OPLR: ALLEVIATING NORMALITY BIAS IN PPG-TO-ABP RECONSTRUCTION VIA ORDERED PROTOTYPE LATENT REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous arterial blood pressure (ABP) waveforms enable high-resolution hemodynamic monitoring but typically require invasive arterial lines. PPG-to-ABP reconstruction offers a noninvasive alternative, yet long-tailed BP distributions induce a systematic failure we term *Normality Bias*: predictions drift toward the normotensive mean, attenuating hypertensive peaks and elevating hypotensive troughs even when aggregate metrics appear strong. We formalize this bias through a stratified signed-error diagnostic and introduce two model-agnostic metrics: the **Normality Bias Index (NBI)**, which captures monotonic regime-dependent drift, and **TailBias**, which quantifies bias at the clinically critical extremes. Across architecture families, these diagnostics reveal that normality bias is pervasive. We propose **Ordered Prototype Latent Refinement (OPLR)**, a representation-level remedy that imposes an ordered regime geometry (Low $<$ Normal $<$ High) on the latent space and performs routing-guided prototype–token refinement before decoding. This design targets three coupled failure modes of normality bias: regime absorption (rare-regime segments encoded near the dominant Normal mode), routing monopoly (routing collapsing onto the Normal prototype), and attenuation of extreme waveform amplitudes. Under strict subject-wise, calibration-free evaluation on PulseDB and zero-shot transfer to two additional cohorts, OPLR achieves the lowest NBI and TailBias among all compared models and the lowest RMSE and MAE in every regime, including the aggregate, with gains significant under subject-level paired tests on PulseDB. The diagnostics are robust to binning choices, and the regime-level results are stable under shifts of the reporting boundaries. Matched-backbone ablations isolate the ordered prototype as the component that moves the bias metrics, show that the advantage is not explained by additional inputs alone, and find that data- and loss-level long-tailed remedies on the same backbone reduce it only marginally.

1 INTRODUCTION

Continuous ABP monitoring is central to perioperative care, yet high-fidelity waveforms require invasive catheterization (Scheer et al., 2002). Photoplethysmography (PPG) is ubiquitous in bedside and wearable monitors (Allen, 2007; Castaneda et al., 2018; Kim & Baek, 2023), motivating PPG-to-ABP reconstruction as a noninvasive alternative (Bian et al., 2024; Ibtehaz et al., 2022; Pan et al., 2024; Tian et al., 2026). While recent models report strong aggregate metrics, clinical utility is particularly sensitive to rare hypotensive and hypertensive extremes, where patient risk is elevated (Saugel & Sessler, 2021; Wesselink et al., 2018).

In retrospective datasets, these extremes form the tails of a long-tailed SBP distribution (Schrumpf et al., 2021; Wang et al., 2023), and models trained with global reconstruction losses systematically drift toward the dominant normotensive range. We term this failure *Normality Bias*. Although the direction of this drift is reminiscent of classical regression toward the mean (Barnett et al., 2005), Normality Bias is structurally different: it is a regime-stratified, signed waveform-error drift arising under long-tailed training distributions and global reconstruction losses, rather than a repeated-measurement artifact. The clearest evidence that it is not an irreducible property of the data is that it moves with the model: on an identical test distribution, restructuring the latent geometry more

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

than halves it while also lowering RMSE, whereas simply stretching predictions would typically increase error. To make this bias measurable, we introduce a stratified signed-error diagnostic and two complementary metrics: the Normality Bias Index (NBI), which captures monotonic regime-dependent drift, and TailBias, which quantifies bias at the clinically critical extremes. Applied to five baselines, NBI reveals that normality bias is pervasive and that aggregate metrics can fail to detect it: U^2-Net achieves the lowest baseline RMSE (11.94 mmHg) yet retains an NBI of 0.39, more than twice that of our method introduced below.

Despite mature long-tailed methods in classification and imbalanced regression (Cui et al., 2019; Menon et al., 2021; Ren et al., 2020; Yang et al., 2021; Gong et al., 2022; Zha et al., 2023; Keramati et al., 2024), the PPG-to-ABP literature has no dedicated solution for this problem. The gap is structural for three reasons: naive resampling duplicates entire waveforms rather than diversifying rare-regime instances; loss reweighting amplifies gradient magnitude but does not reorganize a latent geometry dominated by the normotensive majority; and rare-regime fidelity in waveform reconstruction depends on preserving fine-grained temporal morphology, which interventions on aggregate labels do not directly address. We show empirically (Appendix K) that oversampling (Buda et al., 2018), class-balanced reweighting (Cui et al., 2019), and Label Distribution Smoothing (Yang et al., 2021)—applied individually and in combination on the same backbone—fail to reduce NBI to OPLR's level or consistently improve rare-regime fidelity, indicating that data/loss-level fixes are insufficient for correcting representation-level normality bias. These results suggest that mitigating normality bias requires restructuring the latent representation itself, rather than only changing sample or loss weights.

We therefore propose **Ordered Prototype Latent Refinement (OPLR)**, a representation-level remedy that imposes an ordered regime geometry (Low $<$ Normal $<$ High) on latent representations and performs routing-guided prototype–token refinement before waveform decoding. Rather than using regime only as an auxiliary label or a decoder-side conditioning signal, OPLR uses routing as an ordinal structural prior to reorganize the latent space itself. Ordered routing is designed to reduce regime absorption, distributional regularization discourages routing monopoly, and routing-guided refinement with extrema preservation targets attenuation of extreme waveform amplitudes; these three failure modes are defined in Section 3. Viewed this way, OPLR is not proposed as a new attention primitive, but as a bias-aligned latent restructuring mechanism motivated by the failure modes revealed by NBI and TailBias. Matched-backbone controls (Section 6) support this reading: giving the strongest baseline the same ordinal supervision as an auxiliary head, or the same decoder conditioning, leaves its NBI essentially unchanged, whereas OPLR reduces it by more than half, with or without its extra inputs.

Prior PPG-to-ABP work has focused on average-case reconstruction (Brophy et al., 2021; Huang et al., 2025; Ibtehaz et al., 2022; Pan et al., 2024; Tian et al., 2026) or orthogonal directions such as multimodal cardiovascular signal generation (Chen et al., 2026) and multi-directional waveform conversion (Meyer et al., 2026); furthermore, distribution-dependent BP prediction degradation has been diagnosed in scalar BP estimation (Schrumpf et al., 2021), but not structurally addressed in continuous waveform reconstruction. Our contributions are threefold: we formalize Normality Bias in calibration-free PPG-to-ABP reconstruction; we introduce two simple model-agnostic diagnostics, NBI (a regime-stratified signed-error slope) and TailBias, that expose this failure even when aggregate metrics remain strong; and we show that data/loss-level long-tailed fixes and ordinal supervision alone are insufficient, motivating OPLR as a representation-level remedy that improves rare-regime fidelity under strict subject-wise evaluation and zero-shot transfer to two additional cohorts.

2 Diagnosing Normality Bias

We study calibration-free PPG-to-ABP waveform reconstruction. Given a PPG segment $x \in \mathbb{R}^{1\times T}$, the goal is to predict a continuous ABP waveform $\hat{y} = f(x) \in \mathbb{R}^{1\times T}$ ($T = 1280$ at $f_s = 125$ Hz) without subject-specific calibration. We define $K = 3$ ordinal regimes based on segment-level systolic blood pressure (SBP): Low (SBP $<$ 90 mmHg), Normal ($90 \leq$ SBP ≤ 140 mmHg), and High (SBP $>$ 140 mmHg). These thresholds are used as clinically motivated supervision anchors of hypotension and hypertension (Saugel et al., 2024; Unger et al., 2020). In PulseDB (Wang et al., 2023), the test-set distribution is 79.5% Normal, 18.2% High, and 2.3% Low, reflecting the strong normotensive skew commonly observed in retrospective monitoring datasets.

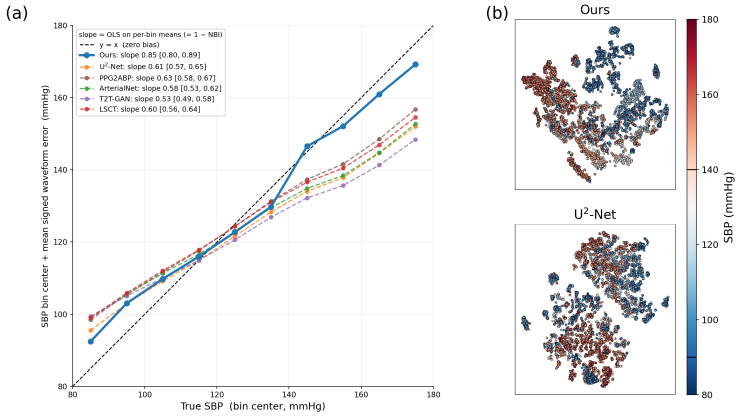

Figure 1: **Normality Bias in reconstruction error and in latent space** (PulseDB test set, calibration-free). **(a)** For each true-SBP bin, $c_m + \mu_m(f)$, the bin center plus the bin-mean signed waveform error of Eq. (1), against true SBP. The dashed diagonal is zero bias; a slope below one is drift toward the population mean. Legend: OLS slope on the bin means, which equals $1 - \mathrm{NBI}$ (Table 2), with 95% subject-bootstrap CI (2,000 resamples). Baselines compress the range (slopes 0.53–0.63); OPLR (Ours) reaches 0.85. **(b)** t-SNE of pooled latent tokens for OPLR and U^2-Net, colored by true SBP (all six models in Appendix D.2). OPLR shows an ordered low-to-high organization; U^2-Net does not.

Standard waveform metrics (RMSE, MAE, PC) aggregate errors across this imbalanced distribution, allowing models that fit the dominant Normal regime to achieve strong aggregate scores while masking systematic errors in rare extremes. To expose this hidden failure, we examine how signed reconstruction error varies as a function of SBP. We partition the SBP range into M fixed-width bins $\{B_m\}_{m=1}^{M}$ with centers c_m, and for each non-empty bin compute the mean signed waveform error:

$$\mu_m(f) = \frac{1}{|\mathcal{I}_m|T} \sum_{i \in \mathcal{I}_m} \sum_{t=1}^{T} \big(\hat{y}_i(t) - y_i(t)\big), \tag{1}$$

where $\mathcal{I}_m$ is the set of segments whose true SBP falls in bin B_m. A positive μ_m indicates systematic overestimation; a negative value indicates underestimation. Figure 1(a) plots $c_m + \mu_m(f)$ against c_m for five baselines and OPLR, so that an unbiased model lies on the diagonal. A striking pattern emerges: baselines sit above the diagonal in hypotensive bins and below it in hypertensive bins, crossing near the normotensive center, with fitted slopes of only 0.53–0.63. This signature is consistent with predictions being pulled toward the population mean regardless of the true regime.

To condense this profile into scalar diagnostics, we introduce two complementary metrics. The Normality Bias Index (NBI) captures the slope of regime-dependent drift: we fit $\mu_m(f) \approx \alpha_f + \beta_f \cdot c_m$ via unweighted least squares and define

$$\mathrm{NBI}(f) = -\beta_f. \tag{2}$$

A larger NBI indicates stronger monotonic drift toward the normotensive mean; $\mathrm{NBI} \approx 0$ indicates no monotonic regime-dependent drift, although constant global bias may remain. TailBias captures the magnitude of bias at the clinically critical extremes:

$$\mathrm{TailBias}(f) = \tfrac{1}{2}\big(|\mu_{m_{\min}}(f)| + |\mu_{m_{\max}}(f)|\big), \tag{3}$$

where $m_{\min}$ and $m_{\max}$ are the lowest and highest non-empty SBP bins. Unless stated otherwise, both are computed on 10 mmHg bins spanning $[80, 180]$ mmHg; sensitivity to this construction is examined in Section 6. NBI and TailBias are complementary: a model could have moderate NBI but large TailBias, or vice versa. Both are model-agnostic post-hoc diagnostics.

As shown in Figure 1 and Table 2, NBI reveals normality bias across all evaluated architecture families. Crucially, NBI is not redundant with aggregate metrics: U^2-Net achieves the lowest baseline RMSE yet retains substantial NBI and TailBias. We therefore evaluate models along four axes: overall fidelity, class-stratified performance, rare-regime performance, and direct bias diagnostics.

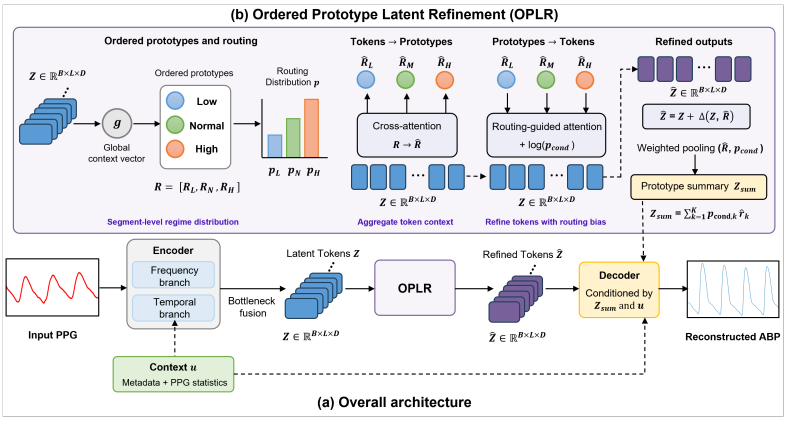

Figure 2: **Overview of Ordered Prototype Latent Refinement (OPLR). (a) Overall architecture:** latent tokens Z are extracted from the input PPG, refined by OPLR, and decoded into the reconstructed ABP waveform using prototype summary Z_{sum} and context features u. **(b) OPLR module:** a global context vector g produces a segment-level routing distribution p over $K = 3$ ordered prototypes (Low, Normal, High). The prototypes first aggregate token context and then refine the tokens through routing-guided prototype-to-token attention. OPLR outputs the refined token sequence $\hat{Z}$ and prototype summary Z_{sum}.

3 METHODS

Three Failure Modes. Under long-tailed BP distributions, a global reconstruction objective is prone to producing a normal-dominant latent geometry. We name three coupled consequences, each of which OPLR targets with one mechanism. *Regime absorption*: rare Low and High segments are encoded close to the dominant Normal mode, so their reconstructions are pulled toward Normal. *Routing monopoly*: any regime-level routing distribution concentrates on the Normal prototype, so Low and High prototypes are rarely used during refinement. *Amplitude attenuation*: the decoder underestimates hypertensive peaks and overestimates hypotensive troughs. Rather than treating regime as an auxiliary label, OPLR imposes an ordered regime geometry on latent representations and uses routing-guided prototype–token interaction to refine latent tokens before decoding (Figure 2).

Backbone and Tokenization. We keep the encoder–decoder backbone fixed and place our contribution in latent refinement. A dual-branch time–frequency encoder extracts complementary PPG features using a 1D Swin-style branch over STFT features (Allen & Rabiner, 1977; Liu et al., 2021) and a 1D ResNet branch over the PPG waveform with FiLM conditioning (He et al., 2016; Perez et al., 2018). The fused bottleneck representation is projected into latent tokens $Z \in \mathbb{R}^{B\times L\times D}$ with $D = 64$. OPLR operates on these tokens, and a 1D U-Net decoder (Ronneberger et al., 2015) reconstructs the ABP waveform from the refined representations, with regime-aware conditioning applied to the temporal skip connections. Full architecture details are in Appendix B.

Ordered Prototype Routing. The first stage of OPLR addresses regime absorption and routing monopoly by organizing the latent space around $K = 3$ learnable regime prototypes $R_0 \in \mathbb{R}^{K\times D}$, corresponding to the clinically ordered regimes Low $<$ Normal $<$ High. Let $R \in \mathbb{R}^{B\times K\times D}$ denote the batch-broadcasted copy of R_0. A pooled global context vector $g \in \mathbb{R}^{B\times D}$ is obtained by attention pooling over Z, and a soft routing distribution $p \in \mathbb{R}^{B\times K}$ is computed through temperature-scaled cosine similarity:

$$p_k = \frac{\exp\left(\frac{\cos(g, r_k)}{\tau}\right)}{\sum_{j=1}^{K} \exp\left(\frac{\cos(g, r_j)}{\tau}\right)}, \qquad \cos(g, r) = \frac{g^\top r}{\|g\|\,\|r\|}, \tag{4}$$

where τ is a temperature parameter scheduled during training via cosine decay. Routing is soft and there are no separate decoding pathways: every token passes through the same attention operation

over all three prototypes, and nothing switches discontinuously as true SBP crosses a threshold (Appendix D.1). The routing distribution serves as an ordinal structural prior for the refinement stage that follows.

Ordered Supervision and Anti-Monopoly Regularization. The routing mechanism differs from standard prototype classification in two ways. First, to encode the ordering Low $<$ Normal $<$ High, we apply an ordinal distance-margin loss to the routing logits, subtracting from each non-target logit an offset proportional to its ordinal distance $|k - c|$ from the ground-truth regime c (Appendix C). Second, to prevent routing monopoly, we regularize the batch-mean routing distribution toward the training-set regime proportions (Appendix C), discouraging collapse onto the Normal prototype.

Routing-Guided Prototype–Token Refinement. Routing alone does not guarantee faithful tail amplitudes: rare-regime segments may still be decoded from normal-dominant latent content. OPLR therefore performs routing-guided prototype–token refinement, in which the prototypes act as latent anchors that first collect regime-specific evidence from the token sequence and then redistribute it back to the tokens under an ordinal routing prior.

Tokens → Prototypes. OPLR updates the prototypes using sample-specific evidence aggregated from the latent token sequence via multi-head attention (Vaswani et al., 2017):

$$\hat{R} = R + \mathrm{MHA}(Q = R,\ K = Z,\ V = Z), \tag{5}$$

followed by LayerNorm and a feed-forward residual block, yielding context-enriched prototype states $\hat{R} \in \mathbb{R}^{B\times K\times D}$.

Prototypes → Tokens. The key refinement step writes regime information back to the token sequence. Let $p_{\mathrm{cond}} \in \mathbb{R}^{B\times K}$ denote the conditioning routing distribution. During training it is a scheduled mixture of the predicted routing distribution and the ground-truth one-hot regime label, with the teacher-forcing probability annealed to zero (Bengio et al., 2015); at convergence and at inference it is the model's own prediction, so there is no train–test mismatch (Appendix B). For token i attending to prototype j, we compute

$$A_{ij} = \frac{(z_i W_Q)(\hat{r}_j W_K)^\top}{\sqrt{d_h}} + \log(p_{\mathrm{cond},j}), \tag{6}$$

$$\alpha_{ij} = \mathrm{softmax}_j(A_{ij}), \qquad \hat{z}_i = z_i + \sum_{j=1}^{K} \alpha_{ij}\,(\hat{r}_j W_V). \tag{7}$$

The log-probability term makes this update routing-guided rather than content-only: tokens are encouraged to consult the prototype consistent with the inferred ordinal regime, reducing the tendency of rare-regime tokens to be absorbed into the Normal representation. Two choices bound the damage from a wrong route: the log-bias is lower-clamped, so a confident routing distribution cannot fully mask the other prototypes, and the update is residual, so a misrouted sample receives an additive correction on top of intact token content rather than having its content replaced. The refined token sequence $\hat{Z}$ replaces Z as the primary input to the decoder.

Lightweight Prototype Summary for Decoder Conditioning. In addition to $\hat{Z}$, we compute a compact prototype summary

$$Z_{\mathrm{sum}} = \sum_{k=1}^{K} p_{\mathrm{cond},k}\,\hat{r}_k \in \mathbb{R}^{B\times D}, \tag{8}$$

which is fused with metadata and PPG statistics and injected into decoder skip connections via FiLM modulation as a global conditioning signal; the primary reconstruction pathway remains $\hat{Z}$.

Failure-Mode-Aligned Objective. The training objective mirrors the three failure modes. Reconstruction fidelity ($\mathcal{L}_{\mathrm{recon}}$) combines time-domain MSE and a frequency-domain STFT loss. Regime-structure regularization ($\mathcal{L}_{\mathrm{regime}}$) enforces ordered routing and discourages routing monopoly by regularizing batch-level routing usage toward the training-set regime proportions. Extrema preservation ($\mathcal{L}_{\mathrm{peak}}$) penalizes errors at the segment-level systolic maximum and diastolic minimum using

differentiable max/min approximations. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{regime}} + \mathcal{L}_{\text{peak}}. \tag{9}$$

Formulations and hyperparameters are in Appendix C. Section 6 shows that $\mathcal{L}_{\text{peak}}$ alone does not reduce the bias metrics; it converts the structure created by the ordered prototype into amplitude-faithful output.

4 Experimental Setup

We evaluate under a strict subject-wise, calibration-free protocol on PulseDB (Wang et al., 2023) (2,779 subjects, 835K segments; test: 91,843 segments from 278 subjects, 79.5% Normal, 18.2% High, 2.3% Low). *Calibration-free* means no subject-specific BP calibration, no BP label, and no regime threshold at inference; the model sees PPG plus fixed demographic and PPG-statistic context. We report zero-shot transfer to two external cohorts with unchanged weights, no calibration, and the same preprocessing. The University of Queensland Vital Signs Dataset (Liu et al., 2012) contains 32 anesthesia cases, of which 11 carry synchronized ABP and PPG; after the same preprocessing and quality control as PulseDB, 4 subjects and 2,825 segments remain (88.6% Normal, 9.6% High, 1.7% Low; the Low regime comprises 49 segments from 3 subjects), so UQ serves as a descriptive stress test. MIMIC-IV (Johnson et al., 2023; Moody et al., 2022) is a later ICU data release from the institution behind MIMIC-III, one of PulseDB's sources; it contributes 30 subjects and 24,022 segments (74.3% Normal, 21.5% High, 4.2% Low). Subject-level disjointness from PulseDB could not be verified (Section 7). We compare against five baselines (Appendix A) spanning supervised (Ibtehaz et al., 2022; Pan et al., 2024), time–frequency latent (Bian et al., 2024), adversarial (Brophy et al., 2021), and cohort-aware (Huang et al., 2025) families, retrained from scratch on the same PulseDB split with identical preprocessing and protocol, using their published architectures, PPG-only inputs, and hyperparameters under our schedule and compute budget. Because OPLR additionally uses age, gender, and segment-level PPG statistics, we report input-parity controls in Section 6. Following the four-axis framework of Section 2, we report Overall, Class-stratified, High/Low, NBI, TailBias, and SBP/DBP error. Significance markers in Tables 1 and 2 summarize paired tests (segment-level on all cohorts; subject-level on PulseDB and MIMIC-IV) detailed in Appendix F; segment-level standard deviations, PRD, and FD are in Appendix E; preprocessing and training details in Appendix B.

5 Results

5.1 Waveform Reconstruction across Cohorts

Table 1 reports RMSE, MAE, and Pearson correlation (PC) across four evaluation settings on all three cohorts. In Tables 1 and 2, $*$ marks cells in which OPLR is significantly better than all five baselines under paired tests over segments, and † cells in which this also holds over subjects (Appendix F). On PulseDB, OPLR attains the lowest RMSE and MAE in all four settings (Overall 11.14 vs. 11.94 for U^2-Net; Class-stratified, where regime proportions are equalized, 12.50 vs. 14.30), with the largest margin in the High regime (14.47 vs. 17.67 for PPG2ABP, the best baseline there) and a smaller but consistent Low advantage (12.89 vs. 13.27 for U^2-Net). Every contrast is significant at the segment level and, because subjects contribute unequal numbers of tail segments, also at the subject level (278 subjects, paired Wilcoxon; Appendix F): against U^2-Net, OPLR lowers subject-mean RMSE by 0.67 mmHg Overall ($p = 3 \times 10^{-5}$), by 4.03 in High ($p = 2 \times 10^{-19}$), and by 1.70 in Low ($p = 6 \times 10^{-6}$). PC, which is invariant to uniform amplitude shifts, remains marginally higher for U^2-Net (0.920 vs. 0.913 Overall) and should be read together with the amplitude-sensitive metrics.

Both external cohorts reproduce the pattern. On UQ, OPLR has the lowest RMSE and MAE in every setting (Overall 11.53 vs. 12.44 for U^2-Net; Low 16.65 vs. 18.92); every segment-level contrast is significant, the narrowest being the High-regime RMSE contrast with U^2-Net ($p = 0.03$, 272 segments), and with 4 subjects we treat the cohort as descriptive. On MIMIC-IV, an ICU cohort with a larger tail (Low: 998 segments from 20 subjects), OPLR again has the lowest RMSE and MAE in every setting (Overall 14.61, High 18.37, Low 12.90, against 15.54 / 18.69 / 14.85 for LSCT, the closest baseline) and the highest PC in three of four; all segment-level contrasts are significant. At the subject level (30 subjects), the Class-stratified contrasts and the Overall and Low MAE contrasts are significant against all five baselines, whereas the Overall RMSE contrast with LSCT falls just

Table 1: Waveform reconstruction on the internal PulseDB test set (calibration-free, 278 unseen subjects) and under zero-shot transfer to UQ (4 subjects) and MIMIC-IV (30 subjects). RMSE and MAE in mmHg; PC = Pearson correlation; best per column and setting in bold. $*$: OPLR significantly better than all five baselines over segments (paired Wilcoxon, $p < 0.05$); †: also over subjects (PulseDB, MIMIC-IV). Markers apply to RMSE and MAE; the UQ Low regime has 49 segments from 3 subjects. SD, PRD, and FD: Appendix E.

		PulseDB (internal)			UQ (external)			MIMIC-IV (external)		
Setting	Method	RMSE	MAE	PC	RMSE	MAE	PC	RMSE	MAE	PC
Overall	T2T-GAN	33.22	27.86	0.405	32.12	27.27	0.470	34.32	28.95	0.357
	ArterialNet	14.76	12.40	0.819	16.51	13.94	0.825	17.44	14.71	0.785
	LSCT	13.10	11.04	0.863	14.92	12.44	0.762	15.54	12.74	0.768
	PPG2ABP	12.53	10.56	0.901	20.39	15.79	0.533	23.19	19.48	0.673
	U^2-Net	11.94	10.16	**0.920**	12.44	10.44	**0.922**	19.54	16.86	0.766
	Ours	**11.14***†	**9.41***†	0.913	**11.53***	**9.58***	0.908	**14.61***	**11.72***†	**0.821**
Class-stratified	T2T-GAN	34.17	27.90	0.403	35.57	29.50	0.451	34.52	28.52	0.332
	ArterialNet	17.49	14.87	0.794	20.62	17.62	0.766	17.79	15.00	0.780
	LSCT	15.92	13.63	0.842	18.84	16.04	0.734	16.14	13.29	0.768
	PPG2ABP	15.02	12.94	0.891	22.91	18.96	0.570	23.34	19.79	0.675
	U^2-Net	14.30	12.31	**0.912**	15.64	13.52	**0.901**	19.87	17.14	0.755
	Ours	**12.50***†	**10.48***†	0.893	**14.12***	**11.96***	0.874	**15.14***†	**12.23***†	**0.822**
High	T2T-GAN	42.82	34.39	0.426	43.74	33.94	0.482	41.65	34.21	0.420
	ArterialNet	22.03	18.54	0.812	21.27	18.42	0.848	20.70	17.48	0.846
	LSCT	19.39	16.22	0.860	23.57	20.70	0.831	18.69	15.12	**0.867**
	PPG2ABP	17.67	14.92	0.907	22.05	18.86	0.697	25.59	21.59	0.737
	U^2-Net	19.61	16.87	**0.913**	17.65	15.43	**0.907**	22.42	19.53	0.796
	Ours	**14.47***†	**12.12***†	0.908	**16.13***	**13.02***	0.853	**18.37***	**14.88***	0.861
Low	T2T-GAN	28.50	22.83	0.389	33.48	29.23	0.420	29.31	23.53	0.235
	ArterialNet	17.51	15.22	0.738	25.16	21.80	0.672	16.15	13.62	0.717
	LSCT	16.92	15.00	0.795	20.16	17.40	0.676	14.85	12.45	0.686
	PPG2ABP	16.24	14.51	0.859	27.02	23.38	0.508	21.80	18.69	0.617
	U^2-Net	13.27	11.59	**0.895**	18.92	17.02	**0.886**	17.98	15.26	0.698
	Ours	**12.89***†	**10.78***†	0.851	**16.65***	**15.09***	0.855	**12.90***	**10.44***†	**0.788**

short ($p = 0.05$) and, in the High regime (18 subjects), only the contrast with T2T-GAN reaches significance (Appendix F).

5.2 Normality Bias Diagnostics

Table 2 reports the bias diagnostics for all six models on all three cohorts. OPLR achieves the lowest NBI and TailBias among all compared models on every cohort (PulseDB: 0.148 / 6.63 vs. next-best 0.369 / 15.85), and every pairwise difference excludes zero under a paired segment bootstrap; on PulseDB the differences also hold under subject resampling (ΔNBI vs. PPG2ABP $+0.221$ [0.195, 0.247]), whereas on MIMIC-IV (30 subjects) the subject-resampled intervals against U^2-Net and PPG2ABP include zero. The ranking is not the RMSE ranking: U^2-Net, the strongest baseline on every aggregate metric, retains an NBI 2.6 times OPLR's. On MIMIC-IV every model drifts less (NBI 0.17–0.42) and OPLR's margin over U^2-Net is narrow; but the two baselines closest to OPLR in NBI carry large constant offsets, which NBI by design does not count: Overall SBP ME of $+6.6$ mmHg for U^2-Net and $+9.5$ for PPG2ABP, against $+0.8$ for OPLR (Appendix G).

Derived SBP/DBP signed error provides a clinical readout of the same effect (Appendix G). OPLR's Overall SBP ME is $+0.5$ mmHg on PulseDB and $+0.8$ on MIMIC-IV, against -5.6 and $+6.6$ for U^2-Net; on UQ every model overshoots (OPLR $+5.7$, U^2-Net $+10.8$). The residual positive bias at the hypotensive end of Figure 1(a) shows that normality bias is reduced substantially at the hypertensive tail and only partially at the hypotensive one.

Qualitative results. In Figure 3's median-RMSE segments, baselines attenuate hypertensive peaks and elevate hypotensive troughs; OPLR stays nearest the reference but retains a hypotensive offset.

Table 2: Normality-bias diagnostics on the Overall test distribution of each cohort (internal PulseDB; external UQ and MIMIC-IV, zero-shot). NBI: slope of the signed-error profile over 10 mmHg SBP bins in $[80, 180]$; TailBias: mean absolute signed error in the two extreme bins (mmHg). Lower is better; best in bold. $*$: OPLR lower than all five baselines under a paired segment bootstrap (95% CI excluding zero); †: also under subject bootstrap. Derived SBP/DBP signed bias is in Appendix G.

	PulseDB (internal)		UQ (external)		MIMIC-IV (external)	
Method	NBI	TailBias	NBI	TailBias	NBI	TailBias
T2T-GAN	0.466	20.51	0.608	24.49	0.418	17.52
ArterialNet	0.421	18.16	0.538	21.83	0.277	12.58
LSCT	0.399	17.25	0.537	22.23	0.321	13.77
PPG2ABP	0.369	15.85	0.478	20.83	0.239	11.80
U^2-Net	0.389	16.77	0.358	16.88	0.184	9.16
Ours	**0.148***†	**6.63***†	**0.251***	**12.09***	**0.174***	**8.24***

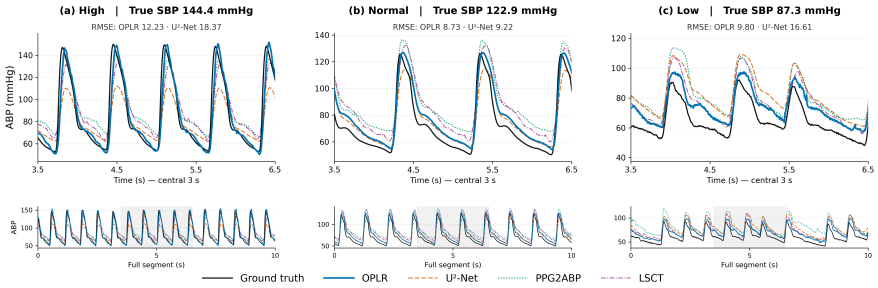

Figure 3: **Qualitative regime-wise comparison (High/Normal/Low).** For each regime, the PulseDB test segment whose OPLR RMSE is closest to the regime median; top: central 3 s, bottom: full segment (window shaded). Baselines with strong aggregate metrics attenuate hypertensive peaks and elevate hypotensive troughs; OPLR tracks the reference most closely in all three, with a residual offset in the Low regime (Section 7).

6 Analysis

Threshold and binning robustness. Re-partitioning the test set under three alternative boundary pairs changes the Low-regime population nearly twenty-fold (564 to 10,912 segments), yet Class-stratified RMSE stays within 12.4–13.0 and High and Low RMSE move smoothly with the definition rather than jumping at the supervision anchors (Appendix I); NBI and TailBias use the continuous SBP profile and do not depend on the boundaries. The diagnostics are also stable: across eleven binning configurations OPLR has the lowest NBI and TailBias in every one, with NBI within 0.15–0.18 and TailBias within 4.3–9.5, while the best baseline never falls below 0.37 and 10.2 (Appendix J).

Ablation and component analysis. Table 3 adds each component alone to a bare backbone, with encoder, decoder, data, and schedule held fixed. Three findings follow. First, $\mathcal{L}_{\text{peak}}$, decoder conditioning, and the frequency encoder added alone leave NBI within 0.363–0.370 against a backbone at 0.366, and TailBias within 0.2 of it: the TailBias reduction is not a direct consequence of the extrema loss. Second, the ordered prototype is the only component that moves either metric (NBI 0.339, TailBias 13.65) and gives the largest Low-regime gain. Third, no single addition reproduces the full model, and the ordered prototype alone does not improve High: it creates regime-separated structure that needs the conditioning path to reach the decoder and $\mathcal{L}_{\text{peak}}$ to become amplitude. No single addition improves Overall RMSE over the bare backbone either (12.17–12.51), whereas the full model reaches 11.14 with NBI 0.148: the components are complementary rather than additive. Removal from the full model gives the converse (Table 3): without the ordinal margin ($m = 0$), NBI returns to the backbone level (0.369) and TailBias to 14.91; without $\mathcal{L}_{\text{dist}}$, NBI is 0.317 and TailBias 12.66, with High RMSE slightly lower (14.09 vs. 14.47). The ordinal margin is necessary; prototypes without it do not reduce the bias.

Table 3: Component analysis on PulseDB (RMSE in mmHg). Top: each component added alone to a bare backbone; bottom: the two regime-structure terms, each removed alone from the full model. Encoder, decoder, data, and schedule are fixed throughout.

Variant	Overall RMSE	High RMSE	Low RMSE	NBI	TailBias
Bare backbone	12.17	17.03	16.37	0.366	15.86
$+\mathcal{L}_{\text{peak}}$ only	12.51	18.18	15.60	0.369	15.88
+ Decoder conditioning only	12.45	18.58	14.78	0.363	15.76
+ Frequency encoder only	12.31	18.60	15.01	0.370	15.93
+ Ordered prototype only	12.18	17.83	13.87	0.339	13.65
Full OPLR (Ours)	**11.14**	**14.47**	**12.89**	**0.148**	**6.63**
Full − ordinal margin ($m = 0$)	11.87	15.70	15.55	0.369	14.91
Full − $\mathcal{L}_{\text{dist}}$	11.99	14.09	13.47	0.317	12.66

Input parity and ordinal supervision alone. Two controls test whether OPLR's advantage is explained by its extra inputs or by ordinal supervision alone (Appendix H, Table 10). Stripping age, gender, and PPG statistics moves OPLR's NBI only from 0.148 to 0.155 and TailBias from 6.63 to 7.71, and PPG-only OPLR still has lower Overall, High, and Low RMSE than PPG-only U^2-Net (11.34 / 15.42 / 13.04 vs. 11.94 / 19.61 / 13.27) at less than half its NBI (0.389); giving U^2-Net the identical FiLM conditioning moves its bias only to 0.369 / 16.06 and worsens its Low RMSE to 16.24. Separately, U^2-Net with an ordinal auxiliary head, supervising the ordering without reorganizing the latent space, reaches NBI 0.375, moves TailBias only to 16.27, and worsens Low-regime RMSE from 13.27 to 14.54, even though it improves U^2-Net's Overall RMSE to 11.73: a model can improve the normotensive-dominated aggregate while leaving the regime-dependent drift untouched.

7 Limitations

Residual hypotensive bias. OPLR still overestimates at the hypotensive end of the signed-error profile (Figure 1(a)); normality bias is reduced substantially at the hypertensive tail and only partially at the hypotensive one. **External evidence.** UQ is our only distinct care setting, and at 4 subjects (49 hypotensive segments from 3 subjects) it is a descriptive stress test, not external validation. MIMIC-IV widens the subject pool but not the setting (ICU, same institution as MIMIC-III), and, with no public subject crosswalk between the two releases, we cannot rule out subject overlap with PulseDB's MIMIC-III portion, so it does not provide definitive external validation. Subject-level inference there is also underpowered at 30 subjects: OPLR's lowest error in every regime is significant against four of five baselines Overall but not against LSCT ($p = 0.05$), and no High-regime contrast (18 subjects) reaches significance except against T2T-GAN. We use it because no other public cohort with synchronized PPG and invasive ABP exists outside our training sources; evaluation on more diverse external cohorts is future work. **Fixed SBP anchors.** The 90 and 140 mmHg boundaries are clinical conventions that enter training only as ordinal routing labels; a cohort whose normotensive range sits elsewhere may need different anchors. Regime metrics are stable under large shifts of the reporting boundaries (Section 6), but we did not retrain with different anchors. **Scope.** All results use fixed 10.24-s segments; the study is retrospective and reconstruction-only and makes no deployment claim, which would require streaming inference and prospective device validation. All models were trained once; reported variability is across segments, not seeds.

8 Conclusion

We formalize Normality Bias in long-tailed physiological waveform reconstruction, introduce NBI and TailBias as model-agnostic diagnostics for it, and propose OPLR, a representation-level remedy built on ordered regime geometry and routing-guided prototype–token refinement. Under strict subject-wise evaluation on PulseDB and zero-shot transfer to two additional cohorts, OPLR attains the lowest NBI and TailBias among all compared models and improves rare-regime fidelity while also lowering aggregate error. Neither data/loss-level rebalancing nor ordinal supervision alone reduces the bias on the same backbone, which points to restructuring the latent geometry itself as an effective remedy.

Ethics Statement

This work performs secondary analysis of publicly available, de-identified physiological monitoring datasets (PulseDB, MIMIC-III, VitalDB, UQ Vital Signs, MIMIC-IV) under their respective access conditions; no new human-subject data were collected and no personally identifiable information is used or released. The method is a reconstruction model evaluated retrospectively; we make no claim of clinical deployment readiness, and premature clinical use without prospective and device validation could cause harm. The authors have read and adhere to the ICLR Code of Ethics.

Reproducibility Statement

The model architecture, training objective, and hyperparameters are specified in Section 3 and Appendices B–C (Table 5); datasets, preprocessing, subject-wise splits, and external-cohort construction are described in Appendix B; the definitions of NBI and TailBias and their binning configuration are given in Section 2 and Appendices B and J; the statistical protocol is in Appendix F; and complete metric tables are in Appendix E. All datasets are public. Code will be released upon acceptance.

AI Use Statement

Large language models were used only for language editing, wording, and LaTeX formatting of the manuscript. They were not used for research ideation, method design, experiment design or execution, data analysis, statistical testing, or the generation of any reported result or reference.

References

John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3):R1–R39, 2007.

Jont B. Allen and Lawrence R. Rabiner. A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564, 1977.

Adrian G. Barnett, Jolieke C. Van Der Pols, and Annette J. Dobson. Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, 34(1):215–220, 2005.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pp. 1171–1179, 2015.

Cheng Bian, Xiaoyu Li, Qi Bi, Guangpu Zhu, Jiegeng Lyu, Weile Zhang, Yelei Li, and Zijing Zeng. Constraint latent space matters: an anti-anomalous waveform transformation solution from photoplethysmography to arterial blood pressure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11087–11095, 2024.

Manuel Blanco-Velasco, Fernando Cruz-Roldán, Juan Ignacio Godino-Llorente, Joaquín Blanco-Velasco, Carlos Armiens-Aparicio, and Francisco López-Ferreras. On the use of PRD and CR parameters for ECG compression. *Medical Engineering & Physics*, 27(9):798–802, 2005.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Eoin Brophy, Maarten De Vos, Geraldine Boylan, and Tomas Ward. Estimation of continuous blood pressure from PPG via a federated learning approach. *Sensors*, 21(18):6311, 2021.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. A review on wearable photoplethysmography sensors and their potential future applications in health care. *International Journal of Biosensors & Bioelectronics*, 4(4):195–202, 2018.

Zehua Chen, Yuyang Miao, Liyuan Wang, Luyun Fan, Danilo P. Mandic, and Jun Zhu. Versatile cardiovascular signal generation with a unified diffusion transformer. *Nature Machine Intelligence*, 8(1):6–19, 2026.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019.

Thomas Eiter and Heikki Mannila. Computing discrete Fréchet distance. Technical Report CD-TR 94/64, Christian Doppler Laboratory for Expert Systems, TU Vienna, 1994.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Yu Gong, Greg Mori, and Frederick Tung. RankSim: Ranking similarity regularization for deep imbalanced regression. In *International Conference on Machine Learning (ICML)*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Sicong Huang, Roozbeh Jafari, and Bobak J. Mortazavi. ArterialNet: Reconstructing arterial blood pressure waveform with wearable pulsatile signals, a cohort-aware approach. *IEEE Open Journal of Engineering in Medicine and Biology*, 7:14–19, 2025.

Nabil Ibtehaz, Sakib Mahmud, Muhammad E. H. Chowdhury, Amith Khandakar, Muhammad Salman Khan, Mohamed Arselene Ayari, Anas M. Tahir, and M. Sohel Rahman. PPG2ABP: Translating photoplethysmogram (PPG) signals to arterial blood pressure (ABP) waveforms. *Bioengineering*, 9(11):692, 2022.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning (ICML)*, pp. 4651–4664, 2021.

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.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Mahsa Keramati, Lili Meng, and R. David Evans. ConR: Contrastive regularizer for deep imbalanced regression. In *International Conference on Learning Representations (ICLR)*, 2024.

Kwang Bok Kim and Hyun Jae Baek. Photoplethysmography in wearable devices: a comprehensive review of technological advances, current challenges, and future directions. *Electronics*, 12(13): 2923, 2023.

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.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning (ICML)*, pp. 3744–3753, 2019.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

David Liu, Matthias Görges, and Simon A. Jenkins. University of Queensland vital signs dataset: Development of an accessible repository of anesthesia patient monitoring data for research. *Anesthesia & Analgesia*, 114(3):584–589, 2012.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 11525–11538, 2020.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021.

Franck Meyer, Kyunghoon Hur, and Edward Choi. MD-ViSCo: A unified model for multi-directional vital sign waveform conversion. *IEEE Journal of Biomedical and Health Informatics*, 2026.

Benjamin Moody, Sicheng Hao, Brian Gow, Tom Pollard, Wei Zong, and Roger Mark. MIMIC-IV Waveform Database (version 0.1.0). PhysioNet, 2022. https://doi.org/10.13026/a2mw-f949.

Jiating Pan, Lishi Liang, Yongbo Liang, Qunfeng Tang, Zhencheng Chen, and Jianming Zhu. Robust modelling of arterial blood pressure reconstruction from photoplethysmography. *Scientific Reports*, 14(1):30333, 2024.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, and Shuai Yi. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:4175–4186, 2020.

Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced MSE for imbalanced visual regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

Bernd Saugel and Daniel I. Sessler. Perioperative blood pressure management. *Anesthesiology*, 134 (2):250–261, 2021.

Bernd Saugel, Nick Fletcher, Tong J. Gan, Michael P. W. Grocott, Paul S. Myles, and Daniel I. Sessler. PeriOperative quality initiative (POQI) international consensus statement on perioperative arterial pressure management. *British Journal of Anaesthesia*, 133(2):264–276, 2024.

Bernd Volker Scheer, Azriel Perel, and Ulrich J. Pfeiffer. Clinical review: Complications and risk factors of peripheral arterial catheters used for haemodynamic monitoring in anaesthesia and intensive care medicine. *Critical Care*, 6(3):199–204, 2002.

Fabian Schrumpf, Patrick Frenzel, Christoph Aust, Georg Osterhoff, and Mirco Fuchs. Assessment of non-invasive blood pressure prediction from PPG and rPPG signals using deep learning. *Sensors*, 21(18):6022, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Zhonghe Tian, Aiping Liu, Junxin Chen, Dan Wang, and Xun Chen. PPG-based continuous arterial blood pressure estimation via multi-scale cross attention fusion. *Biomedical Signal Processing and Control*, 113:108833, 2026.

Thomas Unger, Claudio Borghi, Fadi Charchar, Nadia A. Khan, Neil R. Poulter, Dorairaj Prabhakaran, Agustin Ramirez, Markus Schlaich, George S. Stergiou, Maciej Tomaszewski, Richard D. Wainford, Bryan Williams, and Aletta E. Schutte. 2020 international society of hypertension global hypertension practice guidelines. *Hypertension*, 75(6):1334–1357, 2020.

Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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.

E. M. Wesselink, T. H. Kappen, H. M. Torn, A. J. C. Slooter, and W. A. Van Klei. Intraoperative hypotension and the risk of postoperative adverse outcomes: a systematic review. *British Journal of Anaesthesia*, 121(4):706–721, 2018.

Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International Conference on Machine Learning (ICML)*, pp. 11842–11851, 2021.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-N-Contrast: Learning continuous representations for regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Wei Zong, Thomas Heldt, George B. Moody, and Roger G. Mark. An open-source algorithm to detect onset of arterial blood pressure pulses. In *Computers in Cardiology*, pp. 259–262, 2003.

A Related Work

A.1 PPG-to-ABP Waveform Reconstruction

PPG-to-ABP waveform reconstruction has been studied with diverse deep architectures, typically framed as segment-to-segment mapping. PPG2ABP (Ibtehaz et al., 2022) combines a coarse approximation network with a refinement module via stage-wise objectives. U^2-Net (Pan et al., 2024) integrates recurrent and convolutional blocks to capture individualized dynamics. T2T-GAN (Brophy et al., 2021) adapts cycle-consistency for adversarial waveform-to-waveform translation. LSCT (Bian et al., 2024) operates on STFT representations with codebook-style latent modeling and spectrum-wise enhancement. ArterialNet (Huang et al., 2025) emphasizes inter-subject variability via cohort-aware regularization. These methods optimize average-case reconstruction objectives and do not explicitly address BP regime imbalance.

A.2 Prototypes, Routing, and Structured Latent Refinement

Mixture-of-experts and sparse routing dynamically allocate computation to specialized components (Fedus et al., 2022; Shazeer et al., 2017). Prototype and slot-style representations provide compact latent anchors for structured organization (Locatello et al., 2020; Snell et al., 2017), while attention-based set/latent modules use learned anchors to aggregate and redistribute information (Jaegle et al., 2021; Lee et al., 2019). OPLR builds on these ideas but specializes them for ordered BP-regime structure in continuous waveform reconstruction: prototypes serve as regime-aware refinement nodes with an explicit ordinal geometry (Low $<$ Normal $<$ High), rather than as unordered classification anchors or disjoint expert branches.

A.3 Long-Tailed and Imbalanced Continuous Regression

Long-tailed learning is well-studied in classification, including resampling strategies, class-balanced reweighting, logit adjustment, balanced softmax, and decoupled training (Buda et al., 2018; Cui et al., 2019; Menon et al., 2021; Ren et al., 2020). For continuous targets, Yang et al. (2021) introduced Deep Imbalanced Regression with Label Distribution Smoothing (LDS), which we evaluate in Appendix K, and Feature Distribution Smoothing (FDS), which smooths feature statistics across neighboring label bins but imposes no ordinal structure on the latent space. Feature-level regularizers for imbalanced regression have since been proposed: RankSim (Gong et al., 2022), Rank-N-Contrast (Zha et al., 2023), and ConR (Keramati et al., 2024) align feature-space similarity with label order or distance through auxiliary ranking or contrastive losses, and Balanced MSE (Ren et al., 2022) corrects the regression likelihood under label imbalance. These methods regularize a pooled embedding for scalar targets; OPLR instead operates on the token sequence the decoder consumes, using explicit ordered prototypes and routing-guided refinement so that the ordinal structure is written back into the waveform representation rather than imposed only on a global embedding. In the PPG-to-ABP domain, Schrumpf et al. (2021) diagnosed that long-tailed BP distributions degrade scalar BP estimation but proposed no architectural fix. Recent works on PPG-based vital-sign generation pursue orthogonal directions, multimodal cardiovascular signal generation (Chen et al., 2026) and multi-directional waveform conversion (Meyer et al., 2026), rather than addressing regime imbalance. To our knowledge, no prior work in this domain has proposed a representation-level solution to regime imbalance: existing PPG-to-ABP baselines optimize aggregate fidelity, and the long-tailed remedies we evaluate operate at the data/loss level (Yang et al., 2021). OPLR addresses this gap by restructuring the latent geometry of the waveform representation itself.

A.4 Scope of the Approach Beyond PPG-to-ABP

The requirements for OPLR are an ordered target variable, a heavy-tailed distribution whose extremes are the cases that matter, and a sequence output whose local morphology must be preserved. Other physiological waveform tasks meet all three. We state this as scope rather than as a claim, since we validate only on PPG-to-ABP. NBI and TailBias do not depend on OPLR at all: they require only a signed error and a stratifying variable, and apply post hoc to any reconstruction model.

B Architecture and Implementation Details

B.1 Overview. The proposed method consists of a dual-branch time–frequency encoder, the Ordered Prototype Latent Refinement (OPLR) module, and a 1D U-Net decoder. The main paper reports only the core design; here we provide the full implementation details.

B.2 Frequency Encoder. We compute the short-time Fourier transform (STFT) of the PPG signal with $n_{fft} = 256$, hop size $= 64$, Hann window, and center padding enabled. The complex STFT is represented by stacking real and imaginary parts, producing a $2F$-channel sequence where $F = \frac{n_{fft}}{2} + 1$. This sequence is processed by a hierarchical 1D Swin-style encoder (Liu et al., 2021) with channel widths $[8, 16, 32, 64]$, windowed self-attention (window size 16), and patch merging for downsampling. The frequency bottleneck feature is temporally aligned to the temporal bottleneck resolution via 1D interpolation.

B.3 Temporal Encoder. The temporal encoder operates on the PPG waveform using a 1D ResNet-style backbone (He et al., 2016) with channel widths $[8, 16, 32, 64]$ and residual block configuration $[2, 2, 2, 2]$. To improve calibration-free generalization across subjects, we incorporate conditional modulation using two sources of context: (i) subject metadata (age, gender) and (ii) segment-level PPG statistics (mean, std, min, max, range). These features are embedded by lightweight MLPs and fused into a conditioning vector $u \in \mathbb{R}^{64}$, which is applied through feature-wise affine transformation (Perez et al., 2018):

$$\mathrm{Mod}(h; u) = (1 + \gamma(u)) \odot h + \beta(u), \tag{10}$$

where γ and β are learned affine projections initialized to preserve identity ($\gamma = 0$, $\beta = 0$ at initialization).

B.4 Bottleneck Fusion and Tokenization. Let $H_t \in \mathbb{R}^{C_t \times L}$ and $H_f \in \mathbb{R}^{C_f \times L}$ denote the aligned temporal and frequency bottleneck features. We fuse them by concatenation and a 1×1 convolution:

$$H = \phi([H_t; H_f]) \in \mathbb{R}^{C \times L}. \tag{11}$$

The fused representation is then projected into latent tokens:

$$Z = \mathrm{Proj}(H) \in \mathbb{R}^{B \times L \times D}, \tag{12}$$

where $D = 64$ is the latent dimension used by OPLR.

B.5 OPLR Details. OPLR uses $K = 3$ learnable prototypes. The routing temperature τ is scheduled from 1.0 to 0.7 via cosine decay over the course of training. During training, p_{cond} is a scheduled mixture of the predicted routing distribution and the ground-truth one-hot regime label; the teacher-forcing probability is annealed to zero (Bengio et al., 2015), so by the end of training p_{cond} is the model's own predicted routing on every sample and the mixture serves only to stabilize early optimization while routing is still uninformative. The conditioning temperature is set to $\tau_{\mathrm{cond}} = 1.5$. The log-probability bias in routing-guided attention (Eq. 6) is lower-clamped at -2 to prevent low-probability prototypes from being completely masked. Both the Tokens $\rightarrow$ Prototypes and Prototypes $\rightarrow$ Tokens steps use single-layer multi-head attention with 4 heads, followed by LayerNorm and a feed-forward residual block.

B.6 Decoder. The decoder is implemented as a 1D U-Net (Ronneberger et al., 2015). It receives the OPLR-refined token sequence $\hat{Z}$, which is projected back to a bottleneck feature map via a 1×1 convolution. Skip connections from the temporal encoder are modulated by the final conditioning vector, comprising the regime summary Z_{sum} (Eq. 8), subject metadata embedding, and PPG statistics, through FiLM. The decoder output is $\hat{y} \in \mathbb{R}^{1 \times T}$.

B.7 Training. We train with AdamW (Loshchilov & Hutter, 2019) using learning rate $= 1\mathrm{e}{-3}$, weight decay $= 1\mathrm{e}{-4}$, and batch size $= 256$. The learning rate follows a cosine annealing schedule (Loshchilov & Hutter, 2017) with a plateau-based decay factor of 0.5. All segments use fixed length $T = 1{,}280$ samples at $f_s = 125$ Hz. During training, ABP is scaled by a constant $s = 200$ mmHg for numerical stability. All experiments were run on a single NVIDIA Quadro RTX 8000 GPU with 48 GB memory. A full OPLR training run takes approximately 8 hours; baseline and ablation runs were conducted under the same single-GPU setting. Each model was trained once (Section 7).

B.8 Evaluation Metrics. *Waveform metrics.* RMSE: root mean squared error (mmHg). MAE: mean absolute error (mmHg). PC: Pearson correlation coefficient for waveform shape agreement. PRD: percent root-mean-square difference for scale-normalized distortion (Blanco-Velasco et al., 2005). FD: discrete Fréchet distance for shape-sensitive trajectory discrepancy (Eiter & Mannila, 1994). All $\pm$ values are segment-level descriptive standard deviations; they are not error bars on the mean.

Clinical metrics. SBP and DBP are extracted from each reconstructed ABP waveform via beat-level peak/trough detection with segment-level median aggregation (Zong et al., 2003). We report ME (mean signed error, Pred $-$ True); ME close to zero indicates minimal systematic bias in the derived clinical values.

Bias metrics. NBI and TailBias are defined in Section 2. By default they are computed on $M = 10$ fixed-width SBP bins of width 10 mmHg covering $[80, 180]$ mmHg; segments whose true SBP falls outside this range enter all other metrics but not the bin profile. Sensitivity to the binning construction is reported in Appendix J.

B.9 Datasets. *Internal dataset (PulseDB).* We train and primarily evaluate on PulseDB (Wang et al., 2023), a large benchmark for cuff-less BP estimation derived from MIMIC-III (Johnson et al., 2016) and VitalDB (Lee et al., 2022). PulseDB provides standardized 10-second segments at 125 Hz that have already passed signal-quality filtering. For consistency with our model's fixed input length $T = 1{,}280$ (10.24 s at 125 Hz), we zero-pad each 10-s segment from 1,250 to 1,280 samples for training; all reported metrics, for OPLR and for every baseline, are computed on the 1,250 original samples of each segment. We adopt the official subject-wise split, in which train, validation, and test sets contain mutually exclusive subjects. The test set contains 91,843 segments from 278 subjects: Low 2,112 segments from 117 subjects, Normal 73,044 from 277, High 16,687 from 219; subject counts overlap across regimes and do not sum.

External dataset 1 (UQ Vital Signs). The University of Queensland Vital Signs Dataset (Liu et al., 2012) was recorded from 32 patients undergoing surgical procedures under anesthesia. An arterial line was placed in only some of them; 11 cases carry synchronized ABP and PPG. We apply the same pipeline as PulseDB, 10.24-s segments at 125 Hz with identical filtering and quality screening, after which 4 subjects and 2,825 segments remain, of which 49 fall in the Low regime and come from 3 subjects. No UQ data is used at any stage of training, validation, or model selection, and there is no retraining, fine-tuning, or per-subject calibration. UQ metrics are segment-weighted. With 4 subjects, no subject-level test is attempted on UQ, and we read its figures as descriptive.

External dataset 2 (MIMIC-IV). To obtain a larger second cohort, we evaluate zero-shot on MIMIC-IV (Johnson et al., 2023), a later data release from the institution behind MIMIC-III, one of PulseDB's sources; waveforms are taken from the MIMIC-IV Waveform Database, version 0.1.0 (Moody et al., 2022), hosted on PhysioNet (Pollard et al., 2026). MIMIC-III and MIMIC-IV overlap in admission period and no public subject crosswalk links the two releases, so we could not verify that MIMIC-IV subjects are absent from PulseDB's MIMIC-III portion; we therefore treat MIMIC-IV as a partially external cohort (Section 7) and use it because no other public cohort with synchronized PPG and invasive ABP is available outside our training sources. The cohort comprises 30 subjects and 24,022 segments: Low 998 segments from 20 subjects, Normal 17,854 from 28, High 5,170 from 18 (74.3% / 21.5% / 4.2%). Model weights, architecture, and preprocessing are unchanged. MIMIC-IV widens the subject pool and the data release, not the care setting, since it is ICU data.

Table 4: Summary statistics of the internal and external datasets. PulseDB BP statistics refer to the full dataset; UQ and MIMIC-IV statistics refer to the evaluated segments.

Item	PulseDB (internal)	UQ Vital Signs (external)	MIMIC-IV (external)
Number of subjects	2,779 (test: 278)	4	30
Number of segments	835,819 (test: 91,843)	2,825	24,022
Average SBP (mmHg)	121.70	116.03	121.68
SD of SBP (mmHg)	19.09	17.64	20.95
Average DBP (mmHg)	65.19	63.87	61.08
SD of DBP (mmHg)	10.39	9.12	10.14
Test regime distribution (Normal / High / Low)	79.5 / 18.2 / 2.3 %	88.6 / 9.6 / 1.7 %	74.3 / 21.5 / 4.2 %

C Full Objective Formulations

The main paper presents the training objective as three terms (Eq. 9). Here we provide the complete formulations.

C.1 Reconstruction Fidelity ($\mathcal{L}_{\text{recon}}$)

Time-domain loss. We minimize the pointwise MSE between ground-truth and predicted ABP waveforms, scaled by a constant $s = 200$ for numerical stability:

$$\mathcal{L}_{\text{time}} = \left\| \frac{y}{s} - \frac{\hat{y}}{s} \right\|_2^2. \tag{13}$$

Frequency-domain (STFT) loss. Let $S(\cdot)$ denote the stacked real/imaginary STFT representation computed with the same STFT parameters as the frequency encoder (B.2). We minimize:

$$\mathcal{L}_{\text{freq}} = \left\| S\left(\frac{y}{s}\right) - S\left(\frac{\hat{y}}{s}\right) \right\|_2^2. \tag{14}$$

The frequency-domain loss penalizes spectral discrepancies that pointwise time-domain MSE may underweight, such as harmonic structure and envelope consistency. The reconstruction fidelity term is $\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{time}} + \mathcal{L}_{\text{freq}}$.

C.2 Regime-Structure Regularization ($\mathcal{L}_{\text{regime}}$)

$\mathcal{L}_{\text{regime}}$ combines two complementary mechanisms: ordinal routing supervision and distributional usage regularization.

Ordinal routing supervision. Let $c \in \{0, 1, 2\}$ be the ground-truth regime label and $s \in \mathbb{R}^{B \times K}$ the cosine similarities between the pooled global context g and the K prototypes. We subtract a distance-margin proportional to the ordinal distance $|k - c|$ from each logit, then apply focal cross-entropy (Lin et al., 2017):

$$\ell_k = \frac{s_{b,k} - m\,|k - c_b|}{\tau}, \tag{15}$$

$$\mathcal{L}_{\text{ord}} = \text{FocalCE}(\ell, c; \gamma_f). \tag{16}$$

The offset lowers the training logit of each non-target prototype in proportion to its ordinal distance from the true regime and leaves the target logit unchanged, so the routing objective treats the regimes as ordered (Low $<$ Normal $<$ High) rather than as three unordered classes. Because the gradient of the focal cross-entropy with respect to a non-target logit is proportional to that logit's softmax probability (the proportionality factor is shared by all non-target logits of a sample), which the offset shrinks with ordinal distance, the discriminative pressure of the loss falls mainly on the ordinally adjacent prototype, i.e., on the regime boundary that the sample is actually near; removing the offset ($m = 0$) returns NBI to the backbone level (Table 3). The focal term ($\gamma_f = 2$) downweights easy samples and focuses learning on ambiguous regime boundaries. Teacher forcing affects only the conditioning distribution p_{cond} used in the attention bias (Eq. 6); the routing loss is always computed from model-predicted similarities s, ensuring the router learns to infer regime from PPG alone.

Distributional usage regularization. Under severe class imbalance, the router can degenerate to assign all tokens to the dominant Normal prototype. To prevent this, we regularize the batch-mean routing distribution $\bar{p} = \mathbb{E}_B[p]$ toward a fixed target distribution q using CDF-based distributional matching (Peyré & Cuturi, 2019):

$$\mathcal{L}_{\text{dist}} = \frac{1}{K} \sum_{k=1}^{K} \left| \text{CDF}_{\bar{p}}(k) - \text{CDF}_q(k) \right|. \tag{17}$$

We set $q = (q_L, q_N, q_H) = (0.025, 0.806, 0.169)$, the empirical segment-level regime proportions of the PulseDB training split (validation data excluded). This formulation respects the ordinal structure of the regimes: displacing probability mass from Low to High incurs a larger penalty than displacing it to the adjacent Normal, which is consistent with the clinical ordering.

The combined regime-structure loss is $\mathcal{L}_{\text{regime}} = \lambda_{\text{ord}} \mathcal{L}_{\text{ord}} + \lambda_{\text{dist}} \mathcal{L}_{\text{dist}}$.

C.3 Extrema Preservation ($\mathcal{L}_{\text{peak}}$)

To directly penalize errors at the segment-level systolic maximum and diastolic minimum, we define:

$$\mathcal{L}_{\text{peak}} = |\hat{y}_{\max} - y_{\max}| + |\hat{y}_{\min} - y_{\min}|. \tag{18}$$

Since the argmax and argmin operators are non-differentiable, we approximate them using the LogSumExp function (Boyd & Vandenberghe, 2004):

$$\hat{y}_{\max} \approx \tau_p \log \sum_t \exp(\hat{y}(t)/\tau_p), \qquad \hat{y}_{\min} \approx -\tau_p \log \sum_t \exp(-\hat{y}(t)/\tau_p), \tag{19}$$

with temperature $\tau_p = 0.02$. At this temperature, the approximation is sharp enough to closely track the true max/min while maintaining smooth gradients for backpropagation.

C.4 Overall Objective

$$\mathcal{L} = \mathcal{L}_{\text{time}} + \mathcal{L}_{\text{freq}} + \lambda_{\text{ord}}\mathcal{L}_{\text{ord}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{peak}}\mathcal{L}_{\text{peak}}. \tag{20}$$

All hyperparameters are summarized in Table 5.

Table 5: Hyperparameter summary. All values are fixed across experiments and not tuned per dataset.

Hyperparameter	Value	Role
λ_{ord}	0.2	Ordinal routing supervision weight
λ_{dist}	0.1	Distributional usage regularization weight
q	$(0.025, 0.806, 0.169)$	Target distribution for $\mathcal{L}_{\text{dist}}$ (training regime proportions)
λ_{peak}	0.5	Peak-preserving loss weight
m	0.2	Ordinal distance margin
γ_f	2	Focal loss focusing parameter
τ_p	0.02	LogSumExp temperature for peak detection
τ (init $\rightarrow$ final)	$1.0 \rightarrow 0.7$	Routing temperature (cosine decay)
τ_{cond}	1.5	Conditioning temperature

D Routing and Latent-Space Diagnostics

This appendix examines the two internal quantities on which the method rests: the routing distribution over the ordered prototypes and the latent representation that the decoder consumes. Both are read out from the trained model on the PulseDB test set with no additional training.

D.1 Routing Probability Against True SBP

Figure 4 plots the mean routing probability of each prototype as a function of true SBP in 5-mmHg bins, with pointwise 95% subject-bootstrap bands. The three curves vary smoothly over the whole range: the Low and High curves change near-monotonically, the Low prototype's probability falling from about 0.43 below 90 mmHg to about 0.17 above 160 mmHg and the High prototype's rising from about 0.19 to 0.55, while the Normal prototype's probability peaks near 120 mmHg and declines toward both tails. The most probable prototype changes from Low to Normal near 93 mmHg and from Normal to High near 140 mmHg, but no curve jumps at either point, and every prototype keeps a probability between roughly 0.16 and 0.56 at every SBP. Routing is therefore soft and graded rather than a hard three-way switch, which is the behavior the residual, lower-clamped attention bias of Section 3 is designed around, and no prototype is monopolized or abandoned across the range. The mean routing distribution is considerably more even than the target q of $\mathcal{L}_{\text{dist}}$ (Appendix C); with $\lambda_{\text{dist}} = 0.1$, that term acts as a regularizer against collapse rather than as a constraint that routing usage must satisfy, and removing it raises NBI to 0.317 (Table 3).

D.2 Latent Space Visualization

To qualitatively examine whether OPLR encourages ordinal structure on the latent space, we visualize the pooled latent features from the PulseDB test set across all five baselines and OPLR using t-SNE (van der Maaten & Hinton, 2008), colored by ground-truth SBP on a continuous palette (Figure 5). All visualizations use identical t-SNE hyperparameters and the same segment subset for fair comparison.

The five baselines exhibit no clear ordinal organization: low- and high-SBP segments are interleaved throughout the embedding without systematic spatial arrangement. This pattern is consistent across

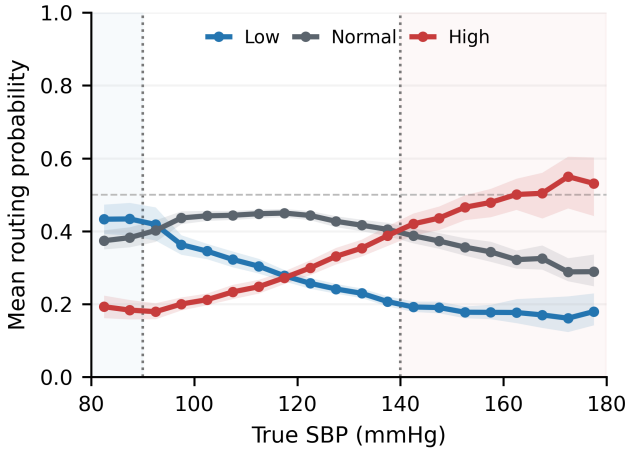

Figure 4: **Routing probability against true SBP** (PulseDB test set). Mean routing probability of the Low, Normal, and High prototypes in 5-mmHg bins of true SBP; shaded bands are pointwise 95% confidence intervals from 1,000 subject-level bootstrap resamples. Dotted lines mark the regime boundaries (90 and 140 mmHg); shaded regions are the Low and High regimes; the dashed line marks 0.5.

architecture families, supervised (PPG2ABP, U^2-Net), adversarial (T2T-GAN), time–frequency latent (LSCT), and cohort-aware (ArterialNet), suggesting that average-case reconstruction objectives may not reliably produce regime-separable latent geometry. OPLR, by contrast, exhibits a visibly smoother gradient from low-SBP (blue) through normal (white) to high-SBP (red) regions, consistent with the ordered regime geometry (Low $<$ Normal $<$ High) imposed by its prototype-centric refinement. The embedding shows a continuous gradient rather than three separated clusters, which is what soft routing predicts: hard assignment at 90 and 140 mmHg would have produced the latter.

E Full Metric Tables

The main paper reports RMSE, MAE, and PC without dispersion (Table 1). Here we provide the complete set of five waveform reconstruction metrics with segment-level descriptive SD ($\pm$) for all models across all evaluation settings on the internal test set and both external cohorts. The Class-stratified setting is a balanced subsample with equal numbers of Low, Normal, and High segments ($3 \times n_{\text{Low}}$, restricted to segments with predictions from all six models: 6,333 segments on PulseDB, 147 on UQ, 2,994 on MIMIC-IV), drawn once and shared by all models.

OPLR has the lowest RMSE and MAE in every setting on all three cohorts; PRD and FD follow the same pattern with three exceptions (PRD and FD in the MIMIC-IV High regime, where LSCT is lower by 0.02 and 0.008, and FD in the PulseDB Low regime, where U^2-Net is lower), and PC is marginally higher for U^2-Net on PulseDB and UQ. Several baselines that perform reasonably on PulseDB degrade sharply under transfer (PPG2ABP from 12.53 to 20.39 and 23.19 Overall RMSE; LSCT from 13.10 to 14.92 and 15.54), whereas OPLR moves from 11.14 to 11.53 and 14.61, suggesting that OPLR, with its full inputs, also transfers more robustly across data sources; given the size of the UQ cohort, we read the UQ figures as descriptive.

F Statistical Analysis

Protocol. Multiple segments from the same subject are not independent experimental units, so we report two levels of inference. At the segment level, per-segment RMSE and MAE of OPLR are compared with each baseline by paired Wilcoxon signed-rank tests; for NBI and TailBias, which are single summary statistics, we form a paired bootstrap of the difference by resampling segments with replacement (2,000 resamples), recomputing both models' diagnostics on each resample, and reading the 95% interval of the difference. These underlie the $*$ markers of Tables 1 and 2, at the

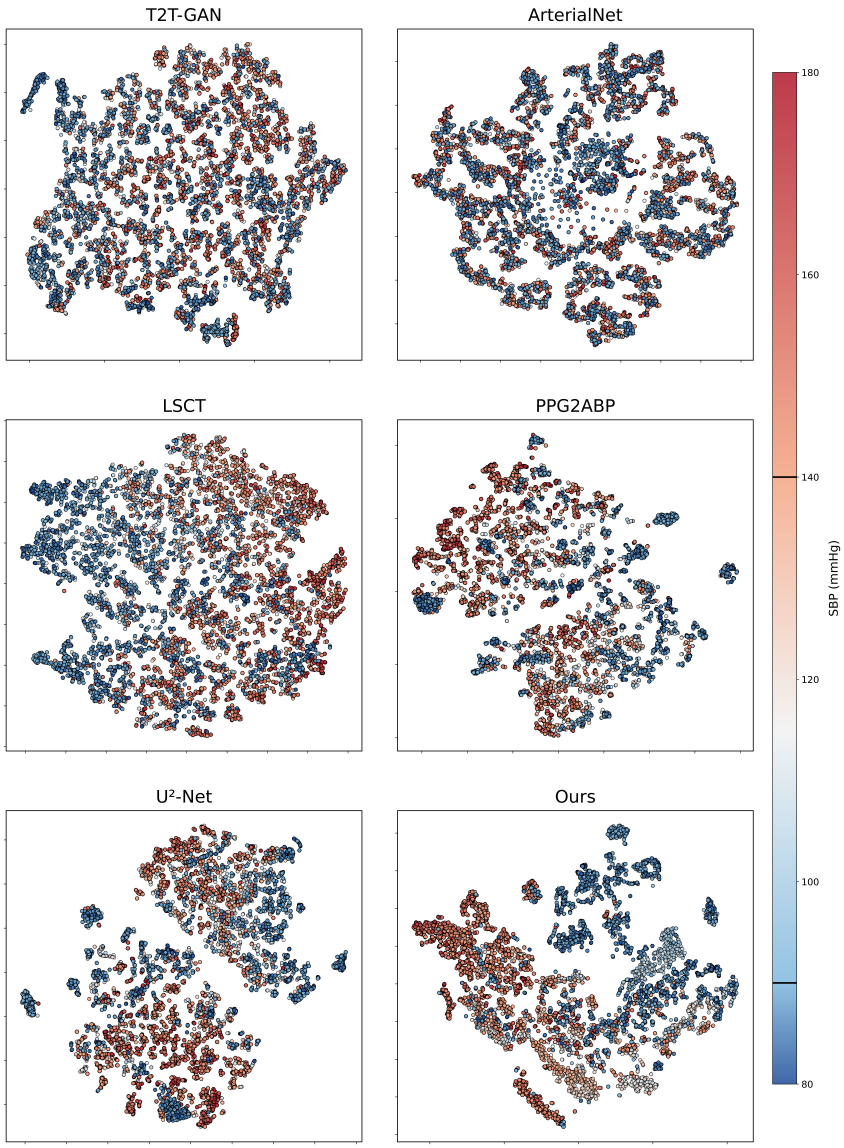

Figure 5: **t-SNE visualization of pooled latent representations** across all baselines and OPLR on the PulseDB test set. Each point corresponds to one test segment, colored by ground-truth SBP (blue = hypotensive, white = normotensive, red = hypertensive; shared color bar, 80–180 mmHg). All panels use identical t-SNE hyperparameters and the same segment subset. Five baselines show no systematic spatial organization by SBP, while OPLR exhibits a visibly more organized ordinal arrangement: low-SBP (blue) and high-SBP (red) regions occupy distinct subregions of the embedding rather than being uniformly intermixed, consistent with its imposed prototype geometry (Low $<$ Normal $<$ High).

conventional $p < 0.05$ (95% interval) level. At the subject level, each subject's mean per-segment RMSE (MAE) within a setting is computed first and the per-subject differences are tested with paired Wilcoxon signed-rank tests; subject resampling gives the corresponding intervals for NBI and TailBias. These underlie the † markers and are reported for PulseDB (278 subjects) and MIMIC-IV (30 subjects); UQ, with 4 subjects, is not tested at the subject level. Segment-pooled and subject-averaged quantities weight subjects differently, and we report both rather than replacing one with the other. Full per-contrast statistics are available with the code release.

Table 6: Internal (PulseDB), all evaluation settings. 91,843 test segments from 278 unseen subjects (79.5% Normal, 18.2% High, 2.3% Low). ± is the segment-level SD.

Method	Setting	RMSE	MAE	PRD (%)	FD	PC
T2T-GAN	Overall	33.22±6.79	27.86±5.92	37.99±5.95	0.214±0.077	0.405
	Class-stratified	34.17±8.21	27.90±7.01	39.96±6.77	0.232±0.075	0.403
	High	42.82±6.20	34.39±5.94	40.54±4.88	0.277±0.092	0.426
	Low	28.50±5.11	22.83±4.41	41.90±8.24	0.220±0.068	0.389
ArterialNet	Overall	14.76±8.21	12.40±7.38	16.76±8.76	0.127±0.066	0.819
	Class-stratified	17.49±8.91	14.87±8.20	20.65±10.42	0.140±0.059	0.794
	High	22.03±9.69	18.54±9.14	20.54±7.95	0.193±0.073	0.812
	Low	17.51±7.76	15.22±7.34	25.79±11.72	0.113±0.049	0.738
LSCT	Overall	13.10±6.97	11.04±6.34	14.98±7.67	0.120±0.070	0.863
	Class-stratified	15.92±7.68	13.63±7.18	19.00±9.54	0.160±0.051	0.842
	High	19.39±8.28	16.22±7.89	18.07±6.69	0.217±0.081	0.860
	Low	16.92±6.89	15.00±6.72	24.95±10.54	0.164±0.024	0.795
PPG2ABP	Overall	12.53±6.16	10.56±5.76	14.36±6.78	0.124±0.059	0.901
	Class-stratified	15.02±6.86	12.94±6.57	18.00±8.62	0.159±0.052	0.891
	High	17.67±7.75	14.92±7.55	16.49±6.33	0.191±0.069	0.907
	Low	16.24±6.01	14.51±5.85	23.93±9.16	0.179±0.044	0.859
U^2-Net	Overall	11.94±7.00	10.16±6.53	13.42±6.94	0.104±0.064	0.920
	Class-stratified	14.30±7.74	12.31±7.35	16.65±8.26	0.123±0.059	0.912
	High	19.61±8.66	16.87±8.57	18.27±7.02	0.175±0.078	0.913
	Low	13.27±6.02	11.59±5.78	19.59±9.19	0.105±0.049	0.895
Ours	Overall	11.14±6.25	9.41±5.77	12.73±6.95	0.097±0.053	0.913
	Class-stratified	12.50±7.88	10.48±7.30	14.90±10.30	0.113±0.062	0.893
	High	14.47±7.30	12.12±6.93	13.63±6.49	0.126±0.062	0.908
	Low	12.89±9.70	10.78±8.99	18.89±14.41	0.126±0.081	0.851

Results. On PulseDB, OPLR is significantly better than all five baselines in every setting and on both metrics at both levels ($p < 0.001$ throughout); the smallest subject-level effect is the Overall contrast with U^2-Net (+0.67 mmHg RMSE, 61% of subjects, $p = 3 \times 10^{-5}$), and the effects are largest in the High regime (+4.03 against U^2-Net, 77% of subjects). On UQ, every segment-level contrast is significant; all but two reach $p < 0.001$, the exceptions being the High-regime RMSE contrast with U^2-Net ($p = 0.03$, 272 segments) and the Low-regime MAE contrast with U^2-Net ($p = 0.001$, 49 segments). On MIMIC-IV, every segment-level contrast reaches $p < 10^{-12}$, and the segment-bootstrap intervals for ΔNBI and ΔTailBias exclude zero against all baselines. At the subject level on MIMIC-IV, the Class-stratified RMSE contrasts and the Overall, Class-stratified, and Low MAE contrasts are significant against all five baselines; the Overall RMSE contrast with LSCT ($p = 0.05$) and the Low RMSE contrast with LSCT ($p = 0.20$) are not; in the High regime (18 subjects) only the contrast with T2T-GAN is; and the subject-resampling intervals for ΔNBI and ΔTailBias include zero against U^2-Net and PPG2ABP. Subject-level significance on an external cohort therefore remains an open item that a larger external subject pool would settle (Section 7).

G Derived BP Signed Bias

SBP and DBP are extracted from each reconstructed waveform (Appendix B, B.8), and ME (Pred − True) summarizes signed bias in the derived clinical values. Table 9 reports Overall SBP/DBP ME for all models on all three cohorts. On PulseDB and MIMIC-IV, OPLR's SBP ME is within 1 mmHg of zero (+0.51 and +0.77) and its DBP ME within 2.2 mmHg, whereas the baselines with the strongest aggregate metrics carry offsets of several mmHg in opposite directions on the two cohorts (U^2-Net: SBP −5.63 on PulseDB and +6.62 on MIMIC-IV; PPG2ABP: −2.18 and +9.54). This is the derived-value counterpart of the profile in Figure 1(a): a model whose reconstructions are pulled toward the normotensive mean acquires a cohort-dependent offset as soon as the cohort's SBP distribution shifts. On the 4-subject UQ cohort every model overestimates SBP; OPLR's offset (+5.71) is about half of U^2-Net's (+10.80), while T2T-GAN and LSCT, whose waveform fidelity is

Table 7: External (UQ Vital Signs), all evaluation settings. 2,825 segments from the 4 subjects retained after the PulseDB preprocessing pipeline (88.6% Normal, 9.6% High, 1.7% Low; the Low regime comprises 49 segments from 3 subjects). Zero-shot evaluation without retraining; segment-weighted; ± is the segment-level SD. Descriptive only.

Method	Setting	RMSE	MAE	PRD (%)	FD	PC
T2T-GAN	Overall	32.12±5.71	27.27±5.45	37.60±6.74	0.174±0.058	0.470
	Class-stratified	35.57±7.37	29.50±5.94	41.88±8.60	0.205±0.058	0.451
	High	43.74±6.15	33.94±6.43	38.58±3.81	0.240±0.075	0.482
	Low	33.48±4.73	29.23±3.89	50.56±7.34	0.210±0.047	0.420
ArterialNet	Overall	16.51±8.02	13.94±7.34	19.66±10.37	0.157±0.059	0.825
	Class-stratified	20.62±9.58	17.62±8.97	25.21±13.18	0.183±0.061	0.766
	High	21.27±10.83	18.42±10.75	18.40±8.07	0.192±0.067	0.848
	Low	25.16±6.70	21.80±6.27	37.97±10.03	0.203±0.059	0.672
LSCT	Overall	14.92±6.55	12.44±6.08	17.34±6.99	0.132±0.053	0.762
	Class-stratified	18.84±7.86	16.04±7.28	22.35±9.45	0.159±0.049	0.734
	High	23.57±10.01	20.70±9.97	20.39±7.29	0.205±0.074	0.831
	Low	20.16±5.25	17.40±4.95	30.42±7.91	0.144±0.025	0.676
PPG2ABP	Overall	20.39±5.08	15.79±4.89	24.30±7.43	0.146±0.045	0.533
	Class-stratified	22.91±6.92	18.96±6.61	28.21±11.34	0.171±0.047	0.570
	High	22.05±10.30	18.86±9.66	19.22±7.84	0.189±0.060	0.697
	Low	27.02±3.63	23.38±3.12	40.79±5.54	0.182±0.041	0.508
U^2-Net	Overall	12.44±5.28	10.44±5.09	14.75±6.43	0.132±0.049	0.922
	Class-stratified	15.64±7.30	13.52±6.91	19.12±10.65	0.151±0.053	0.901
	High	17.65±9.33	15.43±9.14	15.36±7.05	0.170±0.067	0.907
	Low	18.92±7.60	17.02±7.32	28.55±11.49	0.156±0.048	0.886
Ours	Overall	11.53±5.28	9.58±5.06	13.40±5.53	0.125±0.047	0.908
	Class-stratified	14.12±6.13	11.96±5.65	17.21±8.95	0.138±0.047	0.874
	High	16.13±6.27	13.02±5.91	14.26±5.51	0.161±0.060	0.853
	Low	16.65±5.80	15.09±5.26	25.15±8.92	0.135±0.042	0.855

poor or moderate on this cohort, happen to have the smallest mean SBP error, a reminder that ME on a small cohort is a weak criterion on its own. We do not report limits of agreement, which are a device-validation statistic and would invite reading these figures as an accuracy claim.

H Component Controls: Input Parity and Ordinal Supervision

Table 10 collects the component controls on PulseDB that complement Table 3: the additive and subtractive series there ask which component is sufficient, and which is necessary, to move the bias metrics; the controls here ask whether the full model's advantage can be explained by its inputs or by ordinal supervision alone.

Input parity. In Table 1 the baselines use their published PPG-only inputs, whereas the full OPLR model additionally receives age, gender, and segment-level PPG statistics through FiLM conditioning. "OPLR, PPG-only" removes the external inputs but keeps the prototype summary Z_{sum} and the decoder conditioning path through which the ordered latent structure reaches the decoder. Stripping the extra inputs moves OPLR's NBI from 0.148 to 0.155 and TailBias from 6.63 to 7.71, and raises High RMSE from 14.47 to 15.42; PPG-only OPLR nevertheless retains lower Overall, High, and Low RMSE than PPG-only U^2-Net (11.34 / 15.42 / 13.04 vs. 11.94 / 19.61 / 13.27) and less than half its NBI (0.155 vs. 0.389). Giving U^2-Net the identical conditioning barely moves its bias (0.369 / 16.06) and worsens its Low-regime RMSE from 13.27 to 16.24. The extra inputs do not by themselves produce regime-fair reconstruction, and OPLR's advantage is retained without them.

Ordinal supervision without latent restructuring. The simplest version of our idea is to supervise the ordering without reorganizing the latent space around it. We trained the U^2-Net backbone with an ordinal auxiliary head (no prototypes, no routing, no refinement) under the identical protocol.

Table 8: External (MIMIC-IV), all evaluation settings. 24,022 segments from 30 subjects (74.3% Normal, 21.5% High, 4.2% Low; subject-level disjointness from PulseDB could not be verified, Section 7). Zero-shot evaluation without retraining; ± is the segment-level SD.

Method	Setting	RMSE	MAE	PRD (%)	FD	PC
T2T-GAN	Overall	34.32±6.34	28.95±5.71	40.95±7.50	0.346±0.057	0.357
	Class-stratified	34.52±7.32	28.52±6.48	40.89±7.70	0.345±0.058	0.332
	High	41.65±4.70	34.21±4.10	42.48±5.06	0.345±0.060	0.420
	Low	29.31±5.57	23.53±5.04	39.74±9.43	0.345±0.057	0.235
ArterialNet	Overall	17.44±7.61	14.71±7.22	20.79±8.98	0.330±0.069	0.785
	Class-stratified	17.79±7.87	15.00±7.53	21.16±9.18	0.328±0.064	0.780
	High	20.70±8.71	17.48±8.49	20.94±7.97	0.353±0.070	0.846
	Low	16.15±6.80	13.62±6.57	22.05±10.21	0.307±0.056	0.717
LSCT	Overall	15.54±6.76	12.74±6.12	18.48±7.59	0.302±0.040	0.768
	Class-stratified	16.14±6.80	13.29±6.20	19.14±7.41	0.311±0.044	0.768
	High	18.69±8.16	15.12±7.83	18.67±6.48	0.316±0.056	0.867
	Low	14.85±4.85	12.45±4.22	20.32±7.74	0.321±0.043	0.686
PPG2ABP	Overall	23.19±7.59	19.48±7.84	28.03±10.15	0.350±0.059	0.673
	Class-stratified	23.34±7.79	19.79±8.02	28.20±10.52	0.350±0.058	0.675
	High	25.59±8.37	21.59±8.86	26.57±10.28	0.367±0.063	0.737
	Low	21.80±6.98	18.69±7.09	29.76±10.88	0.339±0.052	0.617
U^2-Net	Overall	19.54±8.44	16.86±8.32	23.49±10.69	0.329±0.076	0.766
	Class-stratified	19.87±8.64	17.14±8.61	23.87±10.97	0.331±0.070	0.755
	High	22.42±8.87	19.53±9.05	23.25±10.36	0.356±0.080	0.796
	Low	17.98±8.03	15.26±7.93	24.54±11.73	0.312±0.057	0.698
Ours	Overall	14.61±7.76	11.72±7.46	17.21±8.60	0.293±0.065	0.821
	Class-stratified	15.14±7.89	12.23±7.70	17.80±8.62	0.300±0.061	0.822
	High	18.37±8.59	14.88±9.10	18.69±8.53	0.324±0.067	0.861
	Low	12.90±5.55	10.44±5.08	17.48±7.94	0.292±0.055	0.788

Table 9: Overall derived-BP signed error (ME = Pred − True, mmHg) on the internal PulseDB test set and the two external cohorts (zero-shot). ME closer to zero is better; smallest absolute ME per column in bold.

	PulseDB (internal)		UQ (external)		MIMIC-IV (external)	
Method	SBP ME	DBP ME	SBP ME	DBP ME	SBP ME	DBP ME
T2T-GAN	−4.42	−1.77	**+0.85**	+3.65	**+0.08**	+6.84
ArterialNet	−5.43	**−0.54**	+11.33	+1.67	−1.15	+5.89
LSCT	−4.64	+1.43	+0.94	+2.86	−7.25	+6.26
PPG2ABP	−2.18	+3.38	+3.33	+10.30	+9.54	+13.70
U^2-Net	−5.63	−0.97	+10.80	**+1.64**	+6.62	+9.54
Ours	**+0.51**	+0.57	+5.71	+2.90	+0.77	**+2.17**

The ordinal signal alone does very little to the bias: NBI moves from 0.389 to 0.375, TailBias barely moves, and Low-regime RMSE worsens from 13.27 to 14.54, even though Overall RMSE improves from 11.94 to 11.73. This is the clearest case in our experiments of a model improving the normotensive-dominated aggregate while leaving regime-dependent drift untouched.

I Threshold Sensitivity

The regime definitions used throughout this paper rely on fixed SBP thresholds (Low: <90, High: >140 mmHg) derived from clinical guidelines for hypotension and hypertension (Saugel et al., 2024; Unger et al., 2020). These thresholds enter OPLR only as training-time ordinal labels; inference requires no threshold. To check that the reported regime metrics are not an artifact of where the reporting boundaries are drawn, we re-partition the PulseDB test set under three alternative boundary pairs (High/Low >135/<85, >145/<95, and >150/<100 mmHg) with the trained model and its

Table 10: Component controls on PulseDB. Upper block: input-parity controls. Lower block: ordinal supervision without latent restructuring. RMSE in mmHg.

	Variant	Overall RMSE	High RMSE	Low RMSE	NBI	TailBias
Full model	OPLR (Ours)	11.14	14.47	12.89	0.148	6.63
Input parity	OPLR, PPG-only	11.34	15.42	13.04	0.155	7.71
	U^2-Net, published (PPG-only)	11.94	19.61	13.27	0.389	16.77
	U^2-Net + FiLM conditioning	11.95	17.18	16.24	0.369	16.06
Ordinal sup.	U^2-Net + ordinal head	11.73	18.20	14.54	0.375	16.27

predictions held fixed, and recompute every setting (Table 11). The Overall metrics are unchanged by construction, and NBI and TailBias, which are computed on the continuous SBP profile, are unaffected by the boundaries.

The Low regime ranges from 564 to 10,912 segments and the High regime from 8,102 to 22,823, a nearly 20-fold variation in the Low regime and a 2.8-fold variation in the High regime. Class-stratified RMSE stays within 12.42–12.97 across all configurations. Per-regime RMSE moves smoothly and in the expected direction: High RMSE rises from 13.69 ($>$135) to 15.77 ($>$150) and Low RMSE falls from 15.82 ($<$85) to 11.72 ($<$100) as the boundaries admit progressively more or less extreme segments, with no discontinuity at the 90 and 140 mmHg anchors used for supervision. The $>$135/$<$85 configuration retains only 564 Low segments, the most severe hypotensive cases, for which Low RMSE is 15.82; the broad $<$100 definition includes many near-normotensive segments and correspondingly lower error. These results indicate that OPLR's regime-level behavior is a smooth function of true SBP rather than a property of the particular partition used to report it, consistent with the continuous latent gradient in Figure 1(b). We cannot rule out that a cohort with a different pressure distribution would need the supervision anchors re-set, which would require retraining.

Table 11: Sensitivity of OPLR's reported metrics to the regime boundaries on the PulseDB test set. The trained model and its predictions are fixed; only the High/Low boundaries used to partition the test set change, so the Overall row is identical for every boundary pair and is shown once. The Class-stratified subsample is redrawn for each boundary pair ($3 \times n_{\min}$, without the all-model restriction of Table 1), so its default row is an independent draw from the one in Table 1. n is the number of test segments in each setting (this export differs from Table 1 by at most two segments; the metrics agree to the reported precision); $\pm$ is the segment-level SD over the segments in that setting.

Boundaries (High / Low)	Setting	n	RMSE	MAE	PRD (%)	FD	PC
$>$140 / $<$90 (default)	Overall	91,841	11.14±6.25	9.41±5.77	12.73±6.95	0.097±0.053	0.913
	Class-stratified	6,336	12.58±7.32	10.54±6.80	15.01±9.08	0.114±0.063	0.890
	High	16,686	14.47±7.30	12.12±6.93	13.63±6.49	0.126±0.062	0.908
	Low	2,112	12.89±9.70	10.78±8.99	18.89±14.41	0.126±0.081	0.851
$>$135 / $<$85	Class-stratified	1,692	12.97±7.99	10.94±7.58	16.26±10.33	0.119±0.067	0.887
	High	22,823	13.69±6.98	11.47±6.61	13.20±6.25	0.122±0.059	0.916
	Low	564	15.82±12.07	13.46±11.59	23.97±18.32	0.152±0.096	0.820
$>$145 / $<$95	Class-stratified	16,620	12.42±6.67	10.48±6.25	14.36±7.83	0.109±0.057	0.903
	High	11,897	15.17±7.65	12.76±7.31	13.91±6.61	0.130±0.065	0.903
	Low	5,540	12.07±7.29	10.29±6.80	17.36±10.81	0.110±0.062	0.886
$>$150 / $<$100	Class-stratified	24,306	12.56±6.52	10.63±6.13	14.09±7.30	0.109±0.055	0.907
	High	8,102	15.77±7.92	13.30±7.60	14.10±6.62	0.137±0.067	0.905
	Low	10,912	11.72±6.43	10.08±6.01	16.52±9.44	0.102±0.053	0.896

J Binning Sensitivity of NBI and TailBias

NBI and TailBias are computed by default on 10 fixed-width SBP bins of 10 mmHg over $[80, 180]$ mmHg, and TailBias depends on the two most extreme non-empty bins, which in a long-tailed setting may contain few samples. To test whether the conclusions depend on this construction, we varied it across eleven configurations: bin width 5, 10, and 20 mmHg; range $[80, 180]$ and $[90, 170]$; fixed-width binning and 5- and 10-bin quantile binning; segment-weighted and subject-balanced

weighting; tails defined as the outer one or outer two bins; and a minimum of 50 segments per retained bin (Figure 6). OPLR has the lowest NBI and the lowest TailBias in all eleven configurations. Its NBI stays within 0.15–0.18 and its TailBias within 4.3–9.5, while the best baseline never falls below 0.37 and 10.2; the ordering of the baselines is also largely preserved across configurations. Equal-width binning is the deliberate default, because quantile binning lets the dense normotensive range dominate the fit that NBI is measuring; the quantile and subject-balanced variants are reported to show that the conclusion does not depend on that choice. Sparse tail bins remain the principal weakness of TailBias, which the minimum-count variant only partly addresses; we therefore always report the full signed-error profile (Figure 1) alongside the scalar metrics.

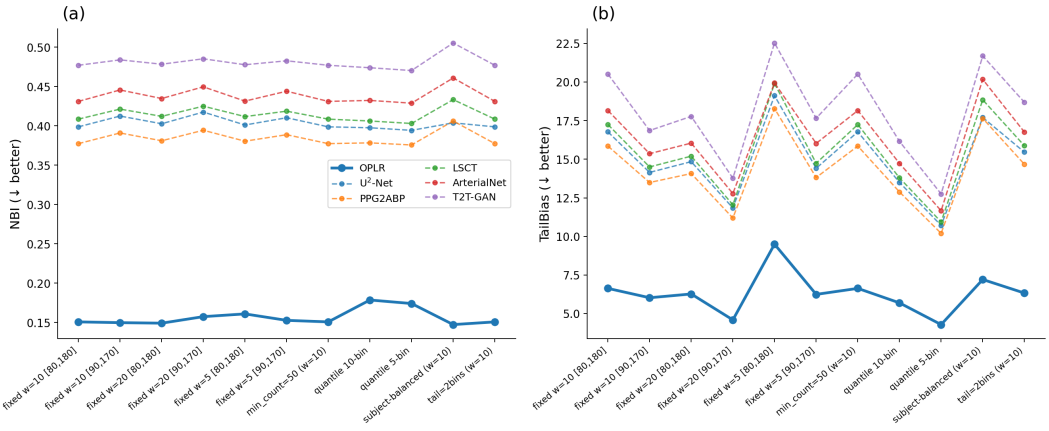

Figure 6: **Binning sensitivity of NBI and TailBias** on the PulseDB test set across eleven binning configurations for all six models: **(a)** NBI, **(b)** TailBias (mmHg). OPLR has the lowest value in every configuration.

K Standard Long-Tailed Fixes

Before developing OPLR, we investigated whether standard long-tailed learning techniques could mitigate normality bias without architectural changes. We applied three established methods to our own time–frequency encoder–decoder backbone (Section 3) with the OPLR module removed but the frequency encoder, decoder conditioning, and $\mathcal{L}_{\text{peak}}$ retained (12.75 Overall RMSE; this differs from the bare backbone of Table 3, which removes all four components): (i) regime-stratified oversampling via a weighted batch sampler (Buda et al., 2018) with inverse-frequency sampling probabilities, so that each batch contains approximately balanced regime proportions; (ii) class-balanced loss reweighting (Cui et al., 2019) that assigns higher weights to High- and Low-regime segments based on inverse frequency; and (iii) Label Distribution Smoothing (Yang et al., 2021), a continuous imbalanced regression technique that smooths the label distribution to rebalance gradient contributions.

Table 12 reports each technique applied individually and in all combinations on the same backbone. Three observations emerge.

First, no individual technique reduces NBI to OPLR's level. The best individual NBI among data/loss-level fixes (loss reweighting: 0.32) remains more than twice OPLR's NBI (0.148), and the best combined fix (Sampler+Reweighting+LDS: 0.28) is still substantially higher.

Second, rare-regime gains are inconsistent and come at a cost. The weighted batch sampler reduces Low-regime RMSE (15.35 $\rightarrow$ 14.20) but worsens High-regime RMSE (18.17 $\rightarrow$ 19.13) and leaves NBI essentially unchanged. Loss reweighting reduces Low-regime error (12.88) at the cost of Overall fidelity. LDS produces marginal gains on isolated metrics but does not address regime-dependent drift.

Third, combinations do not stack. Adding all three techniques together produces $\mathrm{NBI} = 0.28$ but increases Overall RMSE to 14.58: the combined intervention destabilizes the dominant regime without resolving the tails.

By contrast, OPLR (last row) substantially reduces both High-regime RMSE ($18.17 \rightarrow 14.47$) and Low-regime RMSE ($15.35 \rightarrow 12.89$) while reducing NBI by more than half ($0.36 \rightarrow 0.148$), and improves rather than degrades Overall RMSE ($12.75 \rightarrow 11.14$). Data/loss-level fixes change how much each sample counts but do not impose an ordered regime structure on the latent space; OPLR restructures it through ordered prototype routing and regime-aware token refinement.

Table 12: Standard long-tailed fixes applied to our backbone with the OPLR module removed (auxiliary components retained). $\pm$ is the segment-level SD.

Setting	Batch Sampler	Loss Reweighting	LDS	Overall RMSE	High RMSE	Low RMSE	NBI
Backbone				12.75±6.47	18.17±8.01	15.35±5.81	0.36
	✓			12.71±6.70	19.13±8.41	14.20±6.31	0.37
		✓		13.61±7.05	18.91±9.04	12.88±7.88	0.32
			✓	13.63±6.63	17.36±8.31	15.17±7.37	0.34
	✓	✓		13.54±6.88	16.91±8.83	13.35±7.89	0.30
	✓		✓	13.08±6.47	16.56±7.98	14.52±7.27	0.31
		✓	✓	14.26±7.15	17.64±9.10	13.86±9.10	0.30
	✓	✓	✓	14.58±7.24	17.28±8.68	13.76±10.14	0.28
OPLR (Ours)				**11.14±6.25**	**14.47±7.30**	**12.89±9.70**	**0.148**