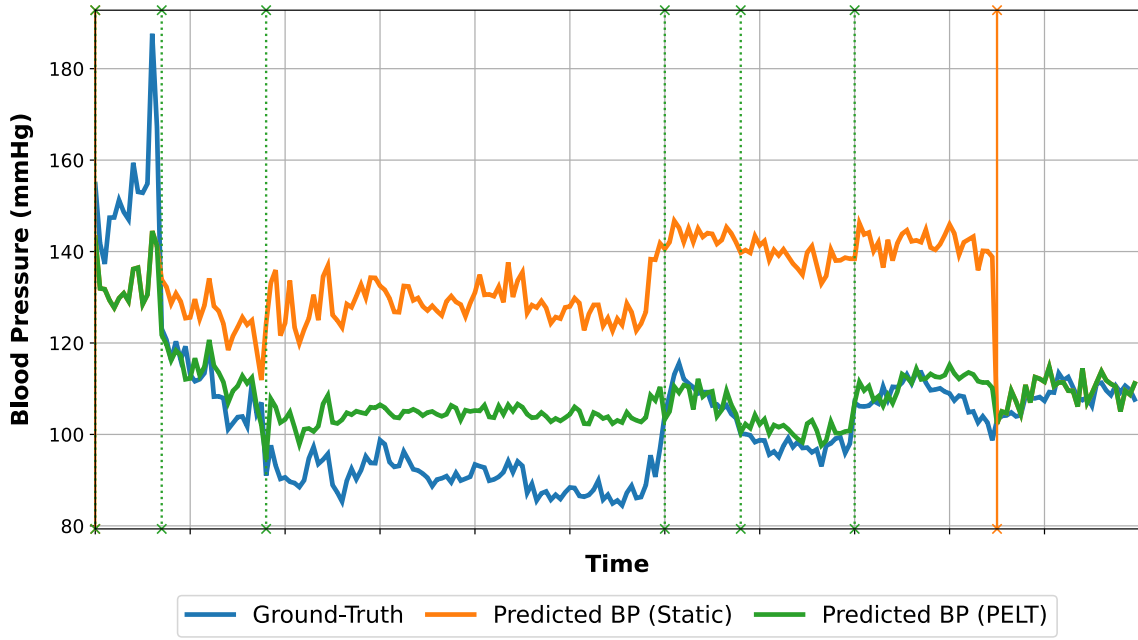


Figure 5: Qualitative results for calibration-based methods using periodic re-calibration (orange) and targeted re-calibration at BP fluctuation points (green).

change point triggering further improves performance across all calibration intervals. Notably, the relative gain from PELT remains evident even under short periodic intervals (e.g., 30 and 10 minutes), suggesting that re-calibration timing aligned with BP fluctuation events provides complementary benefits beyond increasing calibration frequency.

These results indicate that calibration effectiveness depends not only on how often re-calibration is performed, but also on when it is applied, reinforcing the importance of fluctuation-aware re-calibration strategies.

Appendix C. Qualitative Results

Figures 5–7 show representative examples comparing the periodic calibration (Static) baseline and our PELT-based re-calibration against ground-truth BP. In all cases, the periodic calibration exhibits persistent bias when patient hemodynamics shift over time, failing to adapt to sustained BP changes. The PELT model detects these transitions via change point detection and re-calibrates accordingly, maintaining closer agreement with the ground truth throughout the session.

Appendix D. Sensitivity Analysis of PELT Hyperparameters

We selected the PELT hyperparameters via greedy search, using the average blood pressure change within detected segments, $\text{Mean}|\Delta_{\text{seg}}|$, as the selection criterion. This metric favors

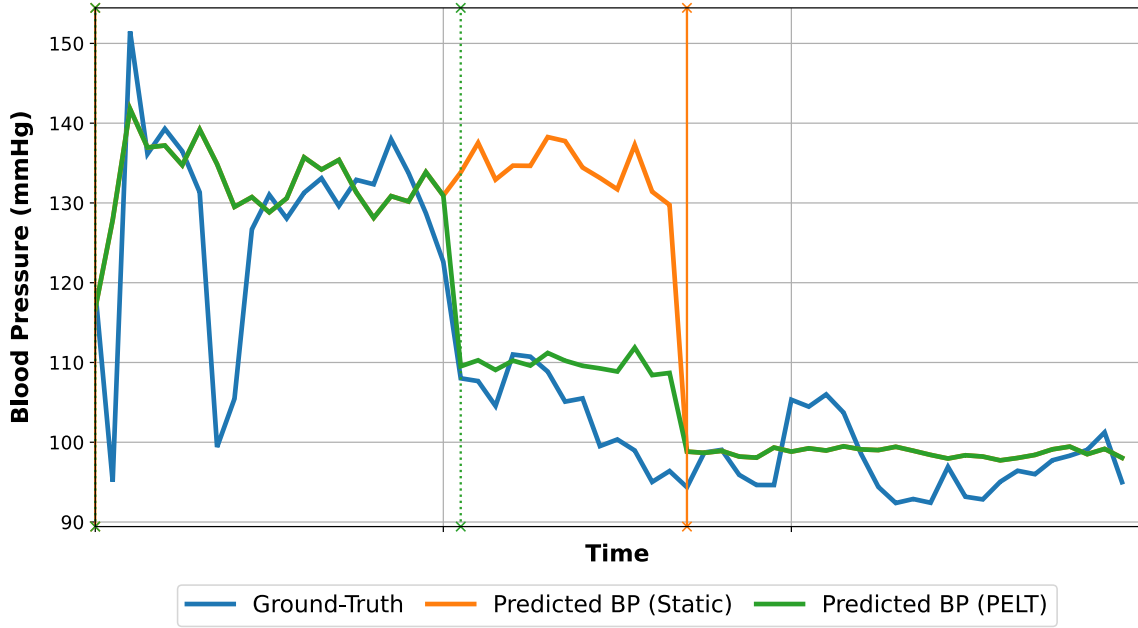


Figure 6: Qualitative results for calibration-based methods using periodic re-calibration (orange) and targeted re-calibration at BP fluctuation points (green).

Table 8: Sensitivity analysis of PELT hyperparameters. The selected default values are marked with an asterisk.

Parameter	Value	Points/Case	$ \Delta_{\text{Seg}} $ SBP	$ \Delta_{\text{Seg}} $ DBP
Penalty	1	26.75	11.69	6.16
	5*	9.09	13.18	6.81
	10	6.00	12.04	6.27
Model	11	56.72	7.52	3.94
	12	121.70	6.32	3.33
	rbf*	9.09	13.18	6.81
	linear	50.35	6.92	3.76

segmentations that capture meaningful BP fluctuations rather than noise-induced over-segmentation.

The selected default configuration provided the best overall trade-off. The cost model had the largest effect on segmentation behavior: 11 and 12 produced many more detected points per case but much smaller segment-level BP changes, indicating sensitivity to noise rather than meaningful BP transitions.

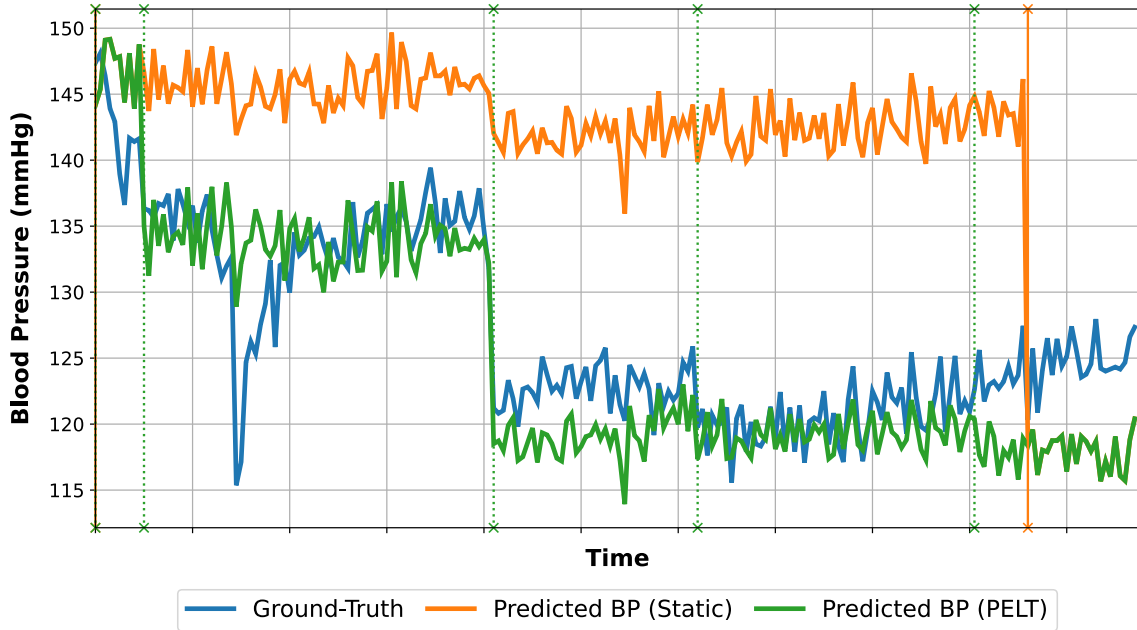


Figure 7: Qualitative results for calibration-based methods using periodic re-calibration (orange) and targeted re-calibration at BP fluctuation points (green).

Appendix E. Hyperparameter Sensitivity of the Online CPD Module

For the online CPD module, hyperparameters were selected based on the intersection-over-union (IoU), which directly reflects the overlap between predicted unstable intervals and the reference intervals. Table 9 summarizes the sensitivity analysis for the main hyperparameters. Overall, the results show a clear trade-off between classification accuracy and interval-level localization quality. For η_{base} , decreasing the value increased accuracy but reduced IoU, indicating that a more permissive threshold improved point-wise detection at the expense of precise interval matching. For λ , larger values caused the threshold to tighten too rapidly, which substantially degraded both accuracy and IoU. Based on these results, we selected $\eta_{\text{base}} = 0.99$ and $\lambda = 0.05$ as the default configuration, as they achieved the best IoU while maintaining stable AUROC and accuracy.

Appendix F. Further Evaluations

F.1. BP Estimation Performance on Unstable BP Intervals with Targeted Re-Calibration

Section 6.2 showed that calibration-free models degrade substantially within unstable BP intervals. Here we examine whether the benefit of targeted re-calibration is retained under the same conditions. Table 10 extends Table 3 by reporting **PPG2BP-Net** performance separately over the whole recording and over unstable BP intervals, defined as in Section 3.2. Every calibration strategy exhibits higher SBP error within unstable intervals than over the

Table 9: Sensitivity analysis of online CPD hyperparameters. The selected default values are marked with an asterisk.

Parameter	Value	AUROC	Accuracy	IoU
η_{base}	0.999	0.7271	0.6412	0.6140
	0.99*	0.7275	0.6457	0.6149
	0.9	0.7190	0.7214	0.5656
λ	1.0	0.7444	0.2391	0.2793
	0.1	0.7401	0.5271	0.5758
	0.05*	0.7275	0.6457	0.6149

Table 10: Impact of targeted re-calibration using ground-truth (offline) blood pressure change points, evaluated with **PPG2BP-Net**. Mean absolute error in mmHg with 95% confidence intervals are reported over the whole recording (*Whole*) and over unstable BP intervals defined in Section 3.2. Points / Case denotes the average number of re-calibration points per case (mean $\pm$ std). The plus sign (“+”) keeps the 120-minute schedule as the base and adds the change points of strategies such as ΔBP , *Random*, and *PELT*. *Random* uses the per-case point count of PELT with uniformly sampled locations.

Calibration Points	Whole		Unstable BP Intervals		Points / Case
	SBP $\downarrow$	DBP $\downarrow$	SBP $\downarrow$	DBP $\downarrow$	
Every 120 minutes	9.33 $\pm$ 0.05	5.21 $\pm$ 0.03	10.93 $\pm$ 0.17	5.78 $\pm$ 0.09	3.30 $\pm$ 1.52
+ ΔBP	8.05 $\pm$ 0.05	4.53 $\pm$ 0.03	9.46 $\pm$ 0.15	3.88 $\pm$ 0.08	17.36 $\pm$ 22.57
+ <i>Random</i>	7.44 $\pm$ 0.04	4.01 $\pm$ 0.03	9.10 $\pm$ 0.15	4.63 $\pm$ 0.09	9.91 $\pm$ 4.02
+ <i>PELT</i>	5.84 $\pm$ 0.04	3.85 $\pm$ 0.02	6.73 $\pm$ 0.13	4.11 $\pm$ 0.07	9.82 $\pm$ 4.02

whole recording, confirming that these segments remain the most difficult regions to track. The relative ordering of the strategies is nevertheless preserved: periodic calibration alone yields 10.93 mmHg SBP MAE, the ΔBP and *Random* strategies reduce the error to 9.46 and 9.10 mmHg, respectively, and *PELT* achieves 6.73 mmHg. Notably, *PELT* also shows the smallest gap between whole-interval and unstable-interval performance (0.89 mmHg, compared with 1.41–1.66 mmHg for the remaining strategies), indicating that aligning re-calibration timing with distributional shifts is most effective precisely where estimation is hardest. For DBP, the differences across strategies are smaller, and the ΔBP criterion attains the lowest error within unstable intervals (3.88 mmHg) despite requiring nearly twice as many calibration points per case; this is consistent with our definition of unstable intervals, which is driven by systolic and mean arterial pressure levels and therefore couples less directly to diastolic dynamics. Overall, these results indicate that the advantage of targeted re-calibration reported in Section 6.2 is not an artifact of averaging over stable periods, but is retained—and for SBP amplified—under clinically critical hemodynamic conditions.

Table 11: Comparison with a naive binary classification baseline for online change point detection. Detection performance and the corresponding downstream BP estimation error using **PPG2BP-Net** are reported.

CPD Backbone	CPD Method	CPD		BP Estimation	
		AUROC $\uparrow$	IoU $\uparrow$	SBP MAE $\downarrow$	DBP MAE $\downarrow$
Inception1D	Classification	0.5332	0.3919	8.31 ± 0.04	4.56 ± 0.03
	Ours	0.7275	0.6149	7.27 ± 0.04	3.87 ± 0.02
PaPaGei-S	Classification	0.5321	0.4552	7.79 ± 0.04	4.13 ± 0.02
	Ours	0.7223	0.6062	7.30 ± 0.04	3.87 ± 0.02

F.2. Classification Baseline for CPD

As no prior work formulates online BP change point detection from PPG signals alone, a directly comparable baseline is not available for the results reported in Table 4. To place these numbers in context, we additionally train a naive binary classifier that predicts, for each incoming 10-second PPG segment, whether the segment corresponds to a BP change point, supervised by the same PELT-derived change point labels. The classifier follows the encoder configuration and training protocol described in Appendix A.3, and replaces the supervised contrastive objective in Section 4.1 and the sequential clustering procedure in Section 4.2 with a binary cross-entropy loss. The predicted change points are then used to trigger targeted re-calibration with **PPG2BP-Net** under the same protocol as Table 5. Table 11 reports the comparison for two backbones, Inception1D and PaPaGei-S. For both backbones, the classification baseline remains close to chance level in AUROC (0.5332 and 0.5321), whereas the proposed method achieves 0.7275 and 0.7223, with the same ordering observed in IoU. The detection gap is likewise reflected in the downstream task: SBP MAE decreases from 8.31 to 7.27 mmHg with Inception1D and from 7.79 to 7.30 mmHg with PaPaGei-S, while DBP MAE decreases from 4.56 and 4.13 mmHg to 3.87 mmHg in both cases. Detection quality thus tracks the gain obtained from targeted re-calibration, and the consistency of this pattern across a from-scratch and a pretrained backbone indicates that the improvement originates from the detection formulation rather than from the choice of encoder.

F.3. Online CPD-Triggered Re-Calibration on Unstable BP Intervals

Table 12 extends Table 5 by reporting **PPG2BP-Net** performance separately over the whole recording and over the unstable BP intervals defined in Section 3.2, and additionally includes the classification baseline described in Appendix F.2 as a trigger. The ordering observed over the whole recording is preserved within unstable intervals: periodic calibration alone yields 10.93 mmHg SBP MAE, the *Classification* and *Random* triggers reduce the error to 8.73 and 8.06 mmHg, and *Online CPD* achieves the lowest error at 7.76 mmHg, with the same ordering in DBP. Two comparisons isolate the source of this gain. First, *Online CPD* and *Random* issue an almost identical number of re-calibration points per case (8.77 for both), so the remaining difference is attributable to when re-calibration is applied rather

Table 12: Performance of targeted re-calibration using the proposed online change point detection. Mean absolute error in mmHg with 95% confidence intervals are reported. Points / Case denotes the average number of re-calibration points per case (mean $\pm$ std). The plus sign (“+”) keeps the 120-minute schedule as the base and adds the change points of strategies such as *Random*, *Classification*, and *Online CPD*. *Random* uses the per-case point count of *Online CPD* with uniformly sampled locations. Both *Classification* and *Online CPD* use Inception1D as the CPD backbone.

Calibration Points	Whole		Unstable BP Intervals		Points / Case
	SBP $\downarrow$	DBP $\downarrow$	SBP $\downarrow$	DBP $\downarrow$	
Every 120 minutes	9.33 $\pm$ 0.05	5.21 $\pm$ 0.03	10.93 $\pm$ 0.17	5.78 $\pm$ 0.09	3.30 $\pm$ 1.52
+ <i>Classification</i>	8.31 $\pm$ 0.04	4.56 $\pm$ 0.03	8.73 $\pm$ 0.09	4.72 $\pm$ 0.05	12.31 $\pm$ 22.59
+ <i>Random</i>	7.60 $\pm$ 0.02	4.07 $\pm$ 0.01	8.06 $\pm$ 0.09	4.33 $\pm$ 0.05	8.77 $\pm$ 3.34
+ <i>Online CPD</i>	7.27 $\pm$ 0.02	3.87 $\pm$ 0.01	7.76 $\pm$ 0.08	4.15 $\pm$ 0.05	8.77 $\pm$ 3.33

Table 13: Analysis on computation cost and latency of online change point detection

CPD Backbone	Params (M)	FLOPs (M)	Forward (ms)	Cluster (ms)
Inception1D	0.5	591.8	21.9	0.03
PaPaGei-S	5.8	76.6	9.2	0.03
Pulse-PPG	28.5	1140.5	31	0.03
AnyPPG	5.9	245.5	26.2	0.04

than how often. Second, the *Classification* trigger performs worse than both despite issuing substantially more points (12.31 per case), indicating that inaccurately localized change points do not compensate for their higher frequency. All triggered strategies also exhibit a smaller degradation from whole-interval to unstable-interval performance (0.42–0.49 mmHg SBP) than periodic calibration alone (1.60 mmHg), confirming that the benefit of targeted re-calibration is retained under clinically critical hemodynamic conditions when change points are detected online.

F.4. Computational Cost

Table 13 reports the computational cost of the online CPD module. At inference, the pipeline performs a single encoder forward pass per incoming 10-second PPG segment, followed by cosine similarity comparisons against the stored centroids, which require $O(K|\mathcal{C}|d)$ operations, where d is the feature dimension, K the buffer size, and $|\mathcal{C}|$ the number of centroids. Across all backbones, the forward pass takes 9.2–31.0 ms while the sequential clustering step remains below 0.05 ms. Since a prediction is required only once every 10 seconds, the end-to-end latency leaves a large margin for real-time operation, and the overall cost is dominated by the encoder, with the proposed detection procedure adding negligible overhead on top of the backbone already used for BP estimation.

Table 14: Distribution of unstable BP segments in VitalDB with respect to vasopressor administration.

Context	Unstable Segments (%)
Vasopressor	6,094 (85.4%)
No vasopressor	1,038 (14.6%)

Appendix G. Clinical Interpretation of Detected Change Points

What transitions PELT captures. To characterize the physiological events underlying the detected change points, we cross-reference the unstable BP intervals defined in Section 3.2 with vasopressor administration records available in VitalDB. As shown in Table 14, the large majority of unstable segments (6,094 segments, 85.4%) occur while a vasopressor is being administered, compared with 1,038 segments (14.6%) in its absence. This indicates that the transitions identified by PELT are not arbitrary statistical fluctuations but concentrate around pharmacologically driven hemodynamic shifts, which are precisely the events during which a stale calibration reference is most likely to mislead the estimation model. This also supports the use of change point-triggered re-calibration as a clinically meaningful, rather than purely data-driven, criterion.

Clinical consequences in unstable intervals. The practical relevance of the residual error in these intervals follows from where the intervals are located. Under our definition, unstable segments lie near the $SBP > 180$ mmHg and $MAP < 65$ mmHg boundaries, and the SBP error observed within them remains approximately 11–14 mmHg for calibration-based models (Table 3). An error of this magnitude at these boundaries may delay the recognition of a hypertensive emergency, for which timely treatment is recommended in the acute care setting (Bress et al., 2024), or the initiation of vasopressor and fluid therapy in the management of sepsis and septic shock (Evans et al., 2021). This motivates reporting estimation accuracy within fluctuation intervals separately, rather than relying on whole-recording averages that mask errors in exactly these regions.