Transformer-Based Cuffless Blood Pressure Estimation Driven by Speckle Contrast Optical Spectroscopy

Tianrui Qi
Department of Biomedical Engineering, Boston University

February 19, 2026

Introduction

Hypertension remains a major cardiovascular risk factor, and blood pressure (BP) assessed outside the clinic is often more predictive than isolated office measurements for long-term outcomes[1]. This is partly because BP is highly dynamic over daily activities and recovery states, so sparse in-clinic measurements can miss clinically meaningful variability. At the same time, cuff-based sphygmomanometry is intermittent and disruptive, which continues to motivate cuffless continuous monitoring approaches[2]. Many existing cuffless pipelines rely primarily on PPG, but robust ambulatory performance across people, activities, and sensing setups is still challenging[3–5].

SCOS provides a different measurement axis by jointly capturing blood-volume and blood-flow dynamics from one optical acquisition. The method builds on diffuse/speckle optical flow sensing[6–8] and has progressed to high-speed waveform-resolved measurements that recover pulsatile blood flow index (BFi) alongside intensity-derived PPG[9]. This dual-modality view is physiologically relevant for BP because pressure regulation reflects coupled effects of vascular resistance, compliance, and flow-volume dynamics rather than a single waveform descriptor[10]. Prior SCOS/SPG studies further suggest that BFi morphology carries information complementary to PPG and can remain informative under lower-perfusion conditions[11;12].

Our previous work established the high-speed SCOS hardware and data-acquisition pipeline and showed that combining BFi and PPG feature sets improves BP estimation over PPG-only features[13]. Building on that foundation, this paper focuses on the computational question: whether representation learning can preserve the latent physiological structure of SCOS waveforms while improving BP prediction across repeated measurements.

Transformer-based sequence modeling is a natural candidate because self-attention can represent long-range temporal dependencies and cross-token interactions[14]. In medical waveform analysis, transformer variants have already shown strong performance in ECG tasks such as arrhythmia modeling[15–17]. Transformer-based designs have also been explored for PPG-driven hemodynamic targets, including BP-related prediction settings[18], and cross-dataset biosignal pretraining frameworks such as BIOT[19]. These developments motivate a transformer formulation for SCOS, but direct transfer is nontrivial because our setting couples optical modality-specific physiology with continuous BP-waveform supervision and downstream calibration.

We therefore present a three-stage SCOS framework with (1) self-supervised optical representation pretraining, (2) BP-supervised waveform finetuning, and (3) lightweight measurement-specific FiLM calibration. We evaluate using measurement-independent train/test splitting, a standard safeguard against leakage in medical time-series evaluation[20], and analyze both endpoint error and latent-space structure to characterize stage-wise behavior.

Method

Data

The data used in this study are paired optical and BP waveforms acquired with the SCOS hardware and processing workflow reported by Garrett *et al.*[13]. SCOS provides two complementary optical pulse waveforms: BFi, a relative index of microvascular flow derived from dynamic speckle-contrast fluctuations, and PPG, which reflects pulsatile blood-volume oscillations. The acquisition geometry includes 808 nm transmission at the finger and 532 nm/808 nm reflectance at the wrist, yielding six possible optical channel combinations. In the prior study, wrist 532 nm BFi and wrist 808 nm PPG were excluded owing to inadequate signal quality. In this study, we further exclude wrist 532 nm PPG because it contributes minimally to BP prediction, and retain three optical inputs: finger 808 nm BFi, finger 808 nm PPG, and wrist 808 nm BFi. The left panel of Figure 1 depicts the acquisition geometry and three retained optical channels. Continuous BP is recorded simultaneously from the contralateral arm with Finapres Nova. BP and optical streams are time-aligned using the camera exposure-active transistor-transistor logic (TTL) signal.

The cohort contains 46 subjects with 82 measurements: 40 non-hypertensive subjects (IDs starting with N/S) and 6 hypertensive subjects (IDs starting with H). All subjects contribute a first measurement under the leg-press perturbation protocol, including a 1-minute baseline (condition 1), a 4-minute isometric leg press (condition 2), and a 1-minute recovery measurement acquired after a 15-minute rest interval (condition 3). A repeat measurement is available for 36 subjects, including all N/H subjects and 20 of 30 S subjects. For S subjects, the second measurement follows the same leg-press protocol. For N/H subjects, the second measurement is an arm-perturbation acquisition. Across protocols, these perturbations are designed to induce within-measurement BP variation.

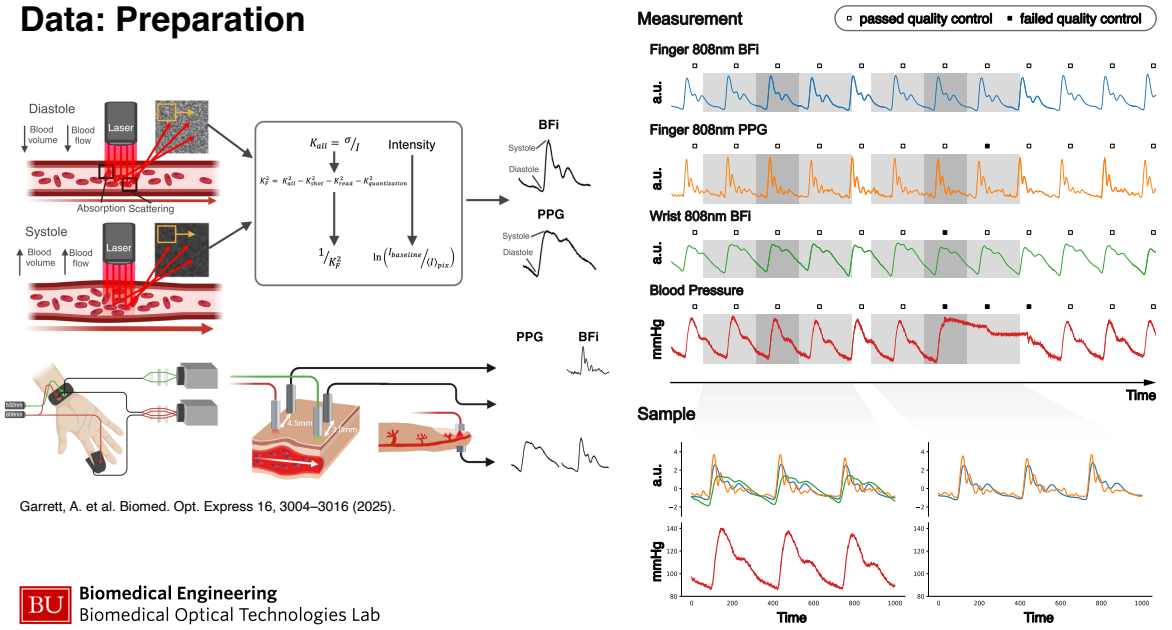

Figure 1: Left Acquisition geometry and three retained optical channels: finger 808 nm blood flow index (BFi), finger 808 nm photoplethysmography (PPG), and wrist 808 nm BFi. **Right** Heartbeat-centered sample construction using length-1000 windows with center-beat quality control (QC) gating. Examples show a complete four-channel sample and a sample in which the blood pressure (BP) channel fails QC, leaving three valid optical channels.

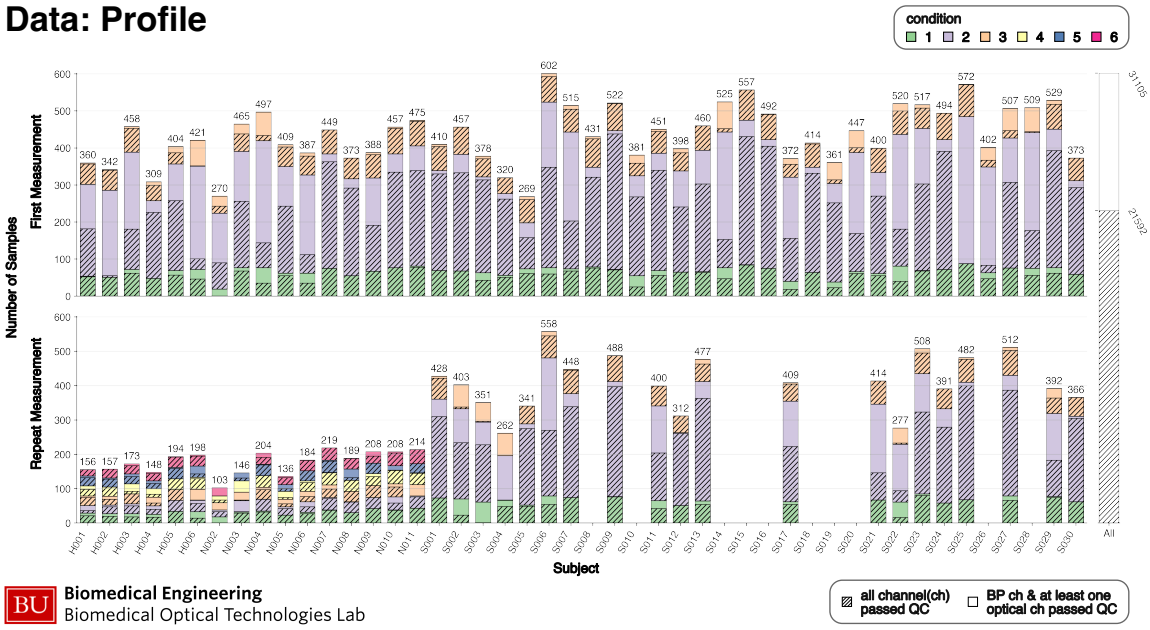

Figure 2: Stacked bars show per-measurement sample counts colored by condition. Measurements with three conditions correspond to the leg-press protocol, whereas measurements with six conditions correspond to the arm-position perturbation protocol. Solid fills denote samples with a valid blood pressure (BP) channel and at least one valid optical channel, $n = 31{,}105$. Hatched overlays denote the subset with all four channels passing quality control (QC), $n = 21{,}592$.

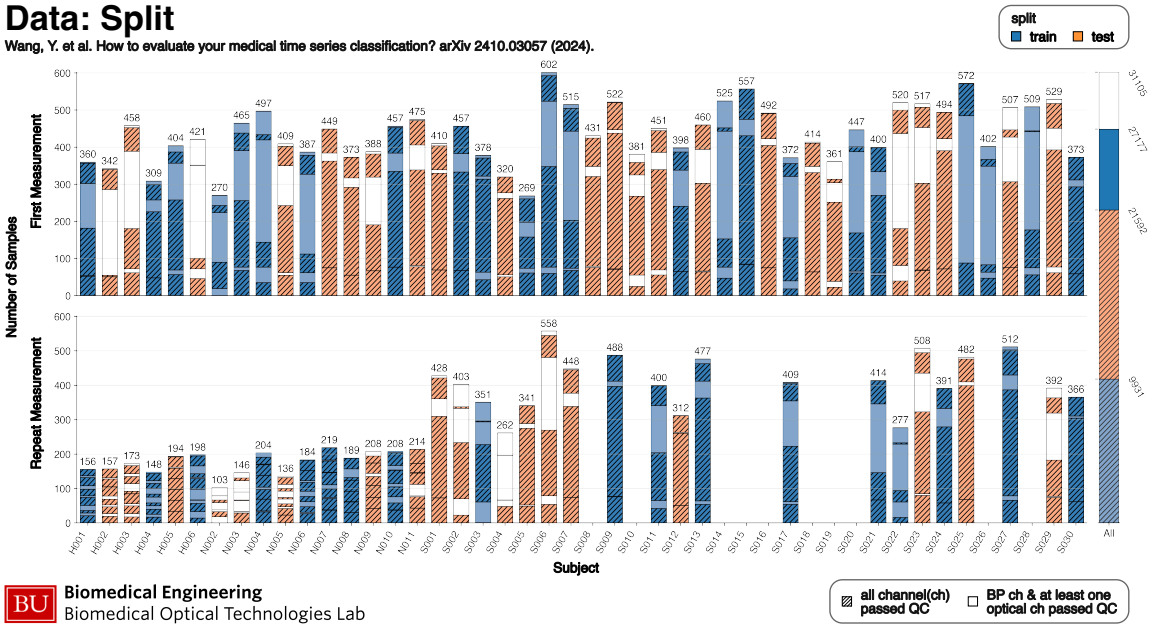

Figure 3: Stacked bars show per-measurement sample counts colored by measurement-level split assignment. Fill patterns indicate sample completeness: valid BP and at least one valid optical channel (solid fill) versus all four channels valid (hatched overlay). The training split includes 9931 complete samples and 5585 partially observed samples; testing is restricted to complete samples ($n = 11{,}661$).

For each measurement, we construct heartbeat-centered samples from four aligned channels (three optical plus BP). Each channel has beat-level quality control (QC) labels. For every heartbeat, we extract a length-1000 sample with that heartbeat at the window center. Channel validity for each sample is determined by center-beat QC: if the center beat passes QC, that channel is kept for the sample even when neighboring beats in the same window fail QC; if the center beat fails QC, that channel is treated as missing. The right panel of Figure 1 shows two examples: a complete four-channel sample (left) and a sample in which the BP channel fails QC, leaving three valid optical channels (right). This preparation yields 31,105 samples with a valid BP channel and at least one valid optical channel; among them, 21,592 have all four channels valid. Figure 2 shows per-measurement sample counts colored by condition, with sample completeness denoted by fill patterns.

To avoid leakage across closely related samples[20], we use a measurement-independent split in which all samples from a given measurement are assigned to either the training set or the test set. Within the training split, we retain both complete samples, i.e., all four channels passing QC, and partially observed samples, i.e., valid BP and at least one valid optical channel. For evaluation, we restrict the test set to complete samples only. Among the 21,592 complete samples, 9931 are assigned to training and 11,661 to testing. Of the remaining 9513 partially observed samples, 5585 fall into the training split, while the rest are excluded. Figure 3 summarizes the resulting train/test split across measurements and sample completeness, denoted by color and fill pattern.

Model

Inspired by BIOT[19], we use a transformer[14] encoder to learn representations from multi-channel optical waveforms. A transformer encoder is an attention-based sequence model that allows each local waveform segment to integrate information from other segments, capturing longer-range temporal context and interactions across channels. Rather than relying on hand-crafted features, we learn a shared representation space in which each sample or each channel is summarized by a fixed-length vector. These representations can be used to visualize latent structure and to support downstream tasks. At a high level, the model comprises four stages: tokenization, embedding, representation encoding, and mean pooling (Figure 4).

Each input sample is a length-1000 window centered on a heartbeat, with a variable subset of the three optical channels available after QC. We tokenize each available channel independently using a length-100 window and a stride size of 25, yielding 37 tokens per channel if the entire channel is valid. Then, each waveform token is mapped to a 256-dimensional embedding by a shared linear projection. To retain channel identity and temporal ordering, we add two auxiliary embeddings: (i) a learnable channel embedding that encodes measurement site and modality, and (ii) a sinusoidal positional embedding that indexes the token location within each channel. The sum of these three components forms the token embedding used by the encoder.

Token embeddings from all available channels are processed jointly by the transformer encoder to produce contextual token representations. The encoder alternates multi-head self-attention and position-wise feed-forward layers, with residual connections and layer normalization. Self-attention updates each token representation as a content-adaptive mixture of other tokens, enabling the model to capture within-channel temporal structure and cross-channel coupling without hand-crafted features. Multiple attention heads allow the encoder to model different relationships in parallel. We use 4 stacked transformer encoder blocks; each block contains multi-head self-attention with 8 heads and a position-wise feed-forward network with hidden dimension 1024. Intermediate representations remain 256-dimensional throughout. Channel and sample representations are obtained by masked mean pooling over token representations within each channel, or across all channels' token.

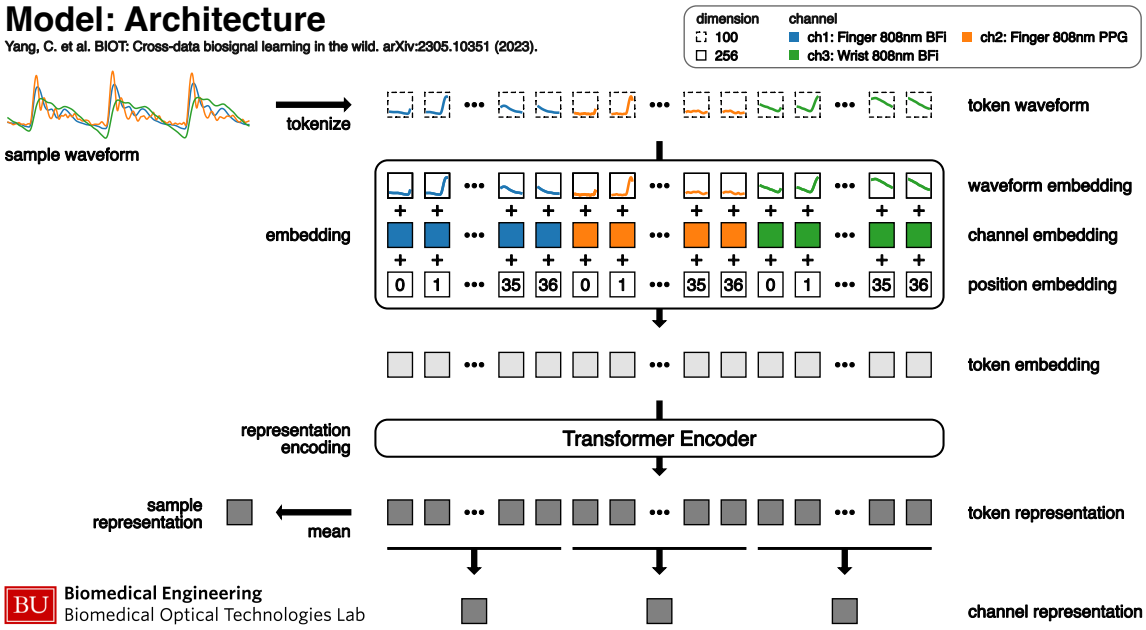

Figure 4: Eeach length-1000 sample waveform from each available optical channel is segmented into overlapping length-100 token waveform with a 75 overlap, yielding a maximum of 37 tokens per channel. Each token is linearly projected to a 256-dimensional waveform embedding and summed with a learnable channel embedding and a sinusoidal positional embedding to form the token embedding. Token embeddings from all channels are processed jointly by a transformer encoder to produce contextual token representations. Masked mean pooling over token representations within each channel and across all channels' token produces channel- and sample-level representations used for downstream tasks.

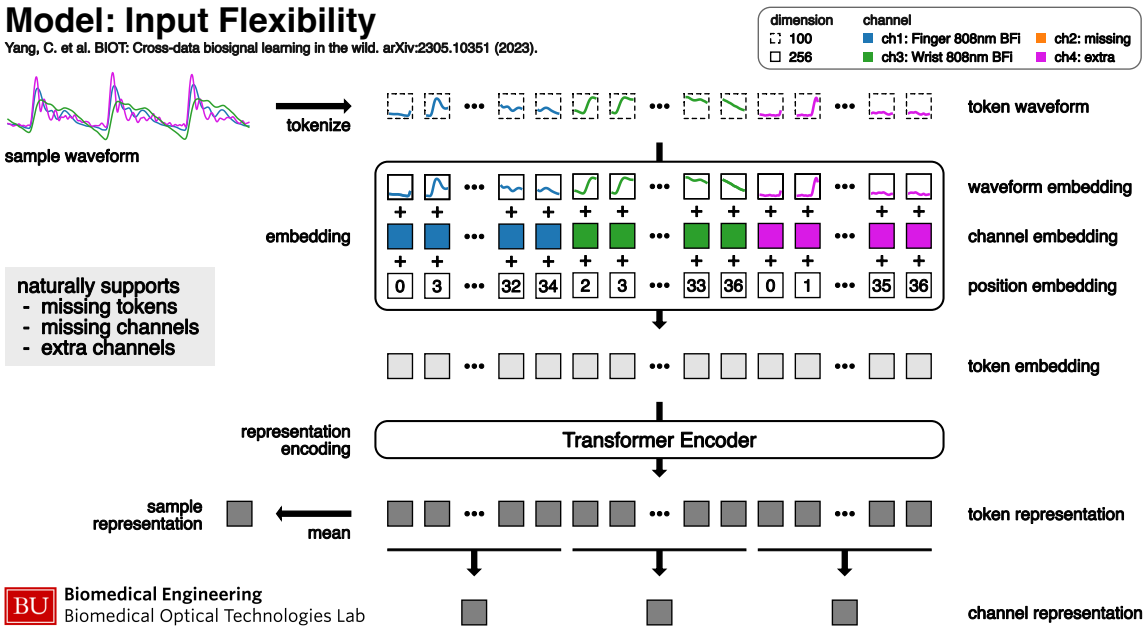

Figure 5: Example input sample waveforms with missing tokens (ch 1 and 3), a missing channel (ch 2), and an additional channel (ch 4), demonstrating the flexible input handling of the model. Tokens with missing values are excluded while channel and positional embeddings preserve token provenance and temporal order.

This architecture is naturally compatible with incomplete and heterogeneous inputs (Figure 5). Tokens that belong to missing channels or contain missing values are excluded from self-attention and pooling, avoiding explicit imputation while preserving relative temporal structure through positional embeddings. This design lets us retain and learn from partially observed samples rather than discarding them, increasing the effective amount of data available for representation learning. Because channel identity is provided by channel embeddings, the same interface can accommodate extended channel sets, e.g., new sensor sites or modalities, by appending tokens from additional waveforms and assigning them new channel identifiers, without changing the transformer backbone.

Training

We train the model in three stages to progressively increase task specificity (Figure 6). Stage 1 performs self-supervised masked reconstruction pretraining on optical waveforms. Stage 2 incorporates BP supervision by predicting BP waveforms from the optical inputs while retaining a reconstruction regularizer. Stage 3 performs lightweight measurement-specific finetuning to calibrate the pretrained model.

Stage 1. We pretrain the encoder using a self-supervised masked reconstruction objective[21] on optical waveforms that assigned to the training split. Since Stage 1 only uses optical inputs and the objective is self-supervised, we further include samples without a valid BP waveform as long as at least one optical channel passes QC, adding ~ 1000 additional samples for pretraining. For simplicity, this case is not shown in Figure 3. Three data augmentations are applied to improve generalization and robustness:

- random permutation of channel order, which encourages model to rely on channel embeddings rather than a fixed input ordering to identify each modality and measurement site;
- stochastic channel dropping, which simulates missing-channel scenarios; and
- global random temporal shift of up to 25 time points applied to all channels with NaN padding, which promotes invariance to small timing offsets in heartbeat-centered windowing without inducing inter-channel misalignment.

We then mask a subset of tokens prior to transformer encoding using a mixture of point-wise masking (mask ratio sampled uniformly up to 0.2 per sample) and a contiguous span mask applied to one randomly selected channel. Span lengths are sampled to cover 0 to 20% of token positions when only one channel is available and 20 to 40% otherwise. Masked tokens are replaced by a learnable mask embedding with 10% of selected tokens left unchanged, and the transformer encodes the corrupted sequence to produce contextual token representations. A lightweight token-wise reconstruction head predicts the original waveform segment for each masked token, and we minimize a smooth L1 loss computed only on the masked tokens. This masked reconstruction objective trains the encoder to infer missing waveform segments from surrounding context and cross-channel correlations, learning a general-purpose representation of the optical waveforms without any label supervision.

Stage 2. We then incorporate BP supervision by initializing from the Stage 1 checkpoint and training on only the final transformer encoder block, while keeping the embedding layers and the first three blocks frozen. We train on samples assigned to the training split as shown in Figure 3, restricted to samples with a valid BP waveform and at least one valid optical channel. Given optical inputs, the transformer produces contextual token representations, which are pooled across channels to form a single token sequence that aggregates information from all available channels. A regression head then predicts the continuous BP waveform from these pooled token representations. We compute the loss as the mean squared error (MSE) between the predicted and measured BP

waveforms. Compared with Stage 1, we use a lower probability of stochastic channel dropping (0.2) and disable the temporal-shift augmentation. We also broaden the span-masking schedule for the reconstruction regularizer to sample span lengths up to 50% of token positions for single-channel inputs and up to 100% for multi-channel inputs. Optimization is dominated by the BP regression objective, with the masked reconstruction loss retained as a downweighted regularizer.

Stage 3. Finally, we perform measurement-specific calibration using a feature-wise linear modulation (FiLM) adapter[22]. Starting from the Stage 2 model, we freeze the embedding layers, the full transformer encoder, and the BP regression head, and insert a lightweight FiLM adapter between the pooled token representations and the regression head. The adapter applies a feature-wise affine transformation $z' = (1 + \gamma) \odot z + \beta$, where γ and β are learnable 256-dimensional vectors shared across tokens. Calibration is performed independently for each measurement: we use samples from the baseline segment, i.e., condition 1, to fit the adapter parameters, and then use the calibrated model to predict BP waveforms for the remaining samples within that measurement as test. Because only the adapter parameters are optimized, calibration is computationally lightweight and helps correct measurement-specific biases without overfitting the backbone. We optimize the adapter using a robust waveform loss based on the smooth L1 distance, including a shift-tolerant shape term (allowing temporal shifts up to 50 time points) and penalties on waveform extrema. Token masking and data augmentations are disabled during this stage.

All experiments were implemented in Python (v3.13.11) using PyTorch[23] (v2.9.1) and trained with PyTorch Lightning[24] (v2.5.6), with experiment configuration managed by Hydra[25] (v1.3.2). Across stages, we optimize with Adam[26] and a step-decay learning-rate schedule (StepLR).

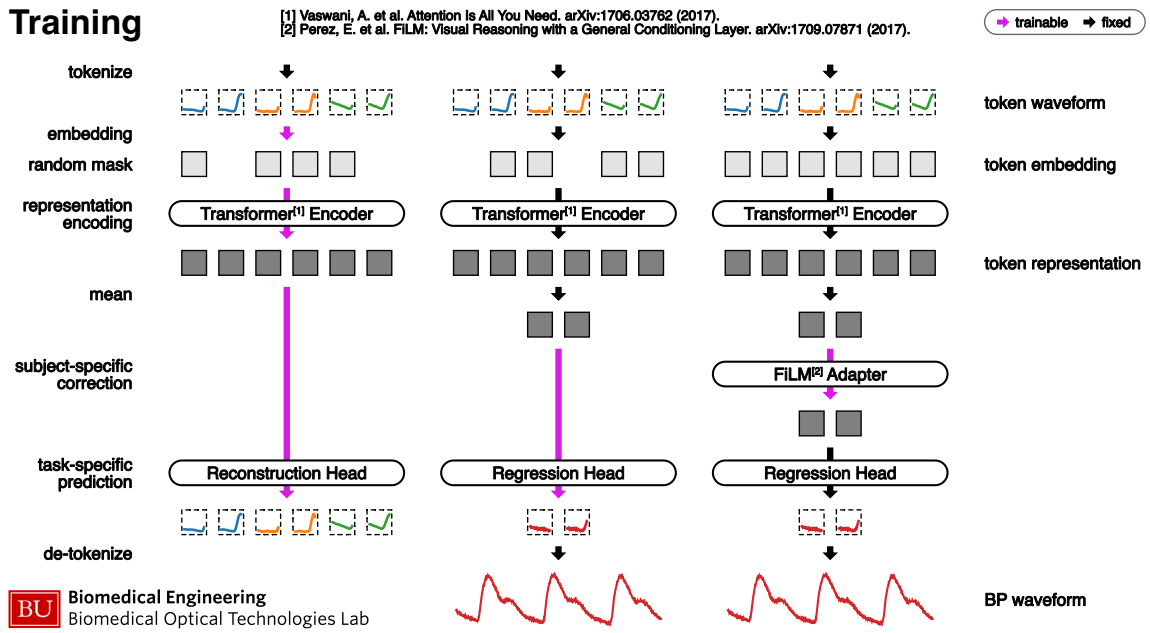

Figure 6: Three-stage training pipeline for representation learning and blood pressure (BP) prediction. Each column summarizes one stage; colored arrows indicate trainable versus frozen modules. Stage 1 pretrains the transformer encoder on optical inputs using masked-token reconstruction, in which a reconstruction head recovers the original waveform segments from randomly masked tokens. Stage 2 incorporates BP supervision by predicting the BP waveform with a regression head applied to channel-pooled token representations. Stage 3 performs lightweight measurement-specific calibration by inserting a feature-wise linear modulation (FiLM) adapter before the regression head and training only the adapter within each measurement.

Results

Stage 1

After Stage 1 self-supervised pretraining, we evaluate the learned sample-level representations by visualizing their latent structure in a two-dimensional space using uniform manifold approximation and projection (UMAP)[27]. We found that samples from the same measurement form coherent local manifolds (Figure 7, left panel). Training and test samples are interleaved within the same manifolds, indicating that the learned representations capture measurement-level structure that generalizes beyond the training samples (Figure 7, right panel). Together, these results suggest that Stage 1 pretraining learns a general-purpose representation of the optical waveforms that captures measurement-level structure without overfitting to the training samples.

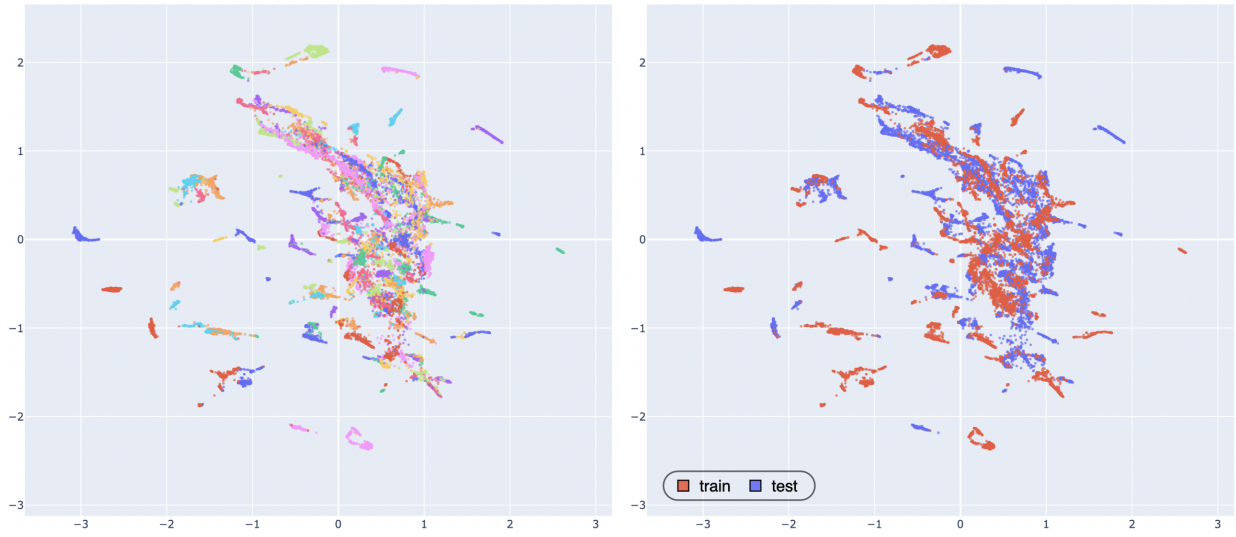

Figure 7: Visualization of the sample-level representation after Stage 1 self-supervised pretraining using uniform manifold approximation and projection (UMAP). The left panel is colored by measurement and the right panel is colored by split. In both panels, the x-axis is UMAP1 and the y-axis is UMAP2. Left panel shows that samples from the same measurement form coherent local manifolds, while right panel shows strong interleaving of train and test samples within the same manifolds without clear split-driven separation.

Subject-level information is also captured by the Stage 1 representation. First and repeat measurements from the same subject cluster together in the same local manifold, even when they are assigned to different splits. Six representative cases are shown in Figure 8: H004/H004R and N006/N006R (both train), S023/S023R and S029/S029R (both test), and S007/S007R plus S011/S011R (cross-split pairs) where R denotes the repeat measurement. Across all three split combinations, the two measurements from the same subject remain close and preserve similar local geometry. We also note nuance that the train/train pairs (H004/H004R and N006/N006R) are typically tighter, whereas the test/test pairs (S023/S023R and S029/S029R) usually retain similar shapes but with slightly larger separation distance.

Physiologically relavant information related to BP is also captured by the Stage 1 representation. In the measurement-level UMAP views for H004 and S002R, diastolic BP changes follow clear geometric trajectories (Figure 9). For H004, higher diastolic BP progresses toward the right side of the manifold; for S002R, higher diastolic BP moves from the lower-right cluster toward the upper-left branch. Comparing the diastolic-color and time-color panels further shows temporal consistency:

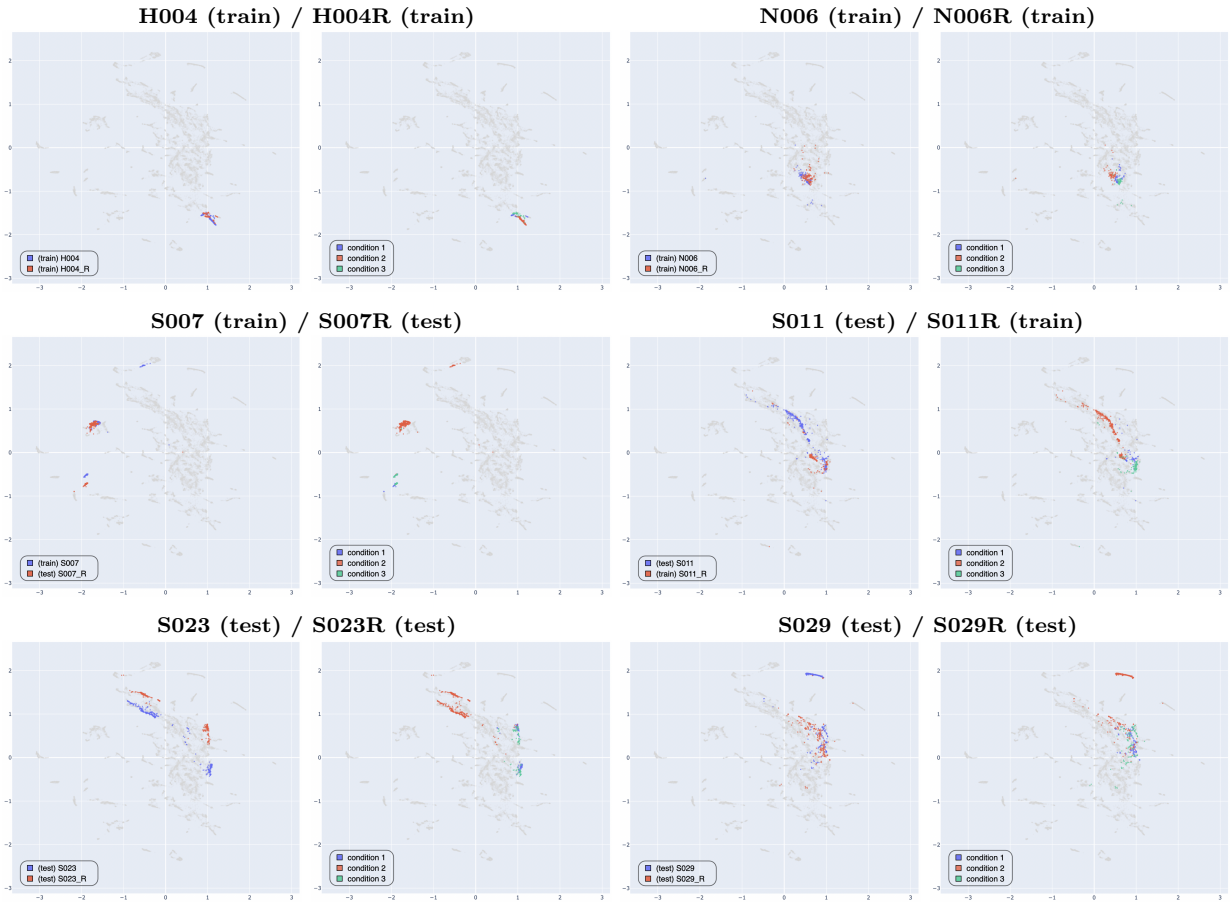

Figure 8: Uniform manifold approximation and projection (UMAP) visualization of the sample-level representation after Stage 1 self-supervised pretraining for six representative subjects, each with two measurements. For each subject, left panel is colored by measurement and right panel is colored by condition. The x-axis is UMAP1 and the y-axis is UMAP2 in all panels. Across all split combinations, two measurements from the same subject remain close and preserve similar local geometry.

low-BP regions are mainly composed of condition 1 and condition 3 samples, while condition 2 traces the rising-BP branch. Importantly, in condition 2-only views, several early condition 2 samples overlap with low-BP regions shared with conditions 1/3, which is physiologically consistent with delayed BP elevation immediately after exercise onset. Complementary principal component analysis (PCA) of Stage 1 test representations provides additional support (Figure 10): along PC1, low and high diastolic BP are clearly separated with a sparse middle region, and this bipolar pattern is also evident in subject-specific examples (S007R and S027), consistent with the rapid BP transition induced by the exercise protocol.

Stage 2

After BP-supervised training, Stage 2 shows a clear train–test generalization gap. For diastolic BP, mean absolute error (MAE) increases from 3.75 mmHg (train) to 11.97 mmHg (test), and for systolic BP from 4.66 mmHg (train) to 17.43 mmHg (test) (Figure 11, left), indicating substantial overfitting to training measurements. At the same time, waveform-level examples show that predicted BP curves preserve the main temporal shape but are shifted in absolute level (Figure 11, right), suggesting that the model learns optical-to-BP dynamics but fails to generalize absolute calibration across

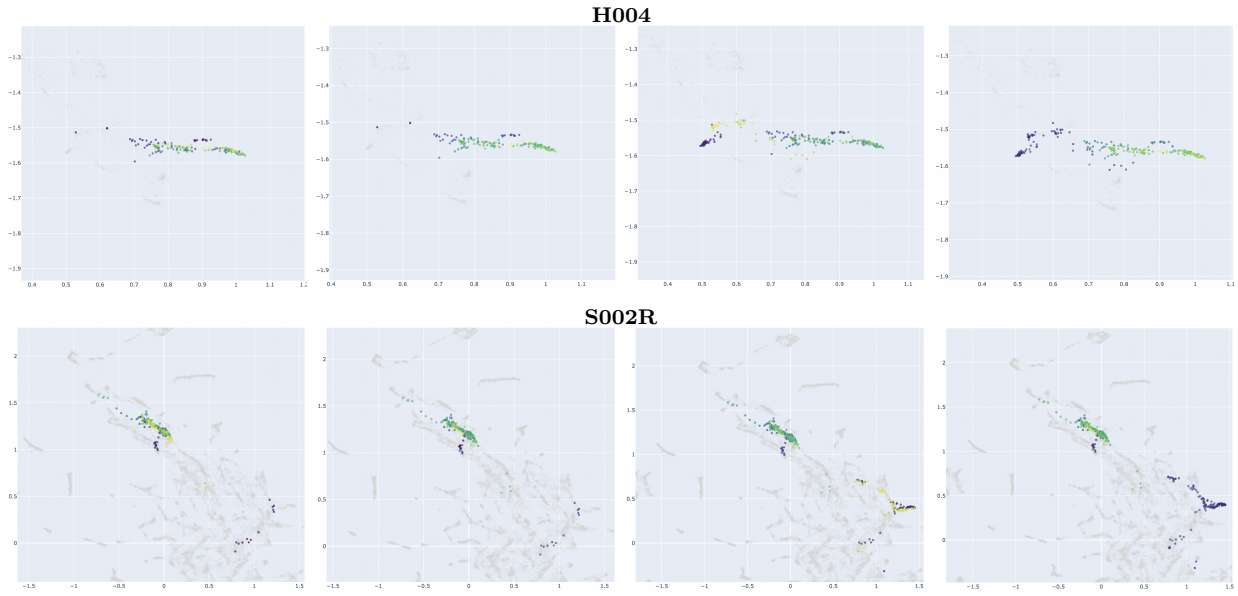

Figure 9: Condition-resolved uniform manifold approximation and projection (UMAP) visualization of the sample-level representation after Stage 1 self-supervised pretraining for two representative measurements, H004 and S002R. Columns (left to right) show condition 2 colored by time, condition 2 colored by diastolic BP, all conditions colored by time, and all conditions colored by diastolic BP. For both measurements, diastolic BP follows structured manifold trajectories, and early condition 2 points partially overlap with low-BP regions associated with conditions 1 and 3, consistent with delayed BP rise after exercise onset.

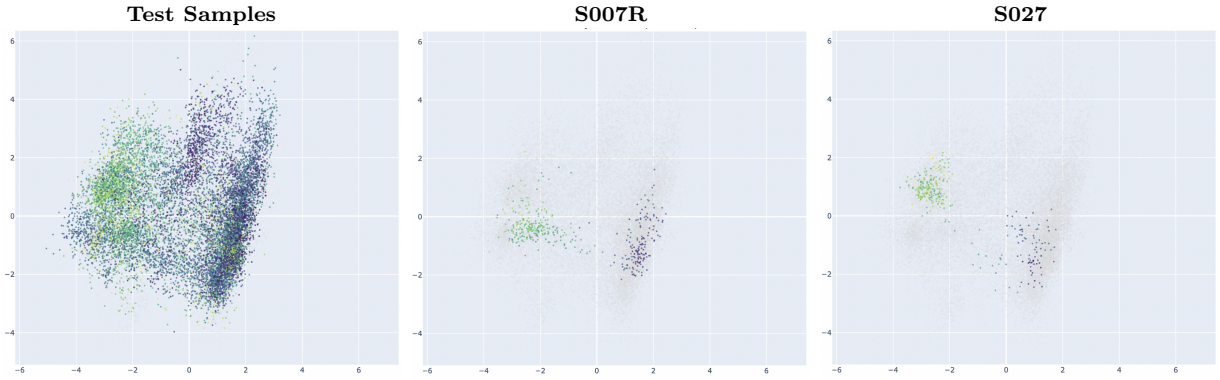

Figure 10: Principal component analysis (PCA) of Stage 1 representations on test samples, colored by diastolic blood pressure (BP). The x-axis is PC1 and the y-axis is PC2 in all panels. Left panel shows all test samples, and the middle and right panels show measurement-specific subsets (S007R and S027). Across panels, low and high diastolic BP form a bipolar structure with sparse occupancy in the middle region, consistent with rapid BP transitions under the perturbation protocol.

measurements, consistent with the observed generalization gap in BP prediction performance.

The Stage 2 representation is more tightly organized by BP, consistent with direct BP supervision (Figure 12). In the global UMAP map, measurement-level clusters remain visible and become more compact than in Stage 1. When coloring test samples by diastolic BP, local trajectories become strongly ordered by BP, and this monotonic structure is clear in zoomed examples from S013R and S025R.

However, this BP-oriented compression also weakens the condition-level structure that was preserved in Stage 1. In representative measurements S002R and S023, the four side-by-side views in Figure 13 (from left to right: S002R colored by diastolic BP, S002R colored by condition, S023 colored by diastolic BP, and S023 colored by condition) show that samples from different conditions are separated into distant islands rather than forming a shared manifold where BP progression links conditions. In particular, early condition 2 points no longer overlap with low-BP regions associated with conditions 1/3 as clearly as in Stage 1, indicating that Stage 2 features become dominated by condition-specific partitioning and reduced cross-condition continuity. This loss of structure is consistent with the poor out-of-measurement test performance observed above.

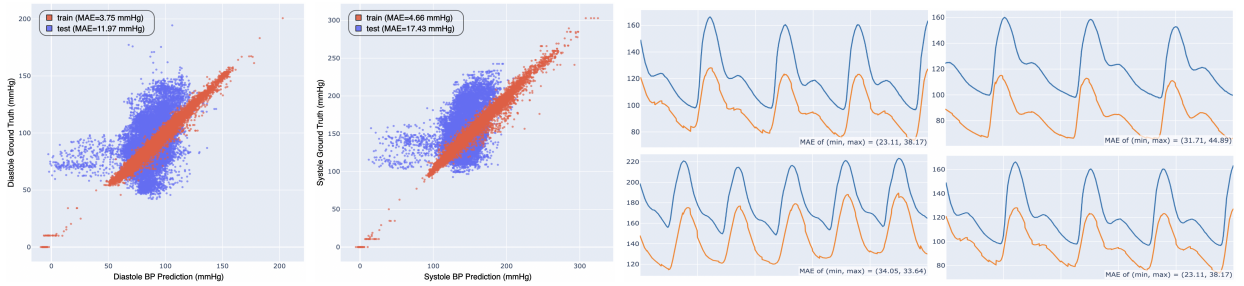

Figure 11: Blood pressure (BP) prediction result after Stage 2 supervised training using blood pressure (BP) waveform. **Left** Scatter plots of predicted versus ground-truth BP for diastolic and systolic pressure. **Right** Four representative waveform predictions showing preserved waveform shape but shifted absolute BP level. Both panels colored by train (red) versus test (blue) split. Together, these panels indicate that Stage 2 captures BP waveform dynamics but does not generalize absolute calibration across measurements.

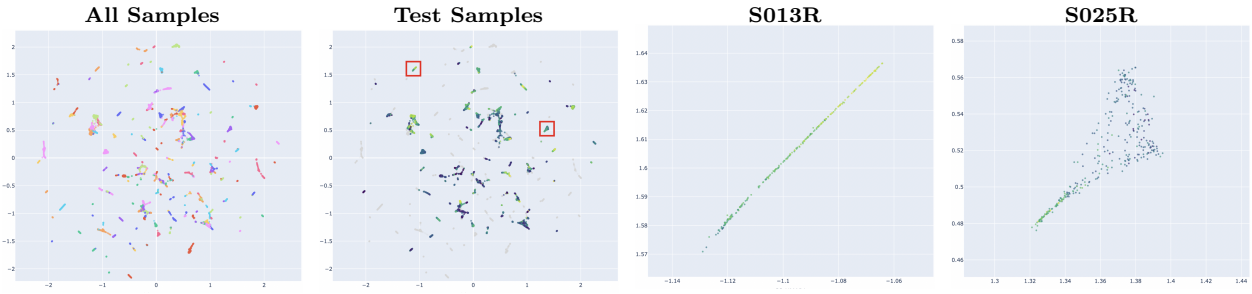

Figure 12: Uniform manifold approximation and projection (UMAP) views of sample-level representations after Stage 2 supervised training using blood pressure (BP) waveform. From left to right, panels show all samples (color by measurement), test samples (color by diastolic BP), and two zoomed measurement examples (S013R and S025R; both color by diastolic BP). In all panels, the x-axis is UMAP1 and the y-axis is UMAP2. Compared with Stage 1, Stage 2 representations are more tightly organized by BP, with local trajectories strongly ordered by diastolic BP, while measurement-level clusters remain visible but more compact.

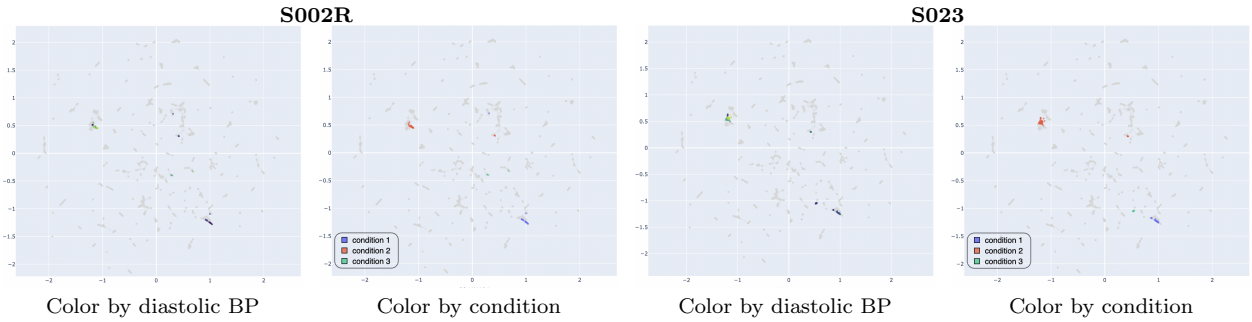

Figure 13: Uniform manifold approximation and projection (UMAP) visualization of the sample-level representation in two representative measurements, S002R and S023, after Stage 2 supervised training using blood pressure (BP) waveform. From left to right, panels show S002R (color by diastolic BP), S002R (color by condition), S023 (color by diastolic BP), and S023 (color by condition). In all panels, the x-axis is UMAP1 and the y-axis is UMAP2. Unlike Stage 1, different conditions within the same measurement are more strongly separated, indicating weakened cross-condition manifold continuity despite BP-related ordering.

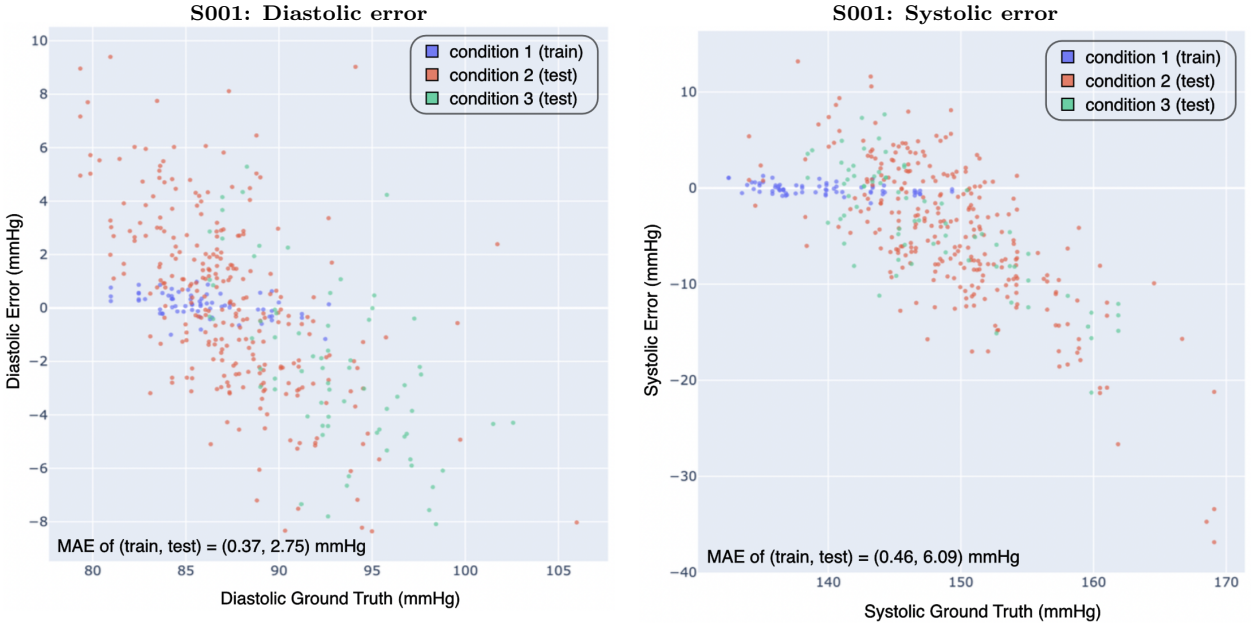

Figure 14: Stage 3 measurement-specific calibration results after feature-wise linear modulation (FiLM) adaptation, using S001 condition 1 for calibration (train) and conditions 2/3 for evaluation (test). Left panel shows diastolic error versus diastolic ground truth; right panel shows systolic error versus systolic ground truth. Colors denote condition 1 (train), condition 2 (test), and condition 3 (test). The resulting mean absolute error (MAE) values are (0.37 mmHg, 2.75 mmHg) for diastolic and (0.46 mmHg, 6.09 mmHg) for systolic in train/test order.

Stage 3

Stage 3 applies measurement-specific FiLM calibration using condition 1 samples as calibration data and evaluates generalization to conditions 2 and 3 within the same measurement. For S001, we fit the FiLM adapter on condition 1 and then predict BP on conditions 2/3. The calibrated model achieves low training error and substantially reduced test error within this measurement: diastolic MAE (train, test) = (0.37 mmHg, 2.75 mmHg) and systolic MAE (train, test) = (0.46 mmHg, 6.09 mmHg) (Figure 14). These results indicate that Stage 3 calibration can effectively correct

measurement-specific offsets.

We then fixed the Stage 3 hyperparameters selected from S001 and applied the same calibration setting to all test measurements. Figure 15 summarizes per-measurement error distributions (boxplots) for diastolic and systolic BP. Across all test measurements, aggregate MAE is 7.50 mmHg for diastolic and 14.54 mmHg for systolic. Error levels vary substantially across measurements, suggesting that a single hyperparameter setting does not fully capture subject- and measurement-specific calibration needs. Further improvement is therefore expected from measurement-adaptive hyperparameter selection and more systematic calibration tuning.

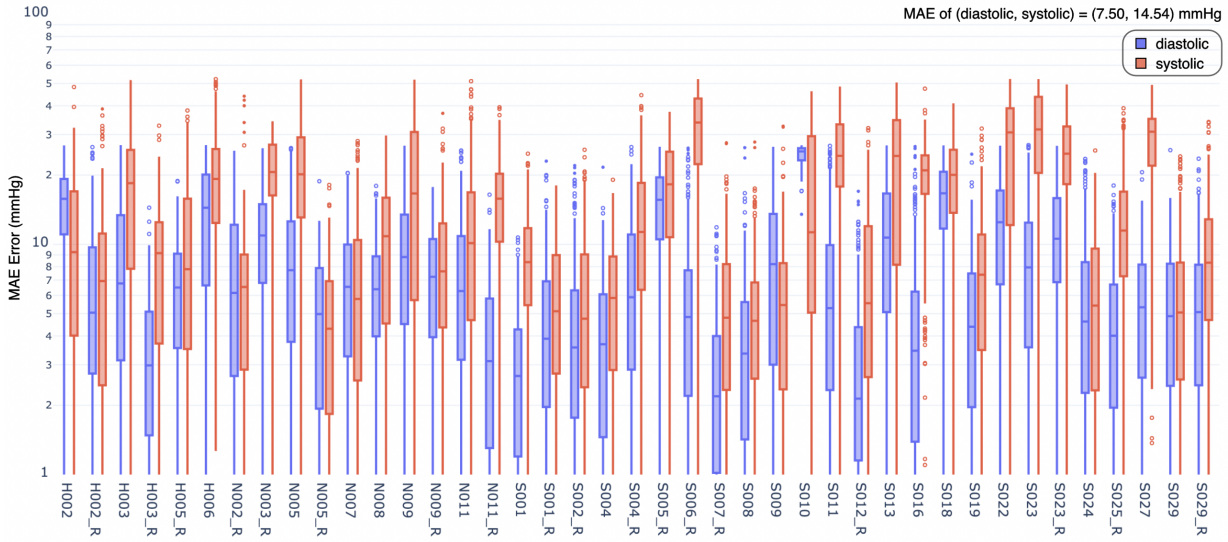

Figure 15: Per-measurement error distributions after Stage 3 feature-wise linear modulation (FiLM) calibration with a shared hyperparameter setting selected on S001 and applied to all test measurements. Blue boxplots denote diastolic error and red boxplots denote systolic error. Boxplots summarize the distribution of sample-level absolute errors within each measurement. The aggregate mean absolute error (MAE) reported in the panel is 7.50 mmHg for diastolic and 14.54 mmHg for systolic.

Discussion

This study shows that self-supervised representation learning on optical waveforms can recover physiologically meaningful structure before explicit BP supervision. Stage 1 yielded the most stable representation geometry across train/test splits, preserving both measurement-level organization and subject-level consistency across repeated measurements. Within this latent space, BP-related trajectories were already visible in both UMAP and PCA views, indicating that the pretrained encoder captures relevant hemodynamic variation.

Adding BP supervision in Stage 2 increased BP ordering in local manifolds but also introduced a clear train–test performance gap and weakened cross-condition continuity within measurements. This suggests a tradeoff between task-specific fitting and representation invariance: the model learns BP-correlated structure, but absolute calibration remains vulnerable to measurement- or subject-specific shifts. Stage 3 FiLM calibration partially addressed this issue. In the S001 case, measurement-specific calibration produced low train and test MAE, but using one shared hyperparameter setting across all test measurements still yielded substantial inter-measurement variability.

Several methodological limitations follow from these observations. First, the current measurement-level split is random with respect to repeat measurements, so representation behavior may depend

on how first/repeat measurements are assigned across train/test. Second, Stage 2 updates part of the encoder, which may overwrite favorable Stage 1 geometry. Third, Stage 3 currently relies on a fixed calibration configuration transferred from one measurement.

Future work should therefore focus on three directions. One direction is split-robust evaluation, i.e., systematically varying measurement-assignment strategies and quantifying representation stability and downstream performance under each split. A second direction is to preserve Stage 1 structure during BP supervision, for example by freezing the pretrained encoder and training only lightweight BP heads or by adding constraints that retain Stage 1 geometry. A third direction is to expand calibration and representation analysis: tune calibration hyperparameters per measurement, probe whether Stage 1 features alone can predict systolic/diastolic values with simple readouts, and analyze channel-level/latent-dimension interpretability to link learned representations to known waveform features and potentially discover new biomarkers[28].

Availability of data and code

The dataset used in this study is publicly available on the Open Science Framework (OSF), https://osf.io/yqpht/overview. Code for data preprocessing, model training, evaluation, and figure generation is available at https://github.com/tianrui-qi/SCOS-BP.

Acknowledgements

This work was completed during a PhD lab rotation in the Biomedical Optical Technologies Lab at Boston University. I thank Dr. Darren Roblyer for hosting the rotation, and and Dr. Ariane Garrett and Ana Perez for their support throughout the project.

References

[1] Banegas, J. R., Ruilope, L. M., de la Sierra, J. J. *et al.* Relationship between clinic and ambulatory blood-pressure measurements and mortality. *New England Journal of Medicine* **378**, 1509–1520 (2018).

[2] Mukkamala, R., Stergiou, G. S. & Avolio, A. P. Cuffless blood pressure measurement. *Annual Review of Biomedical Engineering* **24**, 203–230 (2022).

[3] Elgendi, M., Fletcher, R., Liang, D. *et al.* The use of photoplethysmography for assessing hypertension. *npj Digital Medicine* **2**, 60 (2019).

[4] Mieloszyk, R., Twede, H., Lester, G. *et al.* A comparison of wearable tonometry, photoplethysmography, and electrocardiography for cuffless measurement of blood pressure in an ambulatory setting. *IEEE Journal of Biomedical and Health Informatics* **26**, 2864–2875 (2022).

[5] Mukkamala, R., Shroff, S. G., Landry, G. S. *et al.* The microsoft research aurora project: important findings on cuffless blood pressure measurement. *Hypertension* **80**, 534–540 (2023).

[6] Boas, D. A. & Dunn, A. K. Laser speckle contrast imaging in biomedical optics. *Journal of Biomedical Optics* **15**, 011109 (2010).

[7] Valdes, C. P., Varma, H. M., Kristoffersen, A. K. *et al.* Speckle contrast optical spectroscopy, a non-invasive, diffuse optical method for measuring microvascular blood flow in tissue. *Biomedical Optics Express* **5**, 2769 (2014).

[8] Dragojevic, T., Hollmann, J. L., Tamborini, J. P. *et al.* Compact, multi-exposure speckle contrast optical spectroscopy (scos) device for measuring deep tissue blood flow. *Biomedical Optics Express* **9**, 322 (2018).

[9] Bi, R., Du, Y., Singh, A. B. E. *et al.* Fast pulsatile blood flow measurement in deep tissue through a multimode detection fiber. *Journal of Biomedical Optics* **25**, 055004 (2020).

[10] Westerhof, N., Lankhaar, J. W. & Westerhof, B. E. The arterial windkessel. *Medical and Biological Engineering and Computing* **47**, 131–141 (2009).

[11] Ghijsen, M., Rice, T. B., Yang, B. J. *et al.* Wearable speckle plethysmography (spg) for characterizing microvascular flow and resistance. *Biomedical Optics Express* **9**, 3937 (2018).

[12] Dunn, C. E., Monroe, D. C., Crouzet, B. *et al.* Speckleplethysmographic (spg) estimation of heart rate variability during an orthostatic challenge. *Scientific Reports* **9**, 14079 (2019).

[13] Garrett, A. *et al.* Speckle contrast optical spectroscopy for cuffless blood pressure estimation based on microvascular blood flow and volume oscillations. *Biomedical Optics Express* **16**, 3004–3016 (2025).

[14] Vaswani, A. *et al.* Attention is all you need. *arXiv* (2017). 1706.03762.

[15] Che, C. *et al.* Constrained transformer network for ecg signal processing and arrhythmia classification. *BMC Medical Informatics and Decision Making* **21**, 184 (2021).

[16] Hu, R., Chen, J. & Zhou, L. A transformer-based deep neural network for arrhythmia detection using continuous ecg signals. *Computers in Biology and Medicine* **144**, 105325 (2022).

[17] Varghese, J. & De, P. P. Transformer-based temporal sequence learners for arrhythmia classification using ecg signals. *Medical and Biological Engineering and Computing* **61**, 1993–2000 (2023).

[18] Chu, C.-H. *et al.* Real-time blood pressure estimation based on photoplethysmography and personalized transformer model. *BMC Medical Informatics and Decision Making* **23**, 133 (2023).

[19] Yang, C., Westover, M. B. & Sun, J. BIOT: Cross-data biosignal learning in the wild. *arXiv* (2023). 2305.10351.

[20] Wang, Y., Li, T., Yan, Y., Song, W. & Zhang, X. How to evaluate your medical time series classification? *arXiv* (2024). 2410.03057.

[21] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding (2019). 1810.04805.

[22] Perez, E., Strub, F., de Vries, H., Dumoulin, V. & Courville, A. Film: Visual reasoning with a general conditioning layer (2017). 1709.07871.

[23] Ansel, J. *et al.* PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)* (ACM, 2024). URL https://docs.pytorch.org/assets/pytorch2-2.pdf.

[24] Falcon, W. & The PyTorch Lightning team. PyTorch Lightning (2019). URL https://github.com/Lightning-AI/lightning.

[25] Yadan, O. Hydra - a framework for elegantly configuring complex applications. Github (2019). URL https://github.com/facebookresearch/hydra.

[26] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization (2017). 1412.6980.

[27] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction (2020). 1802.03426.

[28] Simon, E. & Zou, J. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods* **22**, 2107–2117 (2025).