VitalFM: A Foundation Model of High-Resolution Vital Signs in the ICU

Anonymous authors
Paper under double-blind review

Abstract

ICU monitors record physiology every few seconds, yet many prediction models use hourly representations that may obscure short-term changes. Clinical prediction using handcrafted summaries requires domain expertise, whereas sequence models learning directly from measurements may underperform when cohorts and positive labels are limited. We therefore introduce **VitalFM**, a foundation model that learns representations from dense, partially observed ICU monitoring through self-supervised pre-training, without input imputation. Its bi-axial encoder combines bidirectional state-space modeling over time with masked attention across channels, and pre-training reconstructs an entirely withheld vital sign while preserving its correlations with remaining signals. The source corpus comprises approximately 13.5 million monitoring hours from 79,000 patients, and we evaluate four outcomes in Site A and MIMIC-III. We compare VitalFM against LightGBM, tabular foundation models, and supervised deep sequence models, evaluating both frozen representations and end-to-end fine-tuning. With TabPFN-3 as a fixed classifier, frozen VitalFM representations improve over raw grids by 3.1 AUROC and 2.5 AUPRC points, outperform handcrafted features on Site A, and achieve comparable performance on MIMIC-III. Fine-tuned VitalFM achieves the highest mean AUROC and AUPRC of all methods: slightly above handcrafted features with tabular foundation models, and 3.2 and at least 7.9 AUROC points above handcrafted LightGBM and supervised deep sequence models, respectively. Pre-training accounts for much of this performance: VitalFM w/o pre-training, which uses the identical architecture trained only on the downstream task, is lower by 14.8 AUROC and 8.4 AUPRC points on average. Fine-tuned models also transfer between health systems without destination-site refitting. These results support self-supervised pre-training on dense monitoring for clinical prediction across outcomes and healthcare systems.

1 Introduction

Bedside monitors record vital-sign numerics at second-to-minute intervals (Moody et al., 2020b), yet established clinical prediction pipelines and benchmarks use hourly representations of clinical measurements (Wang et al., 2020; Harutyunyan et al., 2019), which may obscure short-term physiological changes. Our controlled resolution analysis (Appendix C.3) shows that aggregating inputs to hourly resolution reduces discrimination compared with 2-sec inputs, whereas minute-scale aggregation retains most of the predictive value of the 2-second inputs. Dense monitoring records are also partially observed: continuously measured channels can have gaps, and intermittently sampled ones are unavailable between readings. Many modeling pipelines address this missingness through input imputation, which introduces assumptions about unobserved physiology (Che et al., 2018). Several recent critical-care foundation and self-supervised models operate on hourly representations of clinical data (Burger et al., 2026; Yang et al., 2026; Jagd et al., 2025). Existing approaches therefore face a trade-off between temporal resolution and the duration of jointly modeled context: high-frequency monitoring models typically learn from short local windows (Weatherhead et al., 2022), whereas models that jointly encode longer trajectories commonly aggregate observations hourly and rely on imputation (Burger et al., 2026; Yang et al., 2026; Jagd et al., 2025). Learning from day-long, multichannel records at second-to-minute resolution while accommodating heterogeneous observation patterns remains largely unaddressed.

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.

For clinical prediction, physiological monitoring is commonly represented using handcrafted summaries, which provide strong tabular inputs but require domain expertise to design (Harutyunyan et al., 2019; Yèche et al., 2021). Sequence models instead learn representations directly from measurements. However, when both the representation and predictor are learned from a single outcome-specific cohort, performance can deteriorate if labeled patients or positive events are scarce. These limitations motivate pre-training on larger unlabeled monitoring archives and then adapting the representations to clinical outcomes. Such an approach must accommodate long records, heterogeneous observation patterns, and dependencies among simultaneously monitored physiological signals.

We introduce **VitalFM**, a foundation model for learning reusable representations from high-resolution, partially observed, multi-channel ICU monitoring. Its design addresses three challenges of this setting: capturing within-minute physiological variation, accommodating continuous and episodic measurements without imputation, and jointly modeling long-range temporal dependencies and relationships across vital signs. VitalFM encodes the underlying second-scale measurements within per-channel one-minute patches, uses shared dense and sparse patch encoders to accommodate the two measurement processes, and applies a bi-axial backbone that combines observation-aware temporal modeling with cross-vital attention, limiting the influence of poorly observed patches. For self-supervised pre-training, the model reconstructs an entirely withheld vital-sign channel from the remaining signals, exploiting their co-variation under a shared physiological state while preventing interpolation from the target's own observations. Because pointwise reconstruction does not constrain the target's co-variation with the other vitals, a target-conditioned correlation objective aligns each decoded target–context correlation with its observed counterpart.

We pre-train VitalFM on a corpus of approximately 13.5 million monitoring hours from two U.S. academic health systems (Site A and Site B). We evaluate VitalFM on four outcomes in an internal health-system cohort (Site A) and an external cohort (MIMIC-III). We compare it with LightGBM (Ke et al., 2017), the tabular foundation models TabPFN-3 (Grinsztajn et al., 2026) and TabICLv2 (Qu et al., 2026), and supervised sequence models. The tabular classifiers are evaluated using raw grids and handcrafted features. In frozen evaluation with TabPFN-3 held fixed, VitalFM representations improve mean AUROC/AUPRC over raw six-minute grids by 3.1/2.5 points. Compared with handcrafted features using the same classifier, they perform better on Site A (+0.8/+1.1 AUROC/AUPRC) and slightly lower on MIMIC-III ($-1.1/-0.2$). With end-to-end fine-tuning, VitalFM achieves the highest mean AUROC and AUPRC across the evaluated methods. Its mean AUROC exceeds that of handcrafted features with LightGBM by 3.2 points and that of each supervised sequence baseline by at least 7.9 points. Its aggregate performance is similar to that of the strongest handcrafted-feature tabular foundation-model pipeline, with differences of +0.1/+0.7 AUROC/AUPRC. VitalFM w/o pre-training, which uses the identical architecture trained only on the downstream task, scores 14.8 AUROC and 8.4 AUPRC points lower than fine-tuned VitalFM on average, supporting the contribution of pre-training beyond the architecture alone. In separate cross-cohort transfer experiments, fine-tuned VitalFM models achieved slightly higher mean AUROC and AUPRC than all evaluated baselines when applied directly to the other health system without refitting (Section 4.5). Overall, the strongest competitive alternative is a pipeline that relies on manual feature engineering, whereas VitalFM achieves slightly higher mean performance *directly from raw vital-sign sequences* and substantially outperforms the remaining baselines. Together, these findings support large-scale self-supervised pre-training of high-resolution, partially observed multivital monitoring as a path toward reusable clinical representations across outcomes and healthcare systems.

2 Related Work

ICU prediction has traditionally used handcrafted summaries or outcome-specific sequence models (Harutyunyan et al., 2019; Yèche et al., 2021; van de Water et al., 2024), including architectures that model missing observations (Che et al., 2018; DeVries et al., 2025). Self-supervised learning has also been applied to high-frequency bedside-monitor signals using representations of local temporal windows (Weatherhead et al., 2022). More recent critical-care foundation models represent longer clinical trajectories at hourly resolution (Burger et al., 2026; Yang et al., 2026; Jagd et al., 2025); Jagd et al. (2025) adapt the Bi-Axial Transformer (BAT) introduced by DeVries et al. (2025) for self-supervised learning and use missingness indicators, within-stay forward filling, and training-set mean imputation. By contrast, VitalFM combines *high temporal resolution with long context*: it jointly models 24-hour multichannel records on a shared 2-second grid. At this resolution, imputing

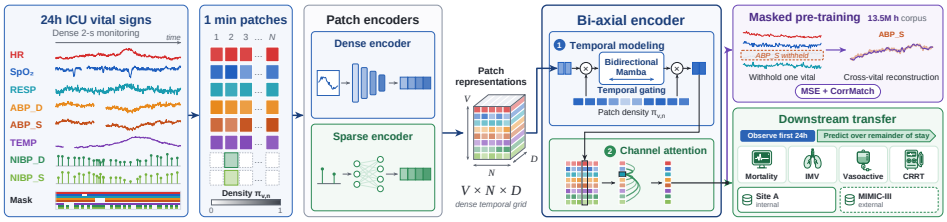


Figure 1: **VitalFM overview.** Per-channel one-minute patches pass through shared dense and sparse patch encoders for continuously monitored signals and episodic NIBP measurements, respectively, followed by a bi-axial backbone for masked pre-training and downstream transfer.

missing inputs would fill long gaps and the intervals between episodic measurements with many artificial values. Instead, VitalFM keeps explicit observation masks, learns representations without imputing inputs, and is designed for reuse across clinical outcomes and health systems.

3 Methods

VitalFM is pre-trained on unlabeled ICU vital-sign data through self-supervised physiological reconstruction and then adapted to clinical prediction using either a frozen encoder or end-to-end fine-tuning. VitalFM comprises four principal components (Figure 1): separate patch encoders for six continuously monitored signals and two episodic non-invasive blood pressure (NIBP) channels; a bi-axial backbone combining bidirectional Mamba over time with attention across vital signs; observation-aware gating of the temporal pathway; and whole-vital masked reconstruction with target-conditioned correlation matching.

3.1 Problem Definition

We represent each 24-hour window of bedside monitoring on a shared 2-second grid (0.5 Hz). The resulting input is a pair $(\mathbf{x}, \mathbf{m})$, where $\mathbf{x} \in \mathbb{R}^{T \times V}$ contains the physiological values and $\mathbf{m} \in \{0,1\}^{T \times V}$ is the corresponding binary observation mask. Here, $T = 43{,}200$ and $V = 8$, and $m_{t,v} = 1$ indicates that channel v is observed at time t. The grid defines a common temporal index but does not imply that every channel is observed at every position.

The eight channels consist of heart rate (HR), peripheral oxygen saturation (SpO_2), respiratory rate (RESP), invasive arterial diastolic and systolic blood pressure (ABP_D and ABP_S), temperature (TEMP), and episodically measured non-invasive diastolic and systolic blood pressure (NIBP_D and NIBP_S). These channels define a shared monitoring schema across the study datasets and capture core cardiorespiratory, temperature, and blood-pressure measurements. We keep invasive and non-invasive pressures as separate channels; NIBP provides blood pressure information for encounters without an arterial line.

3.2 VitalFM Architecture

Dense and sparse patch encoding. We divide each 24-hour window into $N = 1{,}440$ non-overlapping one-minute patches, each spanning 30 positions on the 2-second grid. The six continuously monitored channels use a shared dense patch encoder, while the two episodic NIBP channels use a shared sparse patch encoder. The dense encoder maps each continuous-channel patch to a D-dimensional embedding. For NIBP, the sparse encoder embeds the mean observed value and its mean relative position within the patch; patches without an NIBP measurement receive a learned empty-patch embedding.

Bi-axial temporal and cross-vital modeling. The backbone consists of separate temporal and cross-vital modules (Figure 1) that model temporal dynamics and interactions among vital signs, respectively. The temporal module processes each channel independently across the N one-minute patches using Mamba (Gu & Dao, 2023) in both forward and reverse temporal directions. Vital-sign trajectories contain transient changes, sustained trends, and measurement artifacts, so not every portion of the record is equally informative. Mamba's input-dependent state updates allow the model

to selectively update or preserve its latent state at each one-minute patch, integrating short- and long-range context across the 24-hour record. Because the entire observation window is available at encoding time, bidirectional processing allows each patch to incorporate both preceding and subsequent context. After temporal modeling, the model shifts from the temporal axis to the vital-sign axis: within each one-minute patch, multi-head attention exchanges information among the temporally contextualized representations of the V channels. The patch-availability mask prevents naturally absent channels from contributing as attention keys or values, while a corresponding query gate suppresses the cross-vital update for an absent receiving channel. Thus, each cross-vital update integrates information from available signals within the same patch after their trajectories have been contextualized across the 24-hour record.

Observation-aware temporal gating. The temporal module receives one representation for every one-minute patch, including patches supported by few or no observations. The mask-aware patch encoders determine what each representation contains, but not how much weight it should carry in temporal processing given how much observed data supports it. Without an explicit signal, the temporal model would have to infer this from the representation itself. This is particularly relevant for episodic channels, in which most patches contain no measurement. VitalFM therefore computes patch availability as the fraction of observed samples for continuous channels and as a binary indicator for episodic channels. Before temporal modeling, this signal scales each patch embedding, attenuating inputs with limited support while preserving a weak pathway through empty patches. After Mamba, an output gate combines a learned content-dependent term with the same availability signal and scales the temporal update before it is added to the residual stream. The input and output gates therefore regulate both the information entering temporal modeling and the contextualized update returned at each patch. Full definitions are provided in Appendix B.

3.3 Self-Supervised Pre-Training

VitalFM is pre-trained with two complementary objectives: cross-channel masked reconstruction and target-conditioned cross-vital correlation matching. The combined objective is

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{corr}} \mathcal{L}_{\text{corr}}, \tag{1}$$

where $\mathcal{L}_{\text{recon}}$ is the masked reconstruction loss, $\mathcal{L}_{\text{corr}}$ is the target-conditioned cross-vital correlation loss, and λ_{corr} is a hyperparameter that weights the correlation term.

Cross-channel reconstruction. Because vital signs reflect related, partially coupled physiological processes, the other channels jointly carry information about any single channel withheld. VitalFM exploits this structure through cross-channel masked reconstruction, a variant of masked pre-training (Devlin et al., 2019; He et al., 2022) in which entire channels, rather than random patches, are withheld. During pre-training, all observed values of the selected target channel are withheld across each 24-hour training window and reconstructed from the remaining channels. The encoded representations of the withheld channel's observed patches are replaced by a shared learned mask token $\mathbf{e}_{\text{ssl}}$, while its original observation pattern is retained. The reconstruction loss $\mathcal{L}_{\text{recon}}$ is evaluated only at the originally observed positions of the withheld channel. Withholding the full channel prevents the model from interpolating the target from its own neighboring observations, so reconstruction must rely on the temporal dynamics of the other channels and their relationships to the target. Implementation and objective details are provided in Appendices B.4 and B.5.

Cross-vital correlation matching. A linear decoder produces patch-level outputs for all channels, but reconstruction MSE is applied only to the withheld target. Because the context channels are visible to the encoder, applying MSE to their outputs would introduce an easy self-reconstruction objective that could be satisfied by reproducing information already present at the input. Target-only MSE instead directly supervises recovery of the missing signal, but does not explicitly constrain how it co-varies with the observed vitals. Correlation matching provides this complementary supervision by aligning the correlation between the decoded target and each decoded context channel with the corresponding correlation in the observed signals. For example, when heart rate is withheld, its decoded correlations with the context vitals are matched to their observed counterparts. Thus, MSE supervises recovery of the target values, while correlation matching encourages the decoded target to reproduce its observed pairwise correlations with the context channels. Appendix B.5 provides the objective details.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

Pre-training corpus. The pre-training corpus comprises approximately 13.5 million hours of de-identified bedside-monitor data from Sites A and B, used under institutional approvals. Using the Site A identity-linkage records, we excluded pre-training episodes matched to downstream evaluation patients, and MIMIC-III is reserved exclusively for external evaluation. VitalFM is pre-trained on approximately 589k non-overlapping 24-hour monitoring windows. Additional corpus counts and construction details are provided in Appendix A.2.

Downstream evaluation. We evaluate VitalFM on ICU cohorts from Site A and the MIMIC-III matched-waveform database (Johnson et al., 2016; Moody et al., 2020a), representing internal and external evaluation, respectively. Using the first 24 hours of monitoring, we predict in-hospital mortality and initiation of invasive mechanical ventilation (IMV), vasoactive support, and continuous renal replacement therapy (CRRT). Cohort and task details are provided in Appendix A.4.

Statistical analysis. We computed confidence intervals and two-sided p-values for paired differences (e.g., Figure 2) using a patient-level cluster bootstrap with 10,000 replicates. We report 95% bootstrap percentile intervals and describe differences with $p < 0.05$ as nominally significant; complete procedures are provided in Appendix C.2.

4.2 VITALFM EVALUATION SETTINGS

We evaluate VitalFM through frozen transfer and end-to-end fine-tuning, with randomly initialized controls used to assess the contribution of pre-training.

Frozen transfer. The pre-trained encoder is frozen, its latent representations are mean-pooled over observed patches within each vital sign, and the resulting per-vital representations are concatenated. We evaluate this representation primarily with TabPFN-3 (Grinsztajn et al., 2026) and secondarily with an ℓ_2-regularized logistic regression model, hereafter termed the linear probe. Recent work has used TabPFN to classify embeddings from frozen pre-trained encoders (Zhang et al., 2026). Frozen-transfer and downstream-classifier settings are provided in Appendices C.4 and C.5.1. We also evaluate a joint mean-pooled representation that averages the per-vital representations across observed channels; results are reported in Appendix C.4.1.

Fine-tuning. We initialize the encoder from the pre-trained checkpoint, attach a lightweight head to the vital-pooled representation, and train the full model end-to-end. A lower learning rate for the encoder allows it to adapt to each task and cohort while limiting changes to the pre-trained representation. Implementation details and hyperparameters are given in Appendix C.4.2.

No-pre-training controls. The primary control uses the full VitalFM architecture, randomly initialized and trained end-to-end only on the downstream task. Because a large model without pre-training must learn many parameters from limited task data, we also train smaller randomly initialized variants to check whether the comparison with pre-trained VitalFM depends on model capacity. Full configurations and capacity-sweep results are given in Appendix C.4.3.

4.3 BASELINE MODELS

We compare VitalFM against two baseline families: (1) tabular classifiers applied to flattened raw grids or handcrafted features, and (2) supervised sequence models trained from random initialization on vital-sign sequences.

Tabular models. We evaluate raw vital-sign grids by taking the median within fixed time bins and flattening them into tabular inputs. The primary six-minute grid fits within the feature budgets of all tabular models, allowing each model to receive the complete input. A one-minute grid matches VitalFM's patch granularity and tests whether conclusions depend on temporal aggregation, although its higher dimensionality requires TabPFN-3's built-in feature subsampling.

We then construct a fixed set of handcrafted features from the same eight vital-sign channels available to VitalFM, following the seven-region temporal organization of Harutyunyan et al. (2019): the full observation window together with six nested early- and late-window subsets. Within each region,

Table 1: **Downstream performance.** Five-fold mean ± SD (×100). Best values are blue and bold; second-best values are underlined.

Model	Site A				MIMIC-III			
	Mort.	IMV	Vaso	CRRT	Mort.	IMV	Vaso	CRRT
AUROC								
Raw + TabPFN-3	67.6 ±2.1	63.5 ±0.3	68.9 ±3.0	72.1 ±2.3	74.5 ±4.9	65.4 ±9.5	64.2 ±5.2	70.3 ±11.4
Handcr. + LightGBM	71.3 ±2.1	64.0 ±4.1	71.8 ±1.7	75.2 ±2.7	77.7 ±3.4	62.8 ±5.6	62.1 ±6.4	66.0 ±8.6
Handcr. + TabPFN-3	71.0 ±2.0	66.4 ±2.4	72.5 ±1.8	75.7 ±2.9	79.2 ±3.2	**65.6 ±9.4**	67.0 ±3.7	74.7 ±13.5
Handcr. + TabICLv2	71.1 ±1.8	66.9 ±2.8	73.0 ±2.2	**76.7 ±2.2**	79.6 ±3.9	63.8 ±7.2	**67.4 ±5.4**	77.2 ±9.6
GRU-D	63.9 ±1.9	62.5 ±3.1	66.3 ±2.1	70.0 ±2.3	69.2 ±2.0	58.8 ±5.0	60.5 ±5.2	61.3 ±9.5
TCN	67.9 ±1.8	61.2 ±2.1	63.6 ±2.0	70.3 ±4.1	67.1 ±3.8	57.4 ±7.8	62.1 ±3.6	55.9 ±11.6
Transformer	64.1 ±1.3	60.5 ±2.0	63.5 ±2.0	67.3 ±2.0	66.4 ±2.6	61.0 ±10.0	64.8 ±8.5	58.8 ±13.2
VitalFM w/o pre-training	57.2 ±1.1	56.3 ±3.8	59.0 ±3.7	65.0 ±2.8	56.6 ±9.8	48.8 ±10.6	54.6 ±8.4	60.6 ±8.0
VitalFM (frozen) + TabPFN-3	72.0 ±1.9	67.1 ±3.0	**73.3 ±2.4**	76.6 ±2.9	79.1 ±4.4	61.1 ±6.5	64.9 ±8.0	77.0 ±4.2
VitalFM (fine-tuned)	**73.0 ±1.4**	**67.5 ±4.0**	72.4 ±2.1	75.6 ±4.2	**79.6 ±3.5**	64.6 ±7.8	64.9 ±6.8	**78.4 ±7.5**
AUPRC								
Raw + TabPFN-3	42.1 ±3.8	8.0 ±1.6	14.5 ±2.5	7.3 ±1.4	35.0 ±5.3	10.6 ±3.1	**12.4 ±5.3**	4.2 ±1.8
Handcr. + LightGBM	45.5 ±4.7	8.4 ±2.5	16.2 ±2.3	8.4 ±1.2	36.2 ±7.1	**12.7 ±4.1**	9.2 ±1.9	8.8 ±5.5
Handcr. + TabPFN-3	45.9 ±4.0	8.8 ±1.7	18.0 ±3.6	9.6 ±1.2	39.5 ±6.9	12.0 ±4.8	9.9 ±1.5	6.6 ±3.7
Handcr. + TabICLv2	45.9 ±3.7	8.7 ±1.6	17.8 ±3.7	10.1 ±0.8	39.2 ±7.3	11.4 ±3.7	10.4 ±2.6	7.9 ±5.2
GRU-D	37.5 ±3.3	7.3 ±0.9	15.9 ±3.1	6.4 ±1.1	26.7 ±2.2	11.1 ±2.7	8.4 ±1.8	3.3 ±1.6
TCN	41.7 ±3.7	8.0 ±1.3	12.2 ±1.7	7.2 ±1.5	26.4 ±3.6	7.6 ±2.8	11.9 ±3.1	2.1 ±0.6
Transformer	37.7 ±2.3	7.9 ±1.1	13.1 ±2.3	6.8 ±1.5	26.5 ±4.0	11.0 ±4.5	10.8 ±4.2	2.1 ±0.6
VitalFM w/o pre-training	30.8 ±1.8	6.7 ±1.3	11.0 ±1.7	5.5 ±0.6	17.8 ±1.8	5.5 ±1.1	7.8 ±2.2	4.3 ±3.1
VitalFM (frozen) + TabPFN-3	47.4 ±3.7	9.1 ±1.2	**19.4 ±3.4**	**10.7 ±1.3**	38.1 ±7.2	10.9 ±2.5	8.8 ±2.5	**9.5 ±6.5**
VitalFM (fine-tuned)	**48.5 ±3.7**	**9.9 ±2.7**	17.8 ±2.3	10.2 ±2.2	**39.6 ±6.5**	10.6 ±2.4	11.2 ±2.2	8.9 ±5.0

we compute ten summary statistics per channel. These feature definitions are fixed across cohorts and outcomes. Feature and classifier details are provided in Appendix C.5.1; resolution and raw-grid sensitivities appear in Appendix C.5.2.

Supervised sequence models. We evaluate GRU-D (Che et al., 2018), TCN (Bai et al., 2018), and a Transformer encoder (Vaswani et al., 2017), covering recurrent, convolutional, and attention-based approaches. All are trained from random initialization on vital signs aggregated to one-minute medians with observation masks. This yields 1,440 temporal positions per 24-hour window, matching VitalFM's temporal sequence length while keeping supervised sequence modeling computationally manageable. The baselines use the same eight channels as VitalFM but aggregate within-minute measurements rather than encode them through learned patches. Model configurations and input-resolution details are provided in Appendix C.5.3.

Pre-training ablations. We assess the effects of model size, correlation matching objective, observation-aware temporal gating, bidirectional temporal context, and the temporal backbone. The corresponding variants are VitalFM-S, MSE-only reconstruction, removal of observation-aware temporal gating, unidirectional Mamba, and gated-attention and standard masked Transformer alternatives. We evaluate frozen encoders with TabPFN-3 and the linear probe to assess representation quality without task-specific encoder updates. Full configurations and ablation protocols are provided in Appendix C.6.

4.4 Downstream Prediction Performance

Table 1 summarizes downstream performance across four clinical outcomes in Site A and MIMIC-III, comprising eight cohort-outcome settings. Additional sensitivity analyses are reported in Appendix C.

Raw-grid controls. With TabPFN-3 fixed, frozen VitalFM improves over the six-minute raw grid by 3.1/2.5 AUROC/AUPRC points (higher in 7/8 cells for both metrics) and over the patch-aligned one-minute grid by 3.7/2.7 points, so the comparison is insensitive to grid resolution (Appendix C.5.2).

Handcrafted-feature baselines. With the same TabPFN-3 classifier, handcrafted features and frozen VitalFM reach mean AUROC/AUPRC of 71.5/18.8 and 71.4/19.2, respectively; VitalFM is numerically higher in all eight comparisons at Site A, but in only two of eight comparisons on MIMIC-III. Frozen VitalFM exceeds handcrafted LightGBM by 2.5/1.1 AUROC/AUPRC points (higher in 7/8 and 6/8 cells, respectively) and is comparable on average to the stronger tabular foundation-model pipeline (ΔAUROC -0.8; ΔAUPRC $+0.2$).

Supervised sequence baselines. Fine-tuned VitalFM has numerically higher AUROC than each supervised sequence baseline in all eight cohort-outcome settings. Its mean AUROC exceeds that of

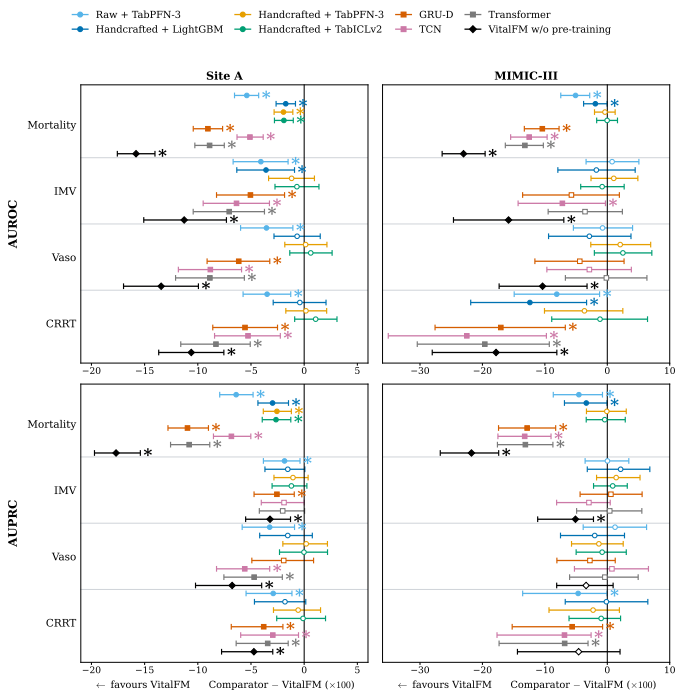


Figure 2: **Paired comparison with Table 1 baselines.** Comparator − fine-tuned VitalFM, $\times 100$; left of zero favors VitalFM. Bars are 95% bootstrap percentile paired patient-bootstrap confidence intervals; asterisks mark nominal $p < 0.05$. Per-setting estimates and p-values: Table C.4.

GRU-D, the sequence baseline with the highest mean AUROC, by 7.9 points. Fine-tuned VitalFM also exceeds the best-performing sequence baseline within each setting in six of eight AUPRC comparisons. These findings are consistent with an advantage from pre-training on the larger unlabeled monitoring corpus before task-specific adaptation in these label-limited cohorts.

Architecture-matched control. The 31.1M-parameter VitalFM w/o pre-training trails fine-tuned VitalFM by 14.8/8.4 AUROC/AUPRC points on average across the eight cohort-outcome settings. In a capacity sweep, the 1.0M-parameter model performs best among randomly initialized models, with performance declining as width increases, a pattern consistent with lower-capacity models being easier to train from limited labeled data. Together, these controls support pre-training, rather than model size alone, as an important contributor to VitalFM's advantage within the evaluated model family (Appendix C.4.3).

NIBP sensitivity. In VitalFM-S w/o pre-training (1.0M parameters), excluding episodic NIBP reduces mean AUROC/AUPRC by 1.5/0.8 points. The dedicated NIBP encoder yields higher AUROC in six of eight settings, while two larger IMV differences favor the shared encoder and AUPRC results are mixed, resulting in similar aggregate performance overall (Table C.6).

External-site adaptation. On MIMIC-III, fine-tuning strengthens VitalFM's performance relative to handcrafted features. With the encoder frozen, VitalFM is numerically higher than handcrafted features with TabPFN-3 in 2 of 8 outcome–metric comparisons (mean differences: $-1.1/-0.2$ AUROC/AUPRC points); after end-to-end fine-tuning on the labeled MIMIC-III training folds, it is higher in 5 of 8 (mean differences: $+0.2/+0.6$), indicating that site-specific adaptation improves its relative performance.

Pooling sensitivity. Frozen VitalFM with TabPFN-3 performs better using concatenated per-vital representations than after averaging representations across channels (71.4/19.2 versus 71.0/18.3 mean AUROC/AUPRC), suggesting that preserving channel-specific information benefits downstream prediction (Appendix C.4.1).

Pairwise comparisons. Figure 2 shows paired contrasts. Fine-tuned VitalFM is numerically higher in 13 of 16 comparisons against raw grids (12 nominally significant), 45 of 48 against supervised se-

Table 2: **Cross-cohort transfer.** Mean ± sample SD (×100) across five source-fold models; no destination-cohort refitting. Best is **blue bold**; second best is underlined.

Model	Site A → MIMIC-III				MIMIC-III → Site A			
	Mort.	IMV	Vaso	CRRT	Mort.	IMV	Vaso	CRRT
AUROC								
Handcr. + TabPFN-3	72.2 ±0.5	63.3 ±1.1	65.6 ±1.8	75.8 ±1.1	65.9 ±0.8	**60.5 ±1.4**	64.5 ±1.2	69.8 ±2.7
Handcr. + TabICLv2	73.4 ±0.3	63.7 ±2.2	**66.3 ±1.3**	75.0 ±0.7	66.7 ±0.7	59.2 ±1.3	65.3 ±1.2	70.1 ±3.1
Handcr. + LightGBM	72.3 ±0.7	59.4 ±3.4	63.2 ±2.8	72.8 ±2.9	66.3 ±1.0	59.5 ±1.0	59.1 ±2.2	64.8 ±2.1
VitalFM (fine-tuned)	**73.8 ±2.1**	**64.1 ±3.1**	64.6 ±2.4	**77.6 ±1.9**	**66.8 ±0.8**	60.4 ±1.9	**65.5 ±0.9**	**71.1 ±4.8**
AUPRC								
Handcr. + TabPFN-3	32.0 ±1.5	**9.3 ±1.0**	9.1 ±1.4	**7.1 ±2.5**	40.4 ±1.2	**7.5 ±0.3**	**13.3 ±0.6**	6.5 ±0.9
Handcr. + TabICLv2	33.1 ±0.9	8.5 ±0.4	9.4 ±0.9	6.1 ±1.0	40.9 ±0.8	6.9 ±0.5	12.7 ±0.6	6.9 ±1.0
Handcr. + LightGBM	30.7 ±1.1	7.2 ±0.8	8.5 ±1.0	5.5 ±1.9	39.6 ±1.0	6.5 ±0.2	10.0 ±1.2	5.6 ±0.6
VitalFM (fine-tuned)	**34.4 ±2.2**	9.0 ±1.1	**9.7 ±1.8**	6.0 ±1.2	**41.5 ±0.9**	7.0 ±0.9	13.1 ±1.1	**7.9 ±1.6**

quence models (34 significant), and all 16 against VitalFM w/o pre-training (14 significant). Against handcrafted-feature pipelines, it is numerically higher in 15 of 16 comparisons with LightGBM (6 significant), 10 of 16 with TabPFN-3 (2 significant), and 12 of 16 with TabICLv2 (2 significant). *No nominally significant comparison favors a comparator.* Full estimates and p-values are reported in Table C.4; confidence intervals are shown in Figure 2.

4.5 Direct Cross-Cohort Transfer

We evaluate cross-cohort transfer against the three strongest non-VitalFM baselines by aggregate performance in Table 1, all of which use handcrafted features. Each source-trained model is applied directly to the other health system without destination-cohort refitting. Table 2 summarizes the five source-fold evaluations as mean ± SD. VitalFM has the highest mean in 10/16 reported comparisons and is numerically higher than TabICLv2, TabPFN-3, and LightGBM in 14/16, 10/16, and 16/16 comparisons, respectively. Small margins against the strongest comparator support comparable cross-site performance with a slight numerical advantage rather than uniform superiority.

4.6 Pre-training Ablations

Table 3 summarizes frozen-transfer effects of temporal design, pre-training objective, and model capacity across the eight cohort–outcome settings in Table 1. To test whether reconstruction quality reflects downstream performance, we compare reconstruction RMSE and PCC with frozen TabPFN-3 performance across the seven configurations. The measures are directionally associated but do not consistently track one another (Figure C.2), so we treat reconstruction metrics as complementary to downstream evaluation. Configurations and per-setting results appear in Appendix C.6.

Table 3: **Pre-training ablations.** Performance averaged across eight cohort–outcome settings (×100). Best values are blue and bold; second-best values are underlined.

Configuration	Model size	TabPFN-3		Linear probe	
		AUROC	AUPRC	AUROC	AUPRC
VitalFM (MSE+CorrMatch)	31.1M	**71.4**	**19.2**	**69.9**	18.3
MSE only	31.1M	70.4	18.1	68.1	17.0
No observation-aware gating	29.9M	71.3	18.9	69.3	17.8
Unidirectional Mamba	23.6M	70.3	18.7	69.3	**18.5**
Gated-attention backbone	30.2M	69.0	17.4	67.3	16.9
Standard masked Transformer	29.0M	69.3	17.0	68.0	17.0
VitalFM-S	1.0M	70.5	18.7	68.0	18.0

Temporal backbone. The bidirectional Mamba backbone achieves stronger aggregate frozen-transfer performance than either approximately parameter-matched attention alternative. With TabPFN-3, the gated-attention and masked-Transformer backbones are lower by 2.3/1.8 and 2.1/2.2 AUROC/AUPRC points, respectively; with the linear probe, the differences are 2.6/1.4 and 1.9/1.3 points. Compared with the smaller unidirectional Mamba (23.6M versus 31.1M parameters), bidirectional Mamba achieves mean AUROC/AUPRC values 1.1/0.6 points higher with TabPFN-3, while the corresponding linear-probe differences are $+0.5/-0.1$ points; the numerical advantage is therefore clearer with the TabPFN-3 readout. These ablations support the Mamba-based temporal backbone, while the directionality comparison is readout-dependent.

Observation-aware temporal gating. Removing observation-aware temporal gating increases normalized reconstruction RMSE and reduces reconstruction PCC on held-out pre-training windows (Appendix C.6.1), although its downstream effects are smaller. Compared with the otherwise matched model without observation-aware temporal gating, VitalFM improves mean AUROC/AUPRC by $0.1/0.3$ points with TabPFN-3 (higher in 5/8 and 7/8 settings) and by $0.6/0.6$ points with the linear probe (4/8 and 5/8). With TabPFN-3, these gains are partly offset by lower Site A vasoactive-support AUROC and lower MIMIC-III IMV-initiation AUROC and vasoactive-support AUROC/AUPRC. Overall, the gating pathway improves reconstruction and provides a modest aggregate downstream benefit (Tables C.13 and C.14).

Correlation matching. With MSE-only pre-training, held-out correlation mismatch increases overall from 0.242 at epoch 1 to 0.427 at the final checkpoint; adding correlation matching reduces the final mismatch to 0.045 while leaving normalized reconstruction MSE nearly unchanged (0.0054 versus 0.0056; Figure 3). Thus, correlation matching improves agreement between decoded and ground-truth target-context correlations without materially increasing target reconstruction error. Relative to MSE-only pre-training, the full objective improves mean AUROC/AUPRC by 1.0/1.1 points with TabPFN-3 and 1.8/1.4 with the linear probe, supporting correlation matching as a component of VitalFM's pre-training objective. Appendix C.6.2 describes the validation protocol.

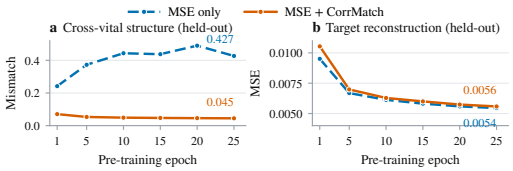


Figure 3: **Correlation matching.** Matched held-out validation trajectories; lower is better and final values are annotated.

Model size. We also evaluate pre-trained VitalFM-S (1.0M parameters), the size that performs best in the no-pre-training sweep. This addresses two questions: whether pre-training still helps at low capacity and whether a compact model can approach VitalFM's (31.1M parameters) performance. VitalFM-S retains most of VitalFM's frozen-transfer performance, achieving mean AUROC/AUPRC of 70.5/18.7 versus 71.4/19.2 for the matched VitalFM reference with TabPFN-3, and 68.0/18.0 versus 69.9/18.3 with the linear probe. Despite remaining frozen during downstream evaluation, the pre-trained VitalFM-S substantially outperforms VitalFM-S w/o pre-training, the same architecture trained end to end from random initialization (62.4/13.3 AUROC/AUPRC), providing evidence that pre-training is beneficial even at low capacity.

5 Limitations

This retrospective study spans three U.S. academic medical centers; generalizability to other geographic and demographic settings and prospective performance remain untested. VitalFM models eight vital-sign channels over 24-hour windows and was evaluated on four binary ICU outcomes; its utility with other clinical modalities, prediction horizons, and downstream tasks remains to be established. Because every pre-training ablation requires full-corpus training, independently tuning each variant or exhaustively searching model width and depth was impractical; we therefore compared targeted configurations under a shared protocol, enabling controlled comparisons without establishing each alternative's best achievable performance.

6 Conclusion

These results show that self-supervised pre-training can extract clinically useful representations from dense, partially observed multivital recordings spanning 24 hours at 2-second resolution. Performance gains over training from scratch, together with the correlation-matching and temporal-backbone ablations, support the contributions of pre-training, correlation matching, and Mamba-based temporal modeling. Cross-cohort transfer further demonstrates reuse across health systems without destination-site refitting, while variation across outcomes and cohorts motivates broader multi-institutional and prospective evaluation.

Ethics Statement

Site A and Site B bedside-monitor data were de-identified and used under institutional approvals. MIMIC-III is a publicly available database requiring credentialed access (Johnson et al., 2016). No new data were collected, and all analyses were retrospective. VitalFM is a research model and was not developed or evaluated as a clinical decision-making system.

Reproducibility Statement

The main text describes the model architecture, pre-training objectives, and evaluation design. The appendix provides cohort and task definitions, preprocessing procedures, model configurations, optimization settings, patient-disjoint downstream partitions, baseline implementations, and statistical analysis procedures. Detailed per-task results and sensitivity analyses are also reported in the appendix. Code for pre-training, fine-tuning, and evaluation will be released upon publication. MIMIC-III is available to credentialed users; Site A and Site B data cannot be publicly shared under institutional data-use agreements.

AI Use Statement

Generative AI tools were used for literature discovery, language editing, manuscript organization, and assistance with checking the consistency of reported calculations, tables, figures, and citations. The authors reviewed and verified all AI-assisted content and take responsibility for the final text, technical claims, analyses, and reported results.

References

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Manuel Burger, Daphné Chopard, Gregor Lichtner, Malte Londschien, Fedor Sergeev, Moritz Fuchs, Hugo Yèche, Rita Kuznetsova, Martin Faltys, Eike Gerdes, Polina Leshetkina, Micha Christ, Moritz Schanz, Nora Göbel, Peter Bühlmann, Elias Grünewald, Felix Balzer, and Gunnar Rätsch. A foundation model for intensive care: unlocking generalization across tasks and domains at scale. medRxiv preprint 2025.07.25.25331635, 2026. URL https://www.medrxiv.org/content/10.1101/2025.07.25.25331635v2.

Finneas J. R. Catling and Anthony H. Wolff. Temporal convolutional networks allow early prediction of events in critical care. *Journal of the American Medical Informatics Association*, 27(3):355–365, 2020. doi: 10.1093/jamia/ocz205.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

Rachael DeVries, Casper Christensen, Marie Lisandra Zepeda Mendoza, and Ole Winther. Biaxial transformers: addressing the increasing complexity of EHR classification. arXiv preprint arXiv:2508.12418, 2025.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. TabPFN-3: Technical report. arXiv preprint arXiv:2605.13986, 2026. URL https://arxiv.org/abs/2605.13986.

Albert Gu and Tri Dao. Mamba: linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6:96, 2019. doi: 10.1038/s41597-019-0103-9. URL https://doi.org/10.1038/s41597-019-0103-9.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Katja Naasunnguaq Jagd, Rachael DeVries, and Ole Winther. Towards self-supervised foundation models for critical care time series. In *NeurIPS 2025 Workshop on Learning from Time Series for Health*, 2025. URL https://openreview.net/forum?id=w1hriS6A90.

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.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Benjamin Moody, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. MIMIC-III Waveform Database Matched Subset. *PhysioNet*, 2020a. doi: 10.13026/c2294b. URL https://physionet.org/content/mimic3wdb-matched/1.0/. Version 1.0.

Benjamin Moody, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. MIMIC-III Waveform Database. *PhysioNet*, April 2020b. doi: 10.13026/c2607m. URL https://doi.org/10.13026/c2607m. Version 1.0.

Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83:112–134, 2018.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. arXiv preprint arXiv:2602.11139, 2026. URL https://arxiv.org/abs/2602.11139.

Robin van de Water, Hendrik Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. Yet another ICU benchmark: A flexible multi-center framework for clinical ML. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ox2ATRM90I.

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*, volume 30, 2017.

Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. MIMIC-Extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 222–235, 2020. doi: 10.1145/3368555.3384469. URL https://doi.org/10.1145/3368555.3384469.

Addison Weatherhead, Robert Greer, Michael-Alice Moga, Mjaye Mazwi, Danny Eytan, Anna Goldenberg, and Sana Tonekaboni. Learning unsupervised representations for ICU timeseries. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 152–168, 2022. URL https://proceedings.mlr.press/v174/weatherhead22a.html.

Yixuan Yang, Mehak Arora, Ryan Zhang, Baraa Abed, Junseob Kim, Tilendra Choudhary, Md Hassanuzzaman, Kevin Zhu, Ayman Ali, Chengkun Yang, Alasdair Edward Gent, Victor Moas, and Rishikesan Kamaleswaran. Clin-JEPA: a multi-phase co-training framework for joint-embedding predictive pretraining on EHR patient trajectories. arXiv preprint arXiv:2605.10840, 2026. URL https://arxiv.org/abs/2605.10840.

Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. HiRID-ICU-Benchmark: a comprehensive machine learning benchmark on high-resolution ICU data. In *NeurIPS Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=SnC9rUeqiqd.

Jingxiang Zhang, Lujia Zhong, Zijie Zhu, Shuo Huang, and Yuang Xu. TabPFN beyond tabular data: calibration and accuracy on multimodal embeddings. *arXiv preprint arXiv:2607.11007*, 2026.

Appendix

VitalFM: A Foundation Model of High-Resolution Vital Signs in the ICU

Contents

A Data and cohort construction . . . 14
A.1 Channel schema and preprocessing . . . 14
A.2 Pre-training corpus . . . 14
A.3 Pre-training window selection . . . 14
A.4 Downstream evaluation cohorts . . . 15
B Model and training details . . . 16
B.1 Encoder and transfer data flow . . . 16
B.2 Dense and sparse patch encoders . . . 16
B.3 Bi-axial block details . . . 16
B.4 Pre-training masking and reconstruction . . . 18
B.5 Objective details . . . 18
B.6 Optimization and configuration . . . 19
C Downstream evaluation and additional results . . . 19
C.1 Downstream evaluation . . . 19
C.2 Statistical analysis . . . 19
C.3 Temporal-resolution sensitivity . . . 20
C.4 VitalFM evaluation settings and controls . . . 22
C.4.1 Frozen VitalFM transfer . . . 22
C.4.2 Fine-tuned downstream contrasts . . . 23
C.4.3 Matched no-pre-training controls . . . 23
C.5 Baseline models . . . 24
C.5.1 Tabular models and representations . . . 24
C.5.2 Raw vital-sign grids . . . 25
C.5.3 Supervised sequence models . . . 26
C.6 Pre-training ablation configurations and results . . . 27
C.6.1 Matched full-validation reconstruction rescore . . . 29
C.6.2 Validation evaluation of the correlation-matching objective . . . 30

A Data and cohort construction

Both pre-training cohorts and both downstream evaluation cohorts use the same preprocessing protocol, applied to raw bedside-monitor streams before model-specific input construction.

A.1 Channel schema and preprocessing

We define eight canonical vital-sign channels in fixed order:

Table A.1: **Canonical vital-sign channels and preprocessing bounds.** Out-of-range measurements are excluded and their positions marked as missing.

Channel	Type	Unit	Valid range
HR	Continuous	bpm	20–250
SpO_2	Continuous	%	60–100
RESP	Continuous	breaths min^{-1}	1–60
ABP_D	Continuous	mmHg	10–150
ABP_S	Continuous	mmHg	20–300
TEMP	Continuous	°C	28–45
NIBP_D	Episodic	mmHg	10–150
NIBP_S	Episodic	mmHg	20–300

Temporal alignment and representation. Each monitoring episode is mapped to a uniform 0.5-Hz grid shared across all vital-sign channels; no interpolation or value-based imputation is applied. Observed values are normalized per channel to $[0, 1]$ using the fixed channel-specific bounds in Table A.1. Unobserved grid positions use zero-valued placeholders paired with an observation mask of zero, allowing the model to distinguish missing positions from observed values.

Artifact filtering. Preprocessing is intentionally conservative. Hardware sentinel values (e.g., unsigned 16-bit maxima encoding signal loss) are removed; physiologically implausible values are set to missing according to Table A.1; and carried-forward NIBP values are removed to preserve the episodic nature of cuff measurements. Carried-forward NIBP values are a documentation convention, not measurements; treating them as observations would lead the model to learn trivial persistence and spurious cross-channel associations.

Blood-pressure channels. Invasive arterial blood pressure and episodic NIBP are retained as separate channels because they have distinct acquisition processes. Arterial pressure is continuously monitored when an arterial line is present, whereas NIBP is typically acquired by an intermittent cuff at intervals of approximately 5–60 minutes and may provide the available blood-pressure measurement when an arterial line is absent. Merging the two would conflate measurement sources, observation densities, and value distributions. VitalFM therefore processes NIBP with a dedicated sparse patch encoder while retaining NIBP as separate streams alongside the continuously monitored signals.

A.2 Pre-training corpus

The pre-training corpus combines bedside-monitor recordings from Site A and Site B; MIMIC-III is reserved for downstream evaluation. A monitoring episode denotes one contiguous bedside-monitor record. The retained corpus comprises 119,560 Site A episodes and 33,609 Site B episodes (153,169 total), representing 57,897 and 21,437 patients, respectively (79,334 total). These episodes span approximately 13.52 million recording hours, including unobserved intervals within records but excluding padding. Pre-training episodes matched to downstream evaluation patients through the Site A identity-linkage records were excluded. The 97/3% training/validation split was performed at the source-record group level rather than the window level: all windows from an episode remained in one partition, and episodes sharing a patient identifier were assigned together.

A.3 Pre-training window selection

Pre-training uses non-overlapping 24-hour windows starting from the beginning of each retained monitoring episode, with shorter terminal segments right-padded and padded positions excluded

Table A.2: **Downstream evaluation cohorts.** Age (years) and weight (kg) are reported as median [IQR]. Task-specific counts show eligible encounters and positive cases, with prevalence calculated among eligible encounters.

Cohort	Encounters (patients)	Age	Female (%)	Weight
Site A	9,035 (7,741)	62 [49–72]	48.1	78 [65–95]
MIMIC-III	2,742 (2,619)	67 [55–79]	44.3	79 [66–93]
Task	**Mortality**	**IMV init.**	**Vasoactive support init.**	**CRRT init.**
Site A, eligible	8,138	6,503	5,850	8,493
positives (%)	1,989 (24.4)	305 (4.7)	394 (6.7)	238 (2.8)
MIMIC-III, eligible	2,697	1,701	2,054	2,700
positives (%)	351 (13.0)	84 (4.9)	99 (4.8)	35 (1.3)

from reconstruction loss. Windows are retained when the mean observed fraction across the six continuous channels is at least 30%; episodic NIBP channels do not contribute to this criterion. This procedure yielded 607,226 windows, comprising 589,201 training and 18,025 validation windows. The reported 13.52 million source recording hours sum the retained episode durations before window admission and exclude padding.

A.4 Downstream evaluation cohorts

We evaluate four binary endpoints—in-hospital mortality, invasive mechanical ventilation (IMV) initiation, vasoactive support initiation, and continuous renal replacement therapy (CRRT) initiation—in Site A and MIMIC-III. Included encounters have at least 24 hours of monitoring span. Models observe the first 24 hours and predict outcomes over the remainder of the stay. Encounters with unavailable outcome labels are excluded from the corresponding task. For intervention endpoints, encounters with treatment already present at ICU entry or initiated during the observation window are also excluded rather than assigned negative labels.

IMV requires evidence of invasive ventilation and excludes non-invasive respiratory support. Vasoactive support includes milrinone, dopamine, dobutamine, isoproterenol, phenylephrine, angiotensin, vasopressin, epinephrine, and norepinephrine. CRRT denotes the first recorded initiation of continuous renal replacement therapy and excludes intermittent hemodialysis and peritoneal dialysis.

Table A.2 summarizes the cohorts, and Figure A.1 shows cohort selection.

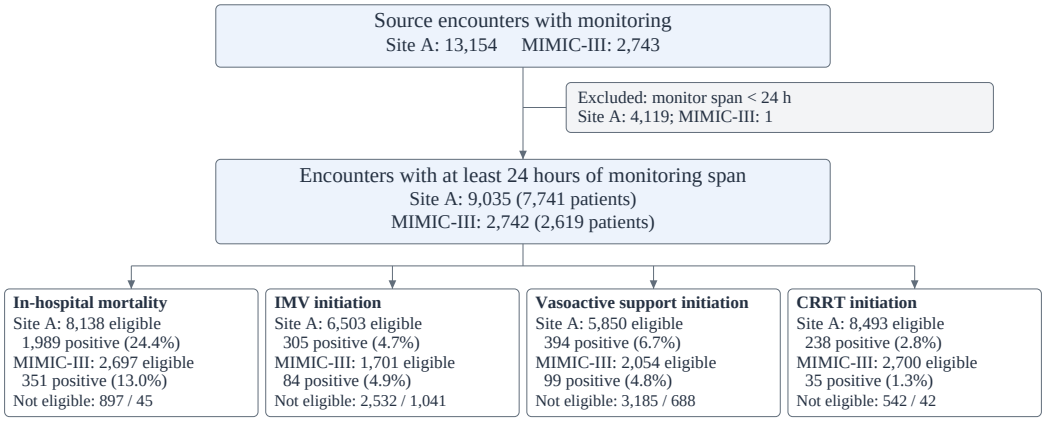


Figure A.1: **Downstream cohort selection.** Source encounters with monitoring, those with at least 24 hours of monitoring span, and task-eligible encounters for each endpoint with positives and prevalence. An encounter is not eligible for a task when its label is unavailable or the intervention was present at ICU entry or initiated within the observation window.

B Model and training details

B.1 Encoder and transfer data flow

Input. A 24-hour, eight-channel ICU vital-sign record and its observation mask, with dense channels (HR, SpO_2, RESP, ABP_D, ABP_S, TEMP) and sparse episodic NIBP channels (NIBP_D, NIBP_S); gaps in dense channels are native missingness and remain unfilled. **Patchification.** The window is split into $N = 1{,}440$ non-overlapping one-minute patches; the per-patch observation density $\pi_{v,n}$ (observed fraction of the 30 samples) is computed for every channel and patch. **Patch encoders.** Separate dense and sparse encoders map each channel-patch to a D-dimensional embedding, yielding $\mathbf{Z}_0 \in \mathbb{R}^{V \times N \times D}$. **Bi-axial blocks.** The encoder stacks n_L bi-axial blocks with separate learned parameters, using $n_L = 2$ in the reported model. Each block applies three operations: *(1) temporal modeling per channel*, a bidirectional Mamba sequence model with observation-aware temporal gating at its input and output; *(2) channel attention per time patch*, multi-head self-attention across the V channels at each time position; and *(3) feed-forward network*. Residual connections and normalization accompany these updates. **Pre-training head.** A shared linear decoder maps each channel–patch representation to 30 reconstructed two-second values. Its outputs are optimized using the combined reconstruction and correlation-matching objective in Eq. 1. **Downstream transfer.** The encoder output is pooled over observed patches within each channel. For frozen transfer, the encoder is held fixed, and the per-vital representations are concatenated and passed to a downstream classifier, such as TabPFN-3 or a linear probe. For end-to-end fine-tuning, the representations are further averaged across vital channels and passed to a prediction layer; the encoder and prediction layer are optimized jointly.

B.2 Dense and sparse patch encoders

Dense channels. For dense channels, the normalized values and observation mask are stacked as a two-channel input to a shared 1-D convolutional encoder. The encoder contains two convolutional layers with 128 output channels, kernel size 5, and padding 2, each followed by group normalization with eight groups and a ReLU activation. Within each patch, the resulting features $\mathbf{h}_l$ are averaged over observed positions:

$$\bar{\mathbf{h}} = \frac{\sum_{l=0}^{L-1} M_l \, \mathbf{h}_l}{\max\left(\sum_{l=0}^{L-1} M_l,\, 1\right)}. \tag{2}$$

where $L = 30$ and M_l indicates whether the position l is observed. The pooled features are linearly projected to a D-dimensional patch embedding. Encoder parameters are shared across the six dense channels.

Sparse channels. For episodic NIBP channels, each non-empty one-minute patch is summarized by its mean observed value and mean relative measurement position, with positions normalized by the patch length. The value summary is embedded using a linear projection followed by layer normalization and ReLU; the position summary is embedded using a separate linear projection. These embeddings are concatenated and linearly projected to a D-dimensional patch embedding. An empty patch receives a learned embedding $\mathbf{e}_\varnothing$. The sparse encoder is shared across the two NIBP channels.

B.3 Bi-axial block details

Figure B.1 shows the three stages within one bi-axial block; the dotted enclosure marked $\times n_L$ indicates sequential repetition of the complete block, with $n_L = 2$ in the reported model. Here, $\mathbf{Z}$ denotes the representation entering a bi-axial block. The temporal update produces $\mathbf{Z}_1$, channel attention produces $\mathbf{Z}_2$, and the feed-forward update produces $\mathbf{Z}'$, including the residual connection and normalization at each stage. For one record, all representations have shape $V \times N \times D$. The output $\mathbf{Z}'$ becomes the input to the next block when another block follows.

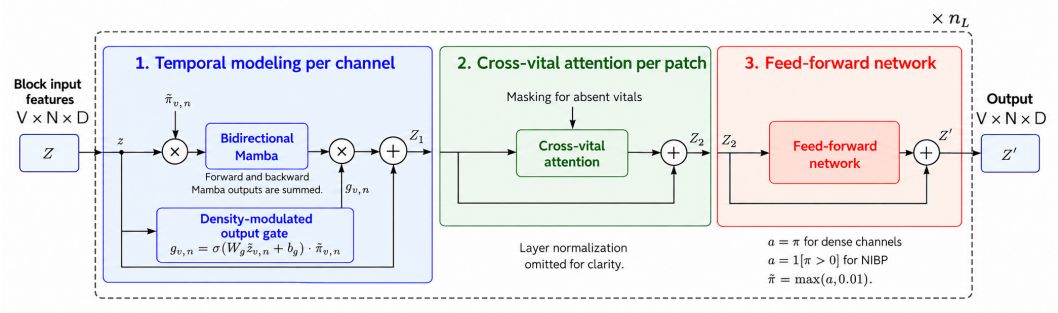


Figure B.1: **Bi-axial block and observation-aware temporal gating.** Per-channel patch density, binarized for NIBP, scales the temporal input before bidirectional Mamba processes each channel across time. A density-modulated output gate then scales the Mamba output before the residual update. Channel attention mixes available vital signs at each patch, followed by a feed-forward network, with residual connections at both stages.

For episodic NIBP channels, patch availability is set to 1 if at least one measurement is observed and to 0 otherwise. This replaces patch density before temporal gating, so a non-empty NIBP patch is not attenuated solely because it contains few measurements.

Temporal sub-block. Let $a_{b,v,n}$ denote the original patch density for dense channels and binary availability for NIBP channels. The temporal input gate scales patch embedding by $\tilde{\pi}_{b,v,n} = \max(a_{b,v,n}, 0.01)$, preserving a weak pathway through fully missing positions; the gated sequence is processed by bidirectional Mamba. Separate forward and backward Mamba branches process the sequence in opposite directions; their outputs are aligned to the original temporal order and summed. The density-modulated output gate combines a learned content term with the same density factor,

$$\mathbf{g}_{b,v,n} = \sigma\Big(W_g \tilde{\mathbf{z}}^{\text{temp}}_{b,v,n} + \mathbf{b}_g\Big)\ \tilde{\pi}_{b,v,n}, \tag{3}$$

where $\tilde{\mathbf{z}}^{\text{temp}}_{b,v,n} = \text{LN}_{\text{temp,pre}}(\mathbf{z}_{b,v,n})$ is the normalized block-input representation before temporal input gating, with $\mathbf{z}_{b,v,n} = (\mathbf{Z})_{v,n,:}$ for record b. The density-modulated output gate $\mathbf{g}$ is applied elementwise to the operator output before the residual update.

Input-density scaling and the density-modulated output gate act at different points in the temporal pathway. The former attenuates the representation presented to Mamba at a low-support patch. The latter, multiplied by the same density factor, separately regulates the temporal update returned at that position, which may contain information propagated from neighboring patches even when the local input has been strongly attenuated. The ablation without observation-aware temporal gating removes input-density scaling, output-density scaling, and the learned temporal content gate; it therefore evaluates the complete observation-aware temporal gating pathway rather than any individual mechanism in isolation.

Vital-channel sub-block. For one record b and time patch n, let $\mathbf{z}_v = (\mathbf{Z}_1)_{v,n,:} \in \mathbb{R}^D$ denote the temporally contextualized representation of channel v before the channel-attention LayerNorm. Suppressing b and n, write $a_v = a_{b,v,n}$ and define $p_v = \mathbf{1}[a_v > 0]$. Each head $h \in \{1, \ldots, H\}$, with $H = 4$, forms queries, keys, and values from $\tilde{\mathbf{z}}_v = \text{LN}_{\text{pre}}(\mathbf{z}_v)$:

$$\mathbf{q}^{(h)}_v = W^{(h)}_Q \tilde{\mathbf{z}}_v, \qquad \mathbf{k}^{(h)}_u = W^{(h)}_K \tilde{\mathbf{z}}_u, \qquad \mathbf{s}^{(h)}_u = W^{(h)}_V \tilde{\mathbf{z}}_u. \tag{4}$$

Here, v indexes the receiving channel and u the source channel. When at least one channel is present, the attention weights are

$$\alpha^{(h)}_{vu} = \text{softmax}_u \left(\frac{\mathbf{q}^{(h)\top}_v \mathbf{k}^{(h)}_u}{\sqrt{D/H}} + \beta_u \right), \qquad \beta_u = \begin{cases} 0, & p_u = 1, \\ -\infty, & p_u = 0. \end{cases} \tag{5}$$

This key-padding mask excludes naturally absent channels as sources; available channels may attend to themselves. The head outputs are combined, and the receiving channel's presence indicator gates

the resulting update:

$$\Delta \mathbf{z}_v = W_O \,\mathrm{Concat}_{h=1}^{H} \left(\sum_{u=1}^{V} \alpha_{vu}^{(h)} \mathbf{s}_u^{(h)} \right), \tag{6}$$

$$\mathbf{z}_v^{\mathrm{vital}} = \mathrm{LN}_{\mathrm{post}} \left(\mathbf{z}_v + p_v \Delta \mathbf{z}_v \right). \tag{7}$$

Here, $\mathbf{z}_v^{\mathrm{vital}} = (\mathbf{Z}_2)_{v,n,:}$. Projection biases are omitted for clarity. Layer normalization acts over features. Absent queries receive no attention update, although normalization and the subsequent feed-forward sub-block can still change their representations. If all channels are absent, the key-padding mask is temporarily cleared to avoid an entirely masked softmax, after which the query gate suppresses every attention update.

Feed-forward network. A position-wise feed-forward network maps each token from D to $4D$ and back to D, with a GELU activation between the two linear projections.

Normalization and residual connections. Layer normalization is applied before temporal input gating and before channel attention. Each temporal and channel-attention update is added to its corresponding residual input and then layer-normalized. The feed-forward network takes the post-attention representation directly; its output is added to that representation and then layer-normalized. The normalization modules at the three stages have separate learned parameters, shared across channels and time patches within each block.

Ordering. Temporal modeling with observation-aware gating is applied separately to each channel before attention mixes information across channels within each time patch.

B.4 Pre-training masking and reconstruction

At each training step, the eligible set comprises continuous channels with at least eight observed two-second samples and NIBP channels with measurements in at least four one-minute patches, counted across the windows processed at that step. One target channel is sampled uniformly from this set and withheld in every window processed at that step. Embeddings corresponding to its originally observed patches are replaced by a shared learned mask token $\mathbf{e}_{\mathrm{ssl}}$, while the original observation-availability indicators are retained.

B.5 Objective details

Let t denote the channel withheld at the current training step, $x_{b,i,v}$ the normalized ground-truth value, and $\hat{x}_{b,i,v}$ the corresponding decoder output. Here, b indexes records, i indexes temporal positions across the full window, and v indexes channels. The original observation mask $m_{b,i,v}$ is retained when the target values are hidden.

Reconstruction. The reconstruction loss averages squared errors over originally observed positions of the withheld channel:

$$\mathcal{L}_{\mathrm{recon}} = \frac{\sum_{b,i} m_{b,i,t} \left(\hat{x}_{b,i,t} - x_{b,i,t} \right)^2}{\sum_{b,i} m_{b,i,t} + \epsilon}. \tag{8}$$

The loss is computed in the normalized value domain, without additional channel-range weighting. Missing positions and visible context channels do not contribute.

Target-conditioned cross-vital correlation matching. For each record b and context channel $j \neq t$, define the jointly observed positions $\mathcal{S}_{b,tj} = \{i : m_{b,i,t} = m_{b,i,j} = 1\}$. Correlations are computed within each record across the full window, using the same jointly observed positions for the decoder outputs and ground-truth signals. For two sequences y and w restricted to a set $\mathcal{S}$, we use

$$r_{\mathcal{S}}(y, w) = \frac{\sum_{i \in \mathcal{S}} (y_i - \bar{y})(w_i - \bar{w})}{\sqrt{\sum_{i \in \mathcal{S}} (y_i - \bar{y})^2 + \epsilon} \sqrt{\sum_{i \in \mathcal{S}} (w_i - \bar{w})^2 + \epsilon} + \epsilon}, \tag{9}$$

where $\bar{y}$ and $\bar{w}$ are means over $\mathcal{S}$ and $\epsilon = 10^{-6}$ provides numerical stability.

Let $\mathcal{P}_t$ contain the record–channel pairs (b, j) with $j \neq t$ and $|\mathcal{S}_{b,tj}| \geq 10$. The correlation-matching loss is

$$\mathcal{L}_{\mathrm{corr}} = \frac{1}{\max(|\mathcal{P}_t|, 1)} \sum_{(b,j) \in \mathcal{P}_t} \left[r_{\mathcal{S}_{b,tj}}(\hat{x}_{b,t}, \hat{x}_{b,j}) - r_{\mathcal{S}_{b,tj}}(x_{b,t}, x_{b,j}) \right]^2, \tag{10}$$

where $x_{b,v}$ denotes the temporal sequence for channel v in record b. Each valid record–channel pair contributes equally; pairs are not weighted by their number of jointly observed samples. If no pair is valid, the correlation loss is zero.

The predicted correlation is computed between two decoder outputs, and the ground-truth correlation is treated as a fixed reference. The two objectives are complementary: reconstruction MSE anchors the withheld target to its observed values, whereas correlation matching provides secondary supervision that encourages the decoded target-context relationships to match those in the observed signals. We do not apply reconstruction MSE to the visible context channels because doing so could reward trivial copying rather than learning cross-vital relationships. Only pairs involving the withheld target contribute. This term encourages agreement in target-context correlation structure rather than measuring agreement between the reconstructed target and its ground truth. We use $\lambda_{\text{corr}} = 0.15$ in Eq. 1; the MSE-only ablation omits this term.

B.6 Optimization and configuration

Table B.1: **Pre-training configuration of the reported backbone.**

Setting	Value
Patch length / stride	30 / 30 samples
Embedding dimension	768
Bi-axial backbone depth	2 blocks
Patch encoders	Dense convolutional; sparse NIBP
Temporal model	Bidirectional Mamba
Mamba state size / convolution width / expansion	16 / 4 / 2
Channel attention	4 heads per block
Objective	Masked MSE + CorrMatch ($\lambda_{\text{corr}} = 0.15$)
Optimizer	AdamW; weight decay 10^{-4}
Learning rate	2.3×10^{-5}; 1,000-step warm-up
Batch size / training epochs	16 / 25

Table B.1 summarizes the architecture and pre-training configuration. Optimization uses AdamW (Loshchilov & Hutter, 2019), mixed precision, gradient checkpointing, and gradient-norm clipping at 1.

C Downstream evaluation and additional results

C.1 Downstream evaluation

Tasks, cohorts, eligibility criteria, and label construction are described in Appendix A.4. All methods use the same predefined, patient-disjoint five-fold partitions and observation windows in a rotating 60/20/20 train/validation/test design. Hyperparameters and checkpoints are selected using validation data only; validation patients are not included in model fitting, and the test fold remains unused until final evaluation.

C.2 Statistical analysis

AUROC and AUPRC are calculated separately in each of five patient-grouped held-out folds and reported as the unweighted mean $\pm$ standard deviation across folds. Paired contrasts are calculated within each fold and then averaged equally across the five folds. Uncertainty is estimated using a paired patient-cluster bootstrap with 10,000 valid replicates. For each replicate, patients are sampled with replacement once within each cohort. The resulting patient selection counts are reused across all outcomes, folds, and methods, with only records eligible for each task included. Metrics are recalculated within each fold before the paired differences are averaged. We report 95% bootstrap percentile confidence intervals and two-sided bootstrap p-values computed by doubling the smaller bootstrap tail probability (Efron & Tibshirani, 1993). Differences with $p < 0.05$ are described as nominally significant. Differences reported in the text are computed from full-precision aggregate means and may differ in the final displayed digit from differences calculated using the rounded values shown in the tables. The bootstrap resamples predictions from the fitted fold models; models are not retrained within replicates.

Table C.1: **Performance across temporal resolutions.** Mean ± standard deviation across five patient-disjoint folds. Native denotes the base 2-second grid. Best entries are shown in bold blue, and second-best entries are underlined.

		AUROC				AUPRC			
Cohort	Outcome	Native	1 min	5 min	1 h	Native	1 min	5 min	1 h
TabPFN-3									
Site A	Mort.	**71.0 ±2.0**	70.2 ±2.0	70.1 ±2.0	68.8 ±2.3	**45.9 ±4.0**	45.0 ±4.3	45.2 ±4.0	43.5 ±3.9
	IMV	**66.4 ±2.4**	65.9 ±3.0	65.7 ±3.1	63.9 ±3.2	**8.8 ±1.7**	8.4 ±1.5	8.1 ±1.2	7.8 ±1.1
	Vaso	**72.5 ±1.8**	71.8 ±1.2	71.7 ±1.2	70.6 ±1.3	**18.0 ±3.6**	17.8 ±3.8	17.6 ±4.2	16.5 ±3.3
	CRRT	75.7 ±2.9	**76.8 ±3.6**	75.9 ±3.3	74.5 ±3.0	9.6 ±1.2	**9.7 ±1.2**	9.0 ±0.8	9.0 ±1.2
MIMIC-III	Mort.	**79.2 ±3.2**	78.9 ±3.4	78.3 ±3.9	75.5 ±3.5	**39.5 ±6.9**	39.2 ±6.7	38.7 ±7.8	36.7 ±6.9
	IMV	**65.6 ±9.4**	64.9 ±8.6	65.5 ±8.5	63.6 ±10.1	**12.0 ±4.8**	11.9 ±4.3	11.2 ±2.9	11.3 ±3.8
	Vaso	**67.0 ±3.7**	65.5 ±4.8	66.0 ±5.3	66.2 ±5.5	9.9 ±1.5	9.2 ±1.8	9.5 ±1.6	**10.1 ±3.1**
	CRRT	74.7 ±13.5	75.7 ±12.4	**76.2 ±11.4**	73.8 ±12.0	6.6 ±3.7	7.5 ±5.1	**8.7 ±5.3**	6.9 ±4.8
TabICLv2									
Site A	Mort.	**71.1 ±1.8**	70.3 ±1.9	70.4 ±1.9	69.2 ±2.0	**45.9 ±3.7**	45.1 ±3.9	45.2 ±3.8	44.0 ±3.7
	IMV	**66.9 ±2.8**	66.3 ±2.9	65.7 ±2.9	63.6 ±3.1	8.7 ±1.6	**8.8 ±1.7**	8.5 ±1.6	7.6 ±1.1
	Vaso	**73.0 ±2.2**	72.2 ±1.5	71.7 ±1.7	70.7 ±2.2	**17.8 ±3.7**	17.7 ±4.3	17.1 ±3.7	16.0 ±2.4
	CRRT	76.7 ±2.2	**76.7 ±3.0**	75.8 ±2.7	74.5 ±3.3	**10.1 ±0.8**	9.9 ±1.1	9.5 ±1.0	9.1 ±1.5
MIMIC-III	Mort.	**79.6 ±3.9**	79.2 ±4.1	78.1 ±4.4	75.5 ±3.9	**39.2 ±7.3**	38.4 ±7.2	37.8 ±7.1	35.3 ±6.5
	IMV	**63.8 ±7.2**	63.3 ±6.8	62.6 ±7.2	62.7 ±9.5	11.4 ±3.7	**11.5 ±3.4**	10.6 ±1.7	10.6 ±1.9
	Vaso	**67.4 ±5.4**	66.8 ±4.9	66.9 ±5.6	67.1 ±5.8	10.4 ±2.6	10.2 ±2.7	10.5 ±2.4	**11.9 ±6.1**
	CRRT	**77.2 ±9.6**	76.4 ±9.9	74.9 ±9.7	74.0 ±11.5	7.9 ±5.2	8.1 ±4.2	**8.4 ±5.2**	7.7 ±4.3
LightGBM									
Site A	Mort.	**71.3 ±2.1**	70.2 ±1.9	69.9 ±2.1	68.4 ±2.2	**45.5 ±4.7**	44.6 ±4.2	44.5 ±4.6	42.8 ±4.4
	IMV	**64.0 ±4.1**	62.6 ±5.2	62.3 ±4.1	63.9 ±3.0	**8.4 ±2.5**	7.9 ±1.1	6.9 ±0.9	8.1 ±1.2
	Vaso	**71.8 ±1.7**	70.9 ±1.9	71.2 ±1.0	67.3 ±1.3	16.2 ±2.3	16.5 ±2.4	**18.0 ±3.3**	14.6 ±1.9
	CRRT	**75.2 ±2.7**	75.1 ±3.8	74.5 ±3.1	71.0 ±4.9	8.4 ±1.2	**8.7 ±1.3**	7.9 ±0.9	8.1 ±2.6
MIMIC-III	Mort.	**77.7 ±3.4**	76.7 ±3.2	76.3 ±4.0	73.8 ±2.7	**36.2 ±7.1**	32.7 ±5.8	33.9 ±5.6	31.9 ±4.4
	IMV	62.8 ±5.6	**66.4 ±4.0**	61.8 ±9.0	62.1 ±10.0	**12.7 ±4.1**	10.2 ±3.9	12.1 ±3.6	11.3 ±3.2
	Vaso	62.1 ±6.4	62.2 ±6.9	62.6 ±7.4	**63.9 ±6.8**	9.2 ±1.9	**9.5 ±2.2**	9.3 ±2.5	9.1 ±1.7
	CRRT	66.0 ±8.6	69.8 ±8.9	**74.3 ±6.7**	65.8 ±10.8	8.8 ±5.5	4.2 ±2.1	**10.5 ±4.7**	4.8 ±2.6

C.3 Temporal-resolution sensitivity

To assess how temporal coarsening affects prediction from bedside monitoring vital signs, we evaluate three downstream classifiers (TabPFN-3, TabICLv2, and LightGBM) using handcrafted features derived from the native 2-second grid and three coarser resolutions. One-minute, five-minute, and one-hour inputs are derived from the same native records by taking the median of observed values within each aggregation bin, while bins without observations remain missing. The handcrafted feature definitions in Table C.8 are applied at every resolution, with undefined statistics retained as missing. Encounters, patient-grouped folds, outcomes, and the downstream fitting protocol are held fixed; the paired comparisons therefore quantify the overall effect of temporal coarsening, including its effects on both signal variation and observation patterns. Table C.1 reports performance at each resolution, Figure C.1 shows paired coarsened-minus-native contrasts with 95% percentile paired patient-bootstrap confidence intervals, and Table C.2 reports the corresponding two-sided bootstrap p-values.

Across the three downstream classifiers, two cohorts, and four outcomes (24 comparisons per resolution and metric), models using native-resolution features achieve higher AUROC than those using one- and five-minute features in 18 and 20 comparisons, respectively, and higher AUPRC in 16 and 18. Five and six AUROC differences and four and three AUPRC differences are nominally significant, respectively; one favors the one-minute input, while the others favor native resolution. At hourly resolution, the separation is clearer: models using native-resolution features achieve higher AUROC in 23 of 24 comparisons and higher AUPRC in 21 of 24, with 14 and 9 nominally significant differences, respectively, none favoring the coarsened input. Overall, models using native-resolution features more often outperformed those using coarsened features, with differences reaching nominal significance most frequently at hourly aggregation.

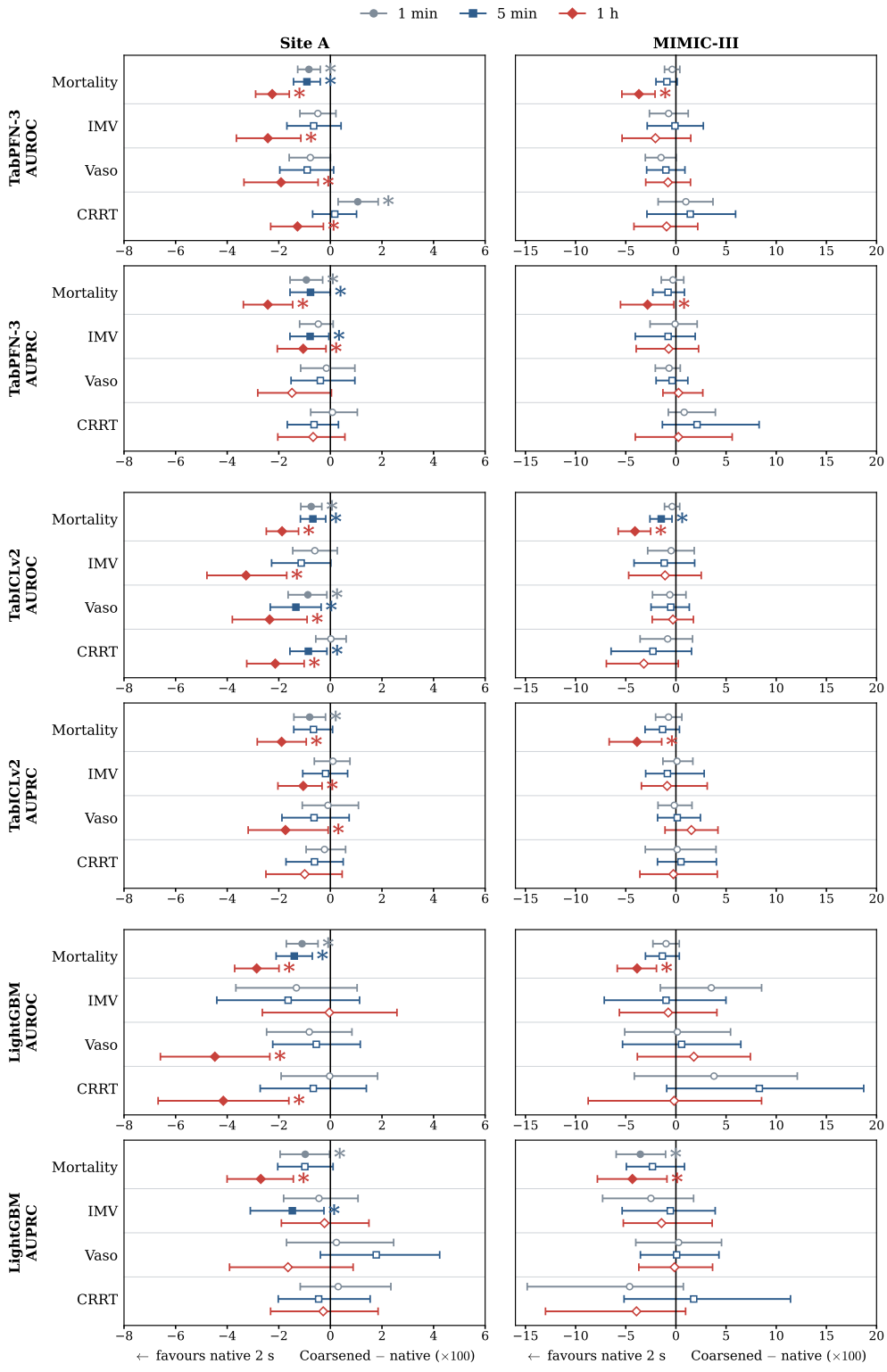


Figure C.1: **Effects of temporal coarsening.** Coarsened input minus native 2-second input ($\times 100$); negative values favor native resolution. Rows show classifier and metric, and colors indicate the one-minute, five-minute, and one-hour resolutions. Bars show 95% bootstrap percentile paired patient-bootstrap confidence intervals; asterisks denote nominal $p < 0.05$.

Table C.2: **Effects of temporal coarsening.** Differences in five-fold mean AUROC and AUPRC (coarsened input − native 2-second input, ×100), arranged as in Table C.1. Negative values favor native resolution; parentheses give two-sided paired patient-cluster bootstrap p-values, with nominal $p < 0.05$ shown in bold.

		AUROC			AUPRC		
Cohort	Outcome	1 min	5 min	1 h	1 min	5 min	1 h
TabPFN-3							
Site A	Mort.	−0.8 (<**0.001**)	−0.9 (<**0.001**)	−2.3 (<**0.001**)	−0.9 (**0.004**)	−0.8 (**0.0498**)	−2.4 (<**0.001**)
	IMV	−0.5 (0.168)	−0.6 (0.220)	−2.4 (<**0.001**)	−0.5 (0.099)	−0.8 (**0.038**)	−1.1 (**0.024**)
	Vaso	−0.8 (0.052)	−0.9 (0.087)	−1.9 (**0.009**)	−0.2 (0.845)	−0.4 (0.611)	−1.5 (0.058)
	CRRT	+1.1 (**0.005**)	+0.2 (0.707)	−1.3 (**0.016**)	+0.1 (0.771)	−0.6 (0.206)	−0.7 (0.301)
MIMIC-III	Mort.	−0.4 (0.342)	−0.9 (0.090)	−3.7 (<**0.001**)	−0.3 (0.552)	−0.8 (0.371)	−2.8 (**0.033**)
	IMV	−0.7 (0.463)	−0.1 (0.945)	−2.0 (0.237)	−0.1 (0.876)	−0.8 (0.539)	−0.7 (0.578)
	Vaso	−1.5 (0.056)	−1.0 (0.298)	−0.8 (0.503)	−0.7 (0.205)	−0.4 (0.589)	+0.3 (0.685)
	CRRT	+1.0 (0.452)	+1.4 (0.499)	−0.9 (0.558)	+0.8 (0.365)	+2.1 (0.218)	+0.2 (0.848)
TabICLv2							
Site A	Mort.	−0.7 (<**0.001**)	−0.7 (**0.008**)	−1.9 (<**0.001**)	−0.8 (**0.011**)	−0.7 (0.084)	−1.9 (<**0.001**)
	IMV	−0.6 (0.178)	−1.1 (0.055)	−3.3 (<**0.001**)	+0.1 (0.904)	−0.2 (0.612)	−1.1 (**0.007**)
	Vaso	−0.9 (**0.023**)	−1.3 (**0.007**)	−2.4 (**0.002**)	−0.1 (0.961)	−0.6 (0.388)	−1.7 (**0.038**)
	CRRT	+0.0 (0.943)	−0.9 (**0.018**)	−2.1 (<**0.001**)	−0.2 (0.741)	−0.6 (0.297)	−1.0 (0.197)
MIMIC-III	Mort.	−0.4 (0.333)	−1.5 (**0.007**)	−4.1 (<**0.001**)	−0.7 (0.295)	−1.3 (0.125)	−3.9 (**0.002**)
	IMV	−0.5 (0.691)	−1.2 (0.444)	−1.1 (0.547)	+0.1 (0.839)	−0.9 (0.866)	−0.9 (0.888)
	Vaso	−0.6 (0.460)	−0.5 (0.602)	−0.3 (0.777)	−0.2 (0.863)	+0.1 (0.783)	+1.5 (0.266)
	CRRT	−0.8 (0.542)	−2.3 (0.242)	−3.2 (0.070)	+0.1 (0.816)	+0.5 (0.686)	−0.3 (0.980)
LightGBM							
Site A	Mort.	−1.1 (<**0.001**)	−1.4 (<**0.001**)	−2.9 (<**0.001**)	−1.0 (**0.044**)	−1.0 (0.074)	−2.7 (<**0.001**)
	IMV	−1.3 (0.273)	−1.6 (0.241)	−0.0 (0.963)	−0.4 (0.634)	−1.5 (**0.018**)	−0.2 (0.797)
	Vaso	−0.8 (0.336)	−0.5 (0.531)	−4.5 (<**0.001**)	+0.2 (0.733)	+1.8 (0.107)	−1.6 (0.212)
	CRRT	−0.0 (0.975)	−0.7 (0.527)	−4.2 (**0.003**)	+0.3 (0.563)	−0.5 (0.713)	−0.3 (0.787)
MIMIC-III	Mort.	−1.0 (0.141)	−1.4 (0.120)	−3.9 (<**0.001**)	−3.5 (**0.008**)	−2.3 (0.164)	−4.3 (**0.015**)
	IMV	+3.5 (0.170)	−1.0 (0.747)	−0.8 (0.754)	−2.5 (0.276)	−0.6 (0.832)	−1.4 (0.720)
	Vaso	+0.1 (0.970)	+0.6 (0.852)	+1.8 (0.531)	+0.3 (0.937)	+0.1 (0.857)	−0.1 (0.970)
	CRRT	+3.8 (0.351)	+8.3 (0.079)	−0.2 (0.966)	−4.6 (0.102)	+1.8 (0.483)	−3.9 (0.139)

C.4 VitalFM evaluation settings and controls

C.4.1 Frozen VitalFM transfer

For frozen transfer, the pre-trained encoder remains fixed. The primary per-vital representation averages observed patch embeddings over time within each of the eight channels and concatenates the resulting 768-dimensional channel representations, producing 6,144 features. This representation is evaluated with TabPFN-3 using its native balanced feature subsampling. As a pooling sensitivity, the channel representations are instead averaged across available channels to produce a single 768-dimensional representation, which is evaluated with TabPFN-3 and TabICLv2. Pre-training ablations additionally evaluate the per-vital representation with the linear probe (Appendix C.6).

Table C.3: **Frozen VitalFM transfer.** Five-fold mean ± SD, ×100. Per-vital pooling with TabPFN-3 is the primary analysis; pooling across available channels is evaluated as a sensitivity with TabPFN-3 and TabICLv2. Best entries are blue and bold; second-best entries are underlined.

		AUROC			**AUPRC**		
		TabPFN-3		TabICLv2	TabPFN-3		TabICLv2
Cohort	Outcome	Per-vital	Pooled	Pooled	Per-vital	Pooled	Pooled
Site A	Mort.	**72.0 ±1.9**	71.7 ±1.7	71.1 ±1.5	**47.4 ±3.7**	46.7 ±3.1	46.3 ±2.7
	IMV	**67.1 ±3.0**	66.9 ±3.3	66.0 ±3.9	9.1 ±1.2	**9.4 ±1.5**	9.3 ±2.5
	Vaso	**73.3 ±2.4**	72.2 ±2.5	72.0 ±1.6	**19.4 ±3.4**	17.7 ±2.5	18.1 ±1.9
	CRRT	76.6 ±2.9	**76.8 ±3.0**	75.9 ±2.7	10.7 ±1.3	**10.8 ±1.1**	10.3 ±0.7
MIMIC-III	Mort.	**79.1 ±4.4**	77.3 ±3.7	77.1 ±3.1	**38.1 ±7.2**	35.2 ±4.5	35.4 ±5.3
	IMV	61.1 ±6.5	**61.6 ±3.9**	61.3 ±3.9	**10.9 ±2.5**	7.8 ±1.7	8.2 ±2.7
	Vaso	**64.9 ±8.0**	64.4 ±8.2	63.4 ±8.3	8.8 ±2.5	**9.4 ±5.1**	8.8 ±3.0
	CRRT	**77.0 ±4.2**	76.8 ±6.1	74.9 ±5.7	9.5 ±6.5	9.8 ±5.0	**9.8 ±8.6**
Mean		**71.4**	71.0	70.2	**19.2**	18.3	18.3

With TabPFN-3, the per-vital representation achieved mean AUROC/AUPRC $0.4/0.9$ points higher than the pooled representation.

C.4.2 Fine-tuned downstream contrasts

Fine-tuning. The encoder is initialized from the pre-trained checkpoint and optimized jointly with a linear prediction head for five epochs on each downstream task using binary cross-entropy loss. Observed patch embeddings are averaged within each channel, and the resulting channel representations are averaged across available channels to form the input to the prediction head. The backbone learning rate is 5×10^{-5} for Site A and 7.5×10^{-5} for MIMIC-III; the head learning rate is 3×10^{-4} in both cohorts. Checkpoints are ranked using a composite validation score that equally weights the AUROC and AUPRC components after normalizing each relative to its random-ranking baseline: $s = \frac{1}{2}(2\,\mathrm{AUROC} - 1) + \frac{1}{2}(\mathrm{AUPRC} - p)/(1-p)$, where both metrics are computed on the validation fold and p is its positive-class prevalence. The same score is used for the no-pre-training control. Fine-tuned VitalFM and its no-pre-training control average the parameters of the three checkpoints with the highest composite validation scores, yielding one model per fold. Because the composite validation score balances normalized AUROC and AUPRC but can remain noisy when validation folds contain few positive events, we average the three highest-ranked checkpoints to reduce reliance on a single potentially unstable validation maximum; this uses no test data and does not ensemble predictions.

Paired downstream contrasts. Table C.4 reports the per-setting differences between each comparator and fine-tuned VitalFM, together with two-sided paired patient-cluster bootstrap p-values. The corresponding 95% bootstrap percentile confidence intervals are shown in Figure 2.

Table C.4: **Fine-tuned VitalFM versus downstream baselines.** Differences in AUROC and AUPRC (comparator − fine-tuned VitalFM, $\times 100$) for each cohort–outcome setting. Negative values favor VitalFM; parentheses give two-sided bootstrap p-values, with nominal $p < 0.05$ shown in bold.

	Site A				MIMIC-III			
Comparison	Mort.	IMV	Vaso	CRRT	Mort.	IMV	Vaso	CRRT
AUROC								
Raw + TabPFN-3	−5.4 (<**0.001**)	−4.1 (**0.002**)	−3.5 (**0.005**)	−3.5 (**0.003**)	−5.1 (<**0.001**)	+0.8 (0.723)	−0.8 (0.742)	−8.1 (**0.026**)
Handcr. + LGBM	−1.7 (<**0.001**)	−3.6 (**0.010**)	−0.7 (0.548)	−0.4 (0.743)	−1.9 (**0.046**)	−1.8 (0.577)	−2.9 (0.399)	−12.4 (**0.006**)
Handcr. + TabPFN-3	−1.9 (<**0.001**)	−1.2 (0.287)	+0.1 (0.891)	+0.2 (0.885)	−0.4 (0.650)	+1.0 (0.590)	+2.1 (0.396)	−3.7 (0.251)
Handcr. + TabICLv2	−1.9 (<**0.001**)	−0.7 (0.512)	+0.6 (0.540)	+1.1 (0.296)	−0.0 (0.962)	−0.8 (0.643)	+2.5 (0.289)	−1.2 (0.771)
GRU-D	−9.1 (<**0.001**)	−5.1 (**0.003**)	−6.2 (<**0.001**)	−5.6 (<**0.001**)	−10.4 (<**0.001**)	−5.7 (0.144)	−4.4 (0.232)	−17.1 (**0.002**)
TCN	−5.1 (<**0.001**)	−6.4 (<**0.001**)	−8.8 (<**0.001**)	−5.3 (<**0.001**)	−12.5 (<**0.001**)	−7.2 (**0.041**)	−2.9 (0.407)	−22.5 (**0.001**)
Transformer	−8.9 (<**0.001**)	−7.0 (<**0.001**)	−8.9 (<**0.001**)	−8.3 (<**0.001**)	−13.2 (<**0.001**)	−3.6 (0.229)	−0.1 (0.963)	−19.6 (<**0.001**)
VitalFM w/o pre-training	−15.8 (<**0.001**)	−11.3 (<**0.001**)	−13.4 (<**0.001**)	−10.6 (<**0.001**)	−23.0 (<**0.001**)	−15.8 (<**0.001**)	−10.4 (**0.005**)	−17.8 (<**0.001**)
AUPRC								
Raw + TabPFN-3	−6.4 (<**0.001**)	−1.8 (**0.013**)	−3.2 (**0.007**)	−2.9 (**0.002**)	−4.6 (**0.019**)	+0.0 (0.976)	+1.2 (0.690)	−4.7 (**0.048**)
Handcr. + LGBM	−3.0 (<**0.001**)	−1.5 (0.060)	−1.5 (0.187)	−1.8 (0.071)	−3.4 (**0.048**)	+2.1 (0.501)	−2.0 (0.369)	−0.1 (0.802)
Handcr. + TabPFN-3	−2.6 (<**0.001**)	−1.1 (0.146)	+0.2 (0.924)	−0.6 (0.542)	−0.1 (0.918)	+1.4 (0.340)	−1.3 (0.484)	−2.3 (0.274)
Handcr. + TabICLv2	−2.6 (<**0.001**)	−1.2 (0.104)	−0.0 (0.959)	−0.1 (0.821)	−0.4 (0.846)	+0.9 (0.753)	−0.8 (0.654)	−1.0 (0.486)
GRU-D	−11.0 (<**0.001**)	−2.6 (**0.003**)	−1.9 (0.178)	−3.8 (<**0.001**)	−12.8 (<**0.001**)	+0.6 (0.808)	−2.8 (0.175)	−5.6 (**0.029**)
TCN	−6.8 (<**0.001**)	−1.9 (0.051)	−5.6 (<**0.001**)	−2.9 (**0.021**)	−13.2 (<**0.001**)	−2.9 (0.078)	+0.7 (0.851)	−6.9 (**0.001**)
Transformer	−10.8 (<**0.001**)	−2.0 (0.055)	−4.7 (<**0.001**)	−3.4 (<**0.001**)	−13.1 (<**0.001**)	+0.4 (0.976)	−0.4 (0.866)	−6.8 (<**0.001**)
VitalFM w/o pre-training	−17.7 (<**0.001**)	−3.2 (**0.002**)	−6.8 (<**0.001**)	−4.7 (<**0.001**)	−21.8 (<**0.001**)	−5.1 (<**0.001**)	−3.4 (0.118)	−4.6 (0.154)

C.4.3 Matched no-pre-training controls

Matched no-pre-training control. The primary no-pre-training control matches the reported model's architecture (31.1M parameters; width D=768). The encoder and prediction-head parameters are initialized randomly. The model is trained end-to-end using binary cross-entropy loss, an effective batch size of 16, and a learning rate of 3×10^{-4}, for up to 60 epochs with patience 10. Checkpoints use the same composite validation score and top-three parameter-averaging procedure as fine-tuned VitalFM. This matched control achieved lower performance than fine-tuned VitalFM in every cohort-outcome cell for both metrics, with average gaps of 14.8 AUROC and 8.4 AUPRC points. Because model width can affect supervised optimization, we also performed a model-size sensitivity analysis under the same no-pre-training protocol (Table C.5). This analysis varies width and total parameter count at fixed depth; it is not an exhaustive search over model depth or architecture. Its purpose is to test whether the architecture-matched comparison is explained by the difficulty of optimizing a large model from limited outcome labels. Mean performance decreased monotonically with model size. The best-performing model in this sensitivity analysis had 1.0M parameters; fine-tuned VitalFM retained higher performance in all eight cells for both metrics, by 9.6 AUROC and 6.3 AUPRC points on average.

Table C.5: **No-pre-training model-size sensitivity.** Five-fold mean ± sample SD, ×100; the final column averages the eight cell means. Rows are labeled by total parameters; the 1.0M, 3.6M, 14.0M, and 31.1M models have widths D=128, 256, 512, and 768 with two bi-axial blocks. Best entry is blue and bold; second-best entry is underlined.

	Site A				MIMIC-III				
Model size	Mort.	IMV	Vaso	CRRT	Mort.	IMV	Vaso	CRRT	Mean
AUROC									
1.0M	**67.0 ±2.0**	56.3 ±1.7	**66.0 ±2.8**	**68.3 ±5.8**	**65.9 ±4.4**	49.8 ±7.4	61.3 ±6.3	64.4 ±12.2	**62.38**
3.6M	64.0 ±1.1	**57.3 ±4.8**	62.2 ±1.1	64.9 ±1.4	63.5 ±3.5	**50.2 ±4.5**	60.8 ±6.8	66.0 ±9.8	61.11
14.0M	55.5 ±2.5	56.0 ±1.7	61.3 ±4.4	66.5 ±3.0	63.6 ±3.2	48.1 ±4.6	**63.6 ±9.7**	**66.4 ±7.3**	60.13
31.1M	57.2 ±1.1	56.3 ±3.8	59.0 ±3.7	65.0 ±2.8	56.6 ±9.8	48.8 ±10.6	54.6 ±8.4	60.6 ±8.0	57.23
AUPRC									
1.0M	**38.9 ±3.0**	6.8 ±1.3	**13.2 ±1.4**	**7.2 ±2.7**	**23.4 ±2.6**	5.4 ±1.5	8.3 ±1.0	3.2 ±1.3	**13.32**
3.6M	36.3 ±2.8	**7.2 ±1.1**	12.3 ±2.0	5.7 ±0.7	20.7 ±3.7	6.0 ±1.6	**8.4 ±1.3**	3.8 ±1.4	12.56
14.0M	29.1 ±3.0	6.3 ±0.5	11.6 ±2.3	7.1 ±2.2	21.0 ±3.8	**6.2 ±2.7**	8.3 ±1.6	3.2 ±0.8	11.62
31.1M	30.8 ±1.8	6.7 ±1.3	11.0 ±1.7	5.5 ±0.6	17.8 ±1.8	5.5 ±1.1	7.8 ±2.2	**4.3 ±3.1**	11.18

Episodic NIBP sensitivity. We evaluated whether episodic NIBP channels contribute to downstream performance and whether they benefit from a dedicated sparse encoder. Because repeating full-corpus pre-training for every NIBP variant would be computationally costly, we conduct this sensitivity analysis using VitalFM-S w/o pre-training, the best-performing randomly initialized configuration in the model-size sweep. We compared the full model with variants that either removed both NIBP channels or processed them using the shared convolutional encoder, while holding the downstream protocol fixed (Table C.6). Removing both NIBP channels reduced the eight-cell macro mean by 1.51 AUROC and 0.78 AUPRC points, indicating that episodic blood-pressure measurements provide useful predictive information overall, although the effect varies across tasks. Compared with processing NIBP through the shared convolutional encoder, the dedicated sparse encoder achieved higher AUROC in all six non-IMV cohort–outcome settings but lower AUROC for IMV in both cohorts. AUPRC differences were more mixed, with each encoder performing better in four of eight settings. Consequently, the two encoders had similar macro performance: the dedicated encoder differed from the shared encoder by -0.15 AUROC and $+0.19$ AUPRC points. We therefore retain the dedicated encoder because it directly reflects the episodic measurement process and performs strongly outside IMV, without claiming a uniform performance advantage across tasks. This experiment is a supervised architectural sensitivity rather than a pre-training ablation.

Table C.6: **NIBP sensitivity.** Five-fold mean ± sample SD (×100) for matched VitalFM-S w/o pre-training configurations (1.0M parameters). Configurations differ only in whether NIBP is processed by the dedicated sparse encoder, processed by the shared convolutional encoder, or excluded. The final column averages the eight cell means.

	Site A				MIMIC-III				
Configuration	Mort.	IMV	Vaso	CRRT	Mort.	IMV	Vaso	CRRT	Mean
AUROC									
Dedicated sparse encoder (reference)	**67.0 ±2.0**	56.3 ±1.7	**66.0 ±2.8**	**68.3 ±5.8**	**65.9 ±4.4**	49.8 ±7.4	**61.3 ±6.3**	**64.4 ±12.2**	62.38
Shared convolutional encoder	65.3 ±4.4	**61.6 ±1.9**	63.8 ±3.2	67.8 ±2.1	63.0 ±1.5	54.3 ±8.9	60.1 ±9.0	64.2 ±9.1	**62.53**
NIBP channels removed	64.7 ±3.4	58.8 ±3.8	57.5 ±2.7	61.3 ±2.6	61.6 ±6.7	**61.2 ±9.5**	58.4 ±5.4	63.5 ±8.6	60.87
AUPRC									
Dedicated sparse encoder (reference)	38.9 ±3.0	6.8 ±1.3	**13.2 ±1.4**	**7.2 ±2.7**	**23.4 ±2.6**	5.4 ±1.5	8.3 ±1.0	3.2 ±1.3	**13.32**
Shared convolutional encoder	**39.3 ±5.5**	7.3 ±1.0	12.2 ±0.6	6.8 ±1.8	20.6 ±2.1	6.8 ±1.6	**9.1 ±2.6**	3.0 ±1.2	13.13
NIBP channels removed	37.5 ±1.8	**7.5 ±1.6**	9.5 ±1.5	5.0 ±0.6	21.5 ±4.9	**8.4 ±3.2**	7.1 ±1.3	**3.8 ±2.6**	12.54

C.5 Baseline models

Baseline selection follows established ICU prediction benchmarks (Purushotham et al., 2018; Harutyunyan et al., 2019; Yèche et al., 2021; van de Water et al., 2024).

C.5.1 Tabular models and representations

For LightGBM and the linear probe, hyperparameters are selected using the equally weighted normalized AUROC–AUPRC validation composite defined in Appendix C.4.2. Evaluating the same

Table C.7: **Downstream-classifier settings.** Preprocessing parameters were fitted on the training partition within each patient-grouped fold, and model selection used the corresponding validation partition. The test partition was not used for preprocessing, selection, or model fitting.

Classifier	Input and preprocessing	Configuration and selection
LightGBM	Handcrafted features or flattened raw grids; all available columns are retained and missing values are handled natively.	Binary objective, no class reweighting, and at most 2,000 trees with early stopping. Fifty Optuna trials per fold search learning rate $[0.005, 0.1]$, leaves $[16, 128]$, depth $[4, 12]$, minimum leaf size $[10, 100]$, ℓ_1 and ℓ_2 penalties $[10^{-3}, 10]$, row subsampling $[0.6, 1]$, feature subsampling $[0.5, 1]$, and minimum split gain $[0, 1]$.
TabPFN-3	Handcrafted features, flattened raw grids, or frozen VitalFM representations; no external feature selection.	Tuning-free 16-estimator base ensemble with balanced native feature subsampling for high-dimensional inputs. For the primary 6,144-feature VitalFM representation, the ensemble expands to 31 estimators to provide complete feature coverage.
TabICLv2	Handcrafted features, six-minute raw grids, or pooled VitalFM representations; columns entirely missing in the training fold are removed.	Tuning-free ensemble with eight estimators.
Linear probe	The complete 6,144-feature per-vital VitalFM representation; training-fold median imputation with missingness indicators and standardization.	Logistic regression with an ℓ_2 penalty and no class weighting. The validation fold selects C from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$.

handcrafted features with LightGBM, TabPFN-3, and TabICLv2 assesses compatibility with distinct downstream learners without changing the input representation.

Handcrafted features. Following the seven-region organization of Harutyunyan et al. (2019), each of the eight monitor channels is summarized over the full 24-hour window and its first 10%, 25%, and 50% and last 50%, 25%, and 10% temporal regions. That benchmark computes six statistics within each region: minimum, maximum, mean, standard deviation, skewness, and measurement count. We retain the five value-based statistics and add the median, 10th and 90th percentiles, slope, and mean absolute change, yielding ten summaries per region. A full-window recency term is also included for each channel, producing $8 \times (7 \times 10 + 1) = 568$ features (Table C.8). Measurement count is excluded because it scales directly with sampling rate, and undefined statistics remain missing rather than being imputed. Table C.9 compares our temporal feature construction with structured baselines used in prior ICU studies.

Table C.8: **Handcrafted feature set.** These prespecified feature definitions are applied unchanged across cohorts, outcomes, folds, and temporal resolutions.

Component	Definition	Count
Channels	Heart rate; SpO_2; respiratory rate; invasive arterial diastolic and systolic pressure; temperature; NIBP diastolic and systolic pressure	8
Windows	Full 24-hour window; first 10%, 25%, 50%; last 50%, 25%, 10%, relative to the observation-window length	7
Statistics	Mean, median, minimum, maximum, 10th and 90th percentiles, standard deviation, slope, mean absolute change, skewness	10
Recency	Hours from the last observation to the prediction cutoff, once per channel	1
Width	$8 \times (7 \times 10 + 1)$	568

This design gives conventional classifiers a rich summary representation. In addition, the raw six-minute and patch-aligned one-minute controls test whether conclusions depend on the particular handcrafted feature set. Any advantage of VitalFM is therefore assessed against both a rich summary representation and minimally processed temporal grids.

C.5.2 Raw vital-sign grids

Table C.10 reports per-cell results for a one-minute raw-grid sensitivity analysis, in which observed values are median-aggregated within one-minute bins; the primary six-minute results are repeated for reference. The six-minute grid contains 1,920 features and is supplied unchanged to LightGBM, TabPFN-3, and TabICLv2. The one-minute grid contains 11,520 features: LightGBM consumes the complete grid, whereas TabPFN-3 uses its native feature subsampling. TabICLv2 is omitted because this dimensionality exceeds its supported feature budget and it lacks an analogous subsampling mechanism. With TabPFN-3 fixed, frozen VitalFM exceeded the six-minute raw grid by 3.1 AUROC

Table C.9: **Feature engineering in ICU prediction studies.** Published structured baselines compared with the handcrafted representation used in this study. The comparison focuses on how longitudinal variables, including vital signs, are summarized for conventional classifiers.

Study	Structured classifier	Temporal feature construction
Harutyunyan et al. (2019)	Logistic regression	Minimum, maximum, mean, standard deviation, skewness, and measurement count over seven temporal regions
YAIB (van de Water et al., 2024)	Logistic regression and LightGBM	Minimum, maximum, mean, and measurement count for each dynamic variable over the available history
HiRID-ICU (Yèche et al., 2021)	Logistic regression and LightGBM	Minimum, maximum, mean, and measurement density for each non-categorical variable over the available history
ICareFM (Burger et al., 2026)	LightGBM	For continuous variables, minimum, 10th percentile, median, 90th percentile, maximum, mean, standard deviation, slope, mean absolute change, and fraction observed over 8-, 24-, and 72-hour histories
Clin-JEPA (Yang et al., 2026)	Ridge regression and LightGBM	Raw clinical features; temporal summarization not specified
VitalFM (ours)	LightGBM, TabPFN-3, and TabICLv2	Minimum, maximum, mean, standard deviation, skewness, median, 10th and 90th percentiles, slope, and mean absolute change over seven temporal regions, plus one recency term per channel

and 2.5 AUPRC points and the one-minute raw grid by 3.7 AUROC and 2.7 AUPRC points on average. The two raw-grid resolutions are sensitivity controls rather than a controlled coarsening comparison, since the one-minute TabPFN-3 input uses feature subsampling. Consistent with the weak separation among minute-scale grids (Appendix C.3), LightGBM, which consumes both grids directly, favored the one-minute grid by 0.7 AUROC points but the six-minute grid by 0.7 AUPRC points on average. VitalFM's advantage therefore holds under both raw-grid constructions.

Table C.10: **Raw vital-sign grids.** Mean ± sample standard deviation across five patient-disjoint folds, ×100. Six-minute primary results and one-minute sensitivity results. Best entry is blue and bold; second-best entry is underlined, with ranks determined before rounding.

	Site A				MIMIC-III			
Model	Mort.	IMV	Vaso	CRRT	Mort.	IMV	Vaso	CRRT
AUROC								
Raw 6-min + TabPFN-3	67.6 ±2.1	**63.5 ±0.3**	**68.9 ±3.0**	**72.1 ±2.3**	**74.5 ±4.9**	**65.4 ±9.5**	**64.2 ±5.2**	70.3 ±11.4
Raw 6-min + TabICLv2	67.3 ±2.1	60.8 ±2.7	66.8 ±2.3	69.9 ±2.8	72.7 ±5.3	61.6 ±8.3	56.6 ±6.2	64.0 ±10.2
Raw 6-min + LightGBM	**67.8 ±2.0**	59.6 ±2.4	65.2 ±2.7	67.8 ±6.3	73.3 ±3.8	60.1 ±6.9	60.5 ±3.3	64.5 ±11.2
Raw 1-min + TabPFN-3	66.9 ±2.3	62.8 ±0.8	68.4 ±2.5	71.7 ±2.0	74.4 ±5.2	63.8 ±8.5	63.5 ±6.0	70.4 ±10.8
Raw 1-min + LightGBM	66.7 ±2.4	58.7 ±2.3	63.3 ±3.3	67.2 ±4.1	72.3 ±4.4	62.4 ±5.5	63.3 ±4.3	**70.5 ±11.7**
AUPRC								
Raw 6-min + TabPFN-3	42.1 ±3.8	**8.0 ±1.6**	**14.5 ±2.5**	7.3 ±1.4	**35.0 ±5.3**	**10.6 ±3.1**	12.4 ±5.3	4.2 ±1.8
Raw 6-min + TabICLv2	41.3 ±4.5	7.0 ±1.5	13.4 ±2.1	6.8 ±1.4	32.5 ±5.6	9.3 ±3.0	7.4 ±1.6	4.5 ±2.5
Raw 6-min + LightGBM	**42.2 ±3.8**	6.8 ±1.3	12.6 ±2.5	7.2 ±2.4	30.8 ±5.5	9.1 ±3.8	**12.7 ±4.1**	**8.6 ±8.0**
Raw 1-min + TabPFN-3	40.9 ±3.6	7.9 ±1.4	14.5 ±2.5	**7.3 ±1.3**	34.2 ±5.5	10.3 ±2.9	11.9 ±5.7	5.1 ±2.4
Raw 1-min + LightGBM	40.9 ±4.2	7.9 ±0.6	12.8 ±2.8	6.1 ±1.2	31.6 ±6.2	8.7 ±1.9	11.8 ±3.6	4.7 ±2.0

C.5.3 Supervised sequence models

We compare VitalFM with three established supervised sequence architectures representing recurrent, convolutional, and attention-based approaches: GRU-D (Che et al., 2018), TCN (Bai et al., 2018; Catling & Wolff, 2020), and a Transformer encoder (Vaswani et al., 2017). We use fixed, literature-informed configurations based on the original architecture descriptions and related sequence-model evaluations in the YAIB ICU benchmark (van de Water et al., 2024).

Vital-sign measurements are represented on a shared two-second grid, so a 24-hour window contains 43,200 temporal positions, which would be computationally demanding for these architectures. We therefore physiologically filter the measurements (Appendix A.1) and median-aggregate them into 1,440 one-minute steps, with separate observation masks and elapsed-time features, allowing each model to process the complete 24-hour window. VitalFM also operates on 1,440 temporal positions, but each position is produced by a learned patch encoder from the 30 two-second samples and their

Table C.11: **Supervised sequence-baseline configurations.** Parameter counts include the task head. One fixed, literature-informed capacity is used per family and held constant across cohorts and outcomes.

Family	Architecture	Fixed capacity	Parameters
GRU-D	One recurrent layer; learned input and hidden decay; dropout 0.2	Hidden width 128	56,977
TCN	Eight causal residual blocks; kernel 4; dilations 1–128; dropout 0.1	128 filters per block	1,002,753
Transformer	Two pre-norm encoder layers plus final LayerNorm; feed-forward width $4D$; dropout 0.2	$D = 128$ with 4 heads	400,129

masks within that minute. The comparison therefore matches channels, observation window, and number of temporal positions, but the models differ in how within-minute information is represented: fixed median aggregation in the baselines versus a learned encoding in VitalFM.

GRU-D, TCN, and Transformer are trained with AdamW using learning rates of 3×10^{-4}, 2×10^{-4}, and 1.5×10^{-4}, respectively, weight decay 10^{-6}, and effective batch size 32. Training continues for at most 60 epochs with early stopping after 10 epochs without improvement. Checkpoints are ranked using the composite validation score defined in Appendix C.4.2, and the highest-scoring checkpoint is retained for the primary evaluation. Fine-tuned VitalFM instead averages the parameters of its three highest-ranked checkpoints. We assess sensitivity to this difference by applying top-three parameter averaging to each sequence baseline and single-checkpoint selection to VitalFM.

Table C.12: **Checkpoint-selection sensitivity.** Macro means across the eight cohort–outcome settings ($\times 100$). The primary protocol uses top-three parameter averaging for VitalFM and the single validation-best checkpoint for each sequence baseline; the sensitivity reverses that choice. Differences are relative to each model's primary result.

Model	Sensitivity readout	Mean AUROC/AUPRC	ΔAUROC/AUPRC
VitalFM	Single validation-best checkpoint	70.36 / 18.39	$-1.64/-1.19$
GRU-D	Top-three parameter average	64.03 / 14.52	$-0.05/-0.06$
TCN	Top-three parameter average	63.44 / 14.57	$+0.26/-0.07$
Transformer	Top-three parameter average	63.80 / 14.63	$+0.51/+0.16$

Under matched top-three parameter averaging, fine-tuned VitalFM retained mean AUROC/AUPRC margins of approximately $8.0/5.1$, $8.6/5.0$, and $8.2/5.0$ points over GRU-D, TCN, and Transformer, respectively. Under single-checkpoint evaluation, VitalFM remained approximately $6.3/3.8$ points above GRU-D, the sequence baseline with the highest primary mean AUROC. These sensitivities indicate that the large aggregate separation from the supervised sequence baselines is not attributable to checkpoint selection alone.

C.6 Pre-training ablation configurations and results

We evaluate seven pre-trained configurations. The reference model is VitalFM with bidirectional Mamba, width $D = 768$, two bi-axial blocks, and MSE+CorrMatch ($\lambda_{\text{corr}} = 0.15$). The six variants test the contribution of correlation matching (MSE only), the full temporal-gating pathway, temporal directionality (unidirectional Mamba), replacement of Mamba with gated self-attention, a masked-Transformer temporal backbone, and reduced model capacity (VitalFM-S; $D = 128$, 1.0M parameters).

The temporal-gating ablation jointly removes input-density scaling, output-density scaling, and the learned temporal content gate; it therefore evaluates the full temporal-gating pathway rather than any individual mechanism in isolation. The gated-attention variant replaces Mamba with multi-head temporal self-attention and a position-wise feed-forward sublayer while retaining VitalFM's temporal gates, providing the closer comparison of temporal operators. The masked-Transformer variant instead uses conventional pre-normalized temporal attention and feed-forward layers, binary temporal key-padding masks, suppression of absent queries, and sinusoidal positional encoding applied once before the block stack. It retains VitalFM's patch encoders and bi-axial scaffold and therefore serves as a literature-aligned temporal-backbone reference rather than an isolated

operator substitution. At the reference width and depth, these alternatives contain 30.2M and 29.0M parameters, respectively, compared with 31.1M for VitalFM.

Table C.13: **Per-vital + TabPFN-3 ablations.** Matched-evaluation frozen-transfer performance ($\times 100$) on the complete $8 \times D$ per-vital representation. TabPFN-3 uses native balanced feature subsampling. Values are five-fold mean ± sample SD across folds. Best entries are blue and bold; second-best entries are underlined, with ranks determined before rounding.

	Site A				MIMIC-III			
Configuration	Mort.	IMV	Vaso	CRRT	Mort.	IMV	Vaso	CRRT
AUROC								
VitalFM (MSE+CorrMatch)	**72.0 ±1.9**	**67.1 ±3.0**	73.3 ±2.4	**76.6 ±2.9**	**79.1 ±4.4**	61.1 ±6.5	64.9 ±8.0	77.0 ±4.2
MSE only	70.5 ±2.3	65.2 ±2.4	70.8 ±2.9	75.6 ±4.2	76.5 ±2.9	**66.7 ±6.0**	64.2 ±7.5	73.3 ±7.3
No observation-aware gating	71.8 ±1.7	66.0 ±2.5	**73.7 ±1.8**	76.5 ±2.7	78.4 ±4.1	62.5 ±8.3	66.1 ±5.3	75.2 ±7.8
Unidirectional Mamba	70.8 ±1.8	65.9 ±2.2	71.4 ±2.5	76.5 ±3.3	76.9 ±3.8	59.1 ±7.4	65.1 ±5.7	76.8 ±8.2
Gated-attention backbone	70.2 ±1.6	64.3 ±3.3	70.9 ±2.7	74.9 ±2.4	77.6 ±4.2	55.6 ±9.6	64.2 ±8.8	74.6 ±6.8
Standard masked Transformer	69.6 ±1.5	66.9 ±2.3	70.7 ±2.3	74.5 ±2.8	71.7 ±3.4	61.1 ±7.7	64.4 ±6.8	75.3 ±13.7
VitalFM-S	70.6 ±1.8	66.0 ±2.0	72.2 ±2.7	74.4 ±3.9	76.7 ±3.0	59.5 ±9.7	**66.1 ±5.7**	**78.3 ±4.2**
AUPRC								
VitalFM (MSE+CorrMatch)	**47.4 ±3.7**	**9.1 ±1.2**	**19.4 ±3.4**	**10.7 ±1.3**	**38.1 ±7.2**	10.9 ±2.5	8.8 ±2.5	9.5 ±6.5
MSE only	46.0 ±4.2	8.7 ±1.5	17.2 ±2.5	9.8 ±2.0	37.0 ±6.3	**10.9 ±3.4**	8.8 ±1.9	6.7 ±3.9
No observation-aware gating	46.9 ±3.6	8.7 ±1.4	19.0 ±3.1	10.1 ±1.1	37.8 ±5.5	9.7 ±3.2	**10.3 ±3.0**	8.7 ±6.2
Unidirectional Mamba	46.2 ±4.0	8.2 ±1.4	17.2 ±2.0	9.9 ±1.2	35.7 ±5.3	8.8 ±2.8	9.6 ±2.8	**13.8 ±7.9**
Gated-attention backbone	44.9 ±3.4	8.0 ±1.7	16.8 ±3.4	9.0 ±1.3	36.6 ±5.9	7.3 ±2.5	8.8 ±1.9	7.7 ±5.4
Standard masked Transformer	44.2 ±3.3	8.3 ±1.6	15.9 ±2.7	9.0 ±1.9	32.1 ±4.2	9.2 ±3.2	9.7 ±2.3	7.7 ±2.2
VitalFM-S	45.0 ±3.7	8.9 ±1.8	18.6 ±2.1	10.0 ±1.8	36.7 ±5.6	9.9 ±4.9	9.9 ±2.3	10.6 ±5.7

All configurations use the same pre-training corpus, downstream cohorts, patient folds, eligible samples, and downstream evaluation protocol. Frozen encoders are evaluated using the complete $8 \times D$ per-vital representation with TabPFN-3 and the linear probe. Except for the component under study, model width and depth are held fixed; VitalFM-S provides the reported capacity sensitivity.

These ablations evaluate targeted configurations rather than independently optimizing every alternative or conducting an exhaustive depth–width search. They therefore characterize the selected designs under a shared training protocol rather than each alternative's best achievable performance. Tables C.13 and C.14 report five-fold mean ± sample standard deviation for the cohort–outcome results summarized in Table 3. The full VitalFM values use the same canonical reference reported in Table 1.

Table C.14: **Per-vital + linear-probe ablations.** Matched-evaluation frozen-transfer performance ($\times 100$) on the complete $8 \times D$ per-vital representation. Values are five-fold mean $\pm$ sample SD across folds. Best entries are blue and bold; second-best entries are underlined, with ranks determined before rounding.

Configuration	Site A				MIMIC-III			
	Mort.	IMV	Vaso	CRRT	Mort.	IMV	Vaso	CRRT
AUROC								
VitalFM (MSE+CorrMatch)	71.7 ±1.3	65.6 ±3.4	72.9 ±3.0	73.2 ±3.1	**76.9 ±4.5**	61.1 ±5.3	62.2 ±4.3	75.3 ±8.2
MSE only	70.5 ±2.0	63.2 ±2.2	70.8 ±2.7	73.6 ±4.0	73.8 ±2.3	**63.9 ±5.6**	56.6 ±8.1	72.4 ±8.8
No observation-aware gating	**71.8 ±1.3**	64.4 ±3.7	**73.2 ±1.6**	74.9 ±2.8	75.9 ±4.8	56.0 ±7.3	**64.4 ±3.8**	73.6 ±13.2
Unidirectional Mamba	71.7 ±1.2	61.5 ±4.5	71.3 ±4.2	**75.5 ±3.7**	75.1 ±4.6	60.0 ±7.1	61.3 ±5.9	**78.3 ±11.3**
Gated-attention backbone	70.4 ±1.3	63.1 ±3.7	70.4 ±3.0	73.0 ±4.1	74.9 ±4.7	54.3 ±7.8	57.7 ±5.7	74.6 ±9.9
Standard masked Transformer	69.2 ±1.8	**66.1 ±4.4**	69.5 ±2.3	72.3 ±1.3	72.1 ±4.7	56.8 ±11.7	60.6 ±6.6	77.3 ±8.4
VitalFM-S	70.8 ±1.7	61.8 ±4.2	71.0 ±4.2	72.6 ±4.6	74.9 ±3.2	56.2 ±10.2	58.9 ±10.2	77.9 ±7.9
AUPRC								
VitalFM (MSE+CorrMatch)	45.8 ±3.8	9.2 ±1.8	18.8 ±2.0	**10.5 ±1.8**	34.7 ±5.9	**10.8 ±3.7**	7.9 ±2.0	8.8 ±9.1
MSE only	45.4 ±4.1	7.6 ±1.2	17.1 ±2.7	9.6 ±1.9	33.2 ±5.2	9.5 ±3.2	7.7 ±2.8	5.6 ±2.4
No observation-aware gating	**46.7 ±4.0**	8.1 ±1.4	**19.2 ±3.7**	9.5 ±0.8	34.0 ±4.9	9.6 ±5.2	**9.6 ±1.3**	5.5 ±2.7
Unidirectional Mamba	46.3 ±3.7	7.4 ±1.7	18.6 ±2.1	9.7 ±2.4	33.5 ±5.6	9.9 ±5.4	8.3 ±2.1	**14.1 ±8.8**
Gated-attention backbone	45.6 ±3.2	8.0 ±1.6	17.4 ±3.4	8.9 ±1.7	33.8 ±5.7	7.6 ±2.8	8.6 ±1.6	5.4 ±1.9
Standard masked Transformer	43.5 ±3.5	**9.9 ±2.9**	15.1 ±2.2	8.5 ±2.5	31.3 ±5.0	9.0 ±4.9	7.7 ±0.9	11.0 ±6.4
VitalFM-S	45.3 ±3.1	8.4 ±1.4	17.6 ±3.1	9.9 ±1.2	**35.2 ±5.3**	7.9 ±3.1	8.3 ±1.6	11.5 ±4.9

C.6.1 Matched full-validation reconstruction rescore

We exhaustively rescored the six matched D=768 configurations and VitalFM-S on the same 3% pre-training validation split. Each vital-sign channel was withheld in turn at all observed patches and reconstructed from the remaining channels and observation masks. Using every target channel and validation window avoids variation from target-channel sampling. Metrics were calculated separately for each target vital and averaged equally across the eight vitals. Normalized RMSE divides each vital's RMSE by its prespecified range, permitting aggregation across signals with different units and numerical scales. Reconstruction PCC is the Pearson correlation between predicted and observed target values and is distinct from the target-conditioned cross-vital correlation-matching loss used during training.

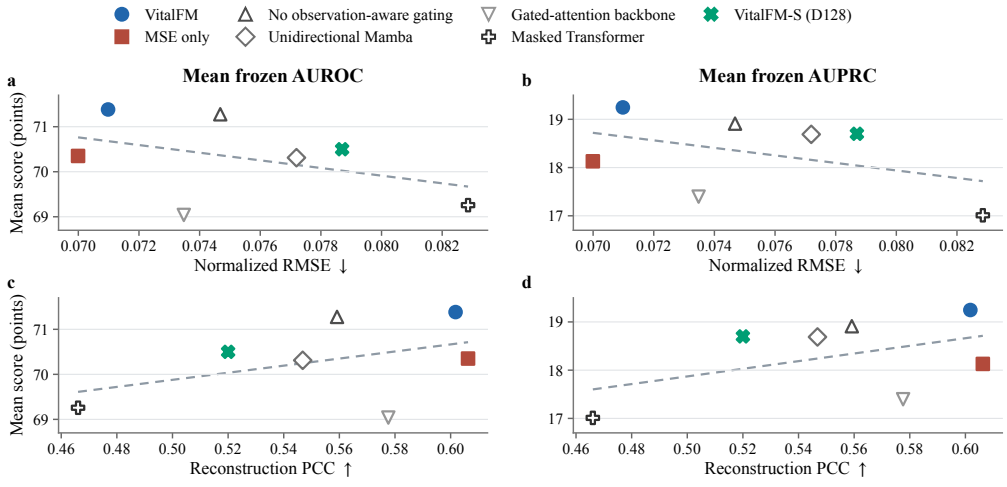


Figure C.2: **Reconstruction quality versus frozen transfer.** Normalized RMSE and reconstruction PCC are plotted against mean frozen-transfer AUROC and AUPRC across eight cohort–outcome settings ($\times 100$). Points represent the six D=768 configurations and VitalFM-S (D=128), evaluated with TabPFN-3. Dashed lines are descriptive least-squares fits; absolute values are reported in Table C.15.

Per-vital reconstruction results for the full VitalFM model are reported in Table C.16.

Table C.15: **Matched validation reconstruction and frozen transfer.** Exhaustive validation reconstruction and mean frozen-transfer performance across eight cohort–outcome settings. The first six configurations have width D=768; VitalFM-S has width D=128. Frozen-transfer metrics use TabPFN-3 and are reported as percentages. Among the reconstruction metrics, best entries are blue and bold and second-best entries are underlined, with ranks determined before rounding.

Configuration	Norm. RMSE ↓	PCC ↑	AUROC	AUPRC
VitalFM (MSE+CorrMatch)	<u>0.0710</u>	<u>0.602</u>	71.38	19.25
MSE only	**0.0700**	**0.606**	70.35	18.13
No observation-aware gating	0.0747	0.559	71.28	18.91
Unidirectional Mamba	0.0772	0.547	70.31	18.69
Gated-attention backbone	0.0735	0.578	69.04	17.40
Standard masked Transformer	0.0829	0.466	69.26	17.01
VitalFM-S (D=128)	0.0787	0.520	70.47	18.69

Table C.16: **Per-vital reconstruction performance.**

Target vital	RMSE ↓	PCC ↑
HR	16.523	0.501
SpO_2	3.196	0.290
Respiratory rate	7.421	0.547
Diastolic ABP	6.874	0.860
Systolic ABP	13.025	0.839
Temperature	1.130	0.250
Diastolic NIBP	9.778	0.742
Systolic NIBP	16.345	0.785

Reconstruction quality was directionally associated with downstream transfer, but did not consistently track configuration rankings (Figure C.2). VitalFM-S further illustrates this distinction, combining relatively weak reconstruction with comparatively strong downstream performance. We therefore treat reconstruction metrics as complementary evidence rather than a substitute for downstream evaluation. Major architectural changes generally affected both exhaustive reconstruction and frozen transfer, but the exact configuration rankings did not coincide. Reconstruction-only pre-training illustrates this distinction: it attained the strongest reconstruction metrics, whereas adding correlation matching improved frozen transfer by 1.03 AUROC and 1.12 AUPRC points. Relative to the no-observation-aware-gating variant, full VitalFM increased reconstruction PCC from 0.559 to 0.602 and reduced normalized RMSE from 0.0747 to 0.0710, with smaller changes in downstream performance. The gated-attention backbone reconstructed more accurately than the standard masked Transformer while attaining slightly lower mean AUROC. Reconstruction metrics therefore diagnose major representational degradations but do not by themselves determine downstream utility.

C.6.2 Validation evaluation of the correlation-matching objective

Figure 3 compares matched 31.1M-parameter models pre-trained with MSE alone and with MSE+CorrMatch. At six checkpoints, both models were evaluated on the same complete held-out pre-training validation split. We report normalized target reconstruction MSE and target-conditioned cross-vital correlation mismatch; the latter is distinct from reconstruction PCC. The plotted values are checkpoint-level validation results.