Curriculum-Aware Interpolate-then-Refine: Learned Physiological Time-Series Imputation under Realistic Missingness

Yu-Chao Huang, Haochen Zhang, Nicholas Konz, and Tianlong Chen

[1] UNITES Lab, University of North Carolina at Chapel Hill

Imputing physiological time series (arterial blood pressure, blood glucose, etc.) is essential for addressing the missingness that pervades clinical data. Yet modern imputation methods perform poorly in this domain: a recent benchmark found that simple linear interpolation outperformed *every* learned imputer on real-world clinical signals with realistic gaps. We show that this reflects two properties of physiological missingness that generic imputers ignore: gaps may occur when the signal is clinically extreme rather than typical, and gap lengths can easily span orders of magnitude. To this end, we introduce Curriculum-Aware Interpolate-then-Refine (CAIR), a two-stage framework for physiological time-series imputation. Our key motivation is to learn a coarse base curve and then repeatedly correct it toward physiological realism, rather than predict a gap in a single pass. Consequently, CAIR couples a bidirectional-GRU interpolator with a Transformer refiner that corrects its own estimate over three successive passes, trained jointly under a broad, signal-agnostic random-gap curriculum. We evaluate imputers stratified by gap length and missingness mechanism (MCAR, MAR, NMAR) rather than by a single average, and CAIR is the most accurate under every mechanism on continuous glucose monitoring (AI-READI) and arterial pressure in intensive care (MIMIC-III). Its margin over the strongest baseline grows with difficulty, from 9% under MCAR to 19% under value-dependent dropout, where generic learned imputers are weakest. We further show low reconstruction error alone does not recover the burden metrics clinicians act on: interpolants matching CAIR's error fail to preserve those metrics, imputers that recover them are far less accurate, and CAIR alone ranks among the best on both axes.

1 Introduction

Physiological time series, such as continuous glucose monitoring (CGM), heart rate, respiration and intensive-care vital signs, are the raw material of modern data-driven medicine, and they are rarely complete (Pratap et al., 2020; Braem et al., 2024; De Andrade et al., 2026). Sensors detach, wearables run out of battery, patients move, and clinical devices are disconnected during procedures (Braem et al., 2024; Bent et al., 2020). Every downstream step, from computing a glycemic burden metric (Kok et al., 2026) to training a risk model (Hurst et al., 2024), first has to decide what the missing values were. Imputation is therefore not a preprocessing detail but a modeling choice that propagates into every clinical conclusion drawn from the data (Cichosz et al., 2025; Poette et al., 2026).

The obvious remedy is a learned imputer, and a rich family now exists (Cao et al., 2018; Du et al., 2022; Tashiro et al., 2021; Wu et al., 2022). Yet a recent real-world benchmark reports a negative result:

✉ Corresponding authors: {morris, haochenz, nick124, tianlong}@cs.unc.edu

Preprint. Under review.

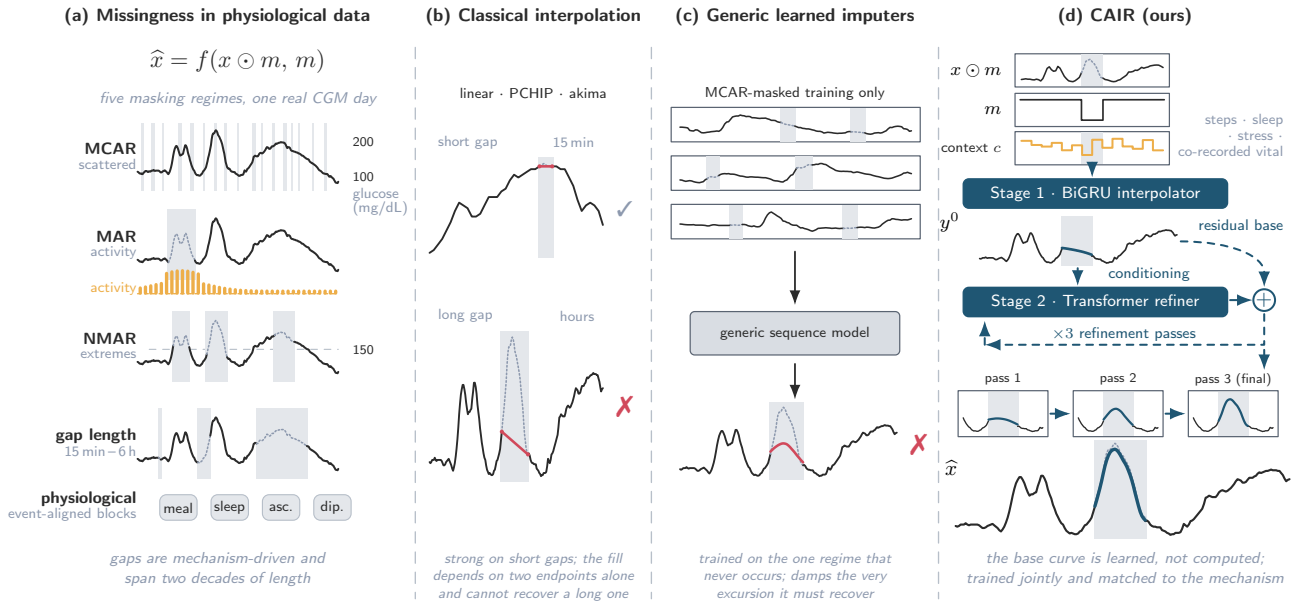

Figure 1: **Method Overview.** *Notation:* $x \in \mathbb{R}^T$ is the complete signal and $m \in \{0,1\}^T$ its observation mask ($m_t = 1$ where x_t is recorded), so the elementwise product $x \odot m$ is the observed (gapped) trace; c collects the co-recorded context channels (steps, sleep, stress, a second vital); f is an imputer and $\widehat{x} = f(x \odot m,\, m)$ its reconstruction; and y^0 is the Stage-1 base curve. Shaded bands mark masked spans throughout. Every trace is recorded AI-READI CGM: (a) is one participant's 24 h day, (b)–(d) zoom into one postprandial excursion from it, and the top row of (b) zooms further so that a 15 min gap is visible. **(a)** The five masking regimes this paper evaluates, all on one signal: MCAR scatters isolated points; MAR deletes contiguous blocks triggered by co-recorded activity; NMAR deletes blocks triggered by the target's own clinically extreme values (here every run above 150 mg/dL); the gap-length protocol sweeps a single block from 15 min to hours; and the physiological protocol masks event-aligned blocks. Dotted grey is the signal each mask removes. **(b)** Deterministic interpolants are the deployed default and are accurate on short gaps, but whatever they place inside a gap is a function of the two endpoints alone, so a long gap is filled by a chord that cannot express the excursion it spans. **(c)** Generic learned imputers are trained under MCAR, the one regime that does not occur, and damp the excursion they must recover. **(d)** CAIR *learns* that base curve instead of computing it: a jointly trained BiGRU interpolator whose base curve y^0 enters a Transformer refiner both as conditioning and as a residual base, unrolled for three passes and conditioned on whatever co-recorded context the deployment setting provides.

on clinical signals with realistic gaps, *linear interpolation* outperforms every method tested, including deep learning models (Toye et al., 2025). We argue this is not evidence that learning cannot help, but a symptom of two properties of physiological missingness that generic imputers and the standard evaluation protocol both ignore:

(C1) **Missingness is mechanism-driven, not random.** We adopt the taxonomy of Rubin (1976), under which missingness is *missing completely at random* (MCAR) when the probability that a sample is lost is independent of the data, *missing at random* (MAR) when that probability depends only on observed quantities such as a co-recorded covariate, and *missing not at random* (NMAR) when it depends on the missing value itself. Physiological sensors fail more often during activity and during the clinical extremes that matter most, so their gaps are MAR or NMAR rather than MCAR. Imputers trained and scored under MCAR are therefore optimized for the one regime that never occurs.

(C2) **Gap length spans orders of magnitude.** A trace contains both single dropped samples and multi-hour holes. Averaging error over a fixed masking protocol lets the rare long gaps dominate the

mean, so a single score can rank a method first while it is strictly worse in the short-gap regime required in clinical practice.

We introduce **Curriculum-Aware Interpolate-then-Refine** (**CAIR**), a two-stage neural imputer designed around these two properties. Our key motivation is that the strength of classical interpolation is a *starting point*, not a ceiling: a coarse curve that respects the observed endpoints is straightforward to construct, and the challenge is to correct it toward physiological realism. A deterministic interpolant, however, fixes that starting point in advance: whatever it places inside a gap is a function of the two endpoints alone, which makes it accurate over a few missing samples but unable to express a multi-hour excursion. CAIR therefore *learns* the base curve rather than computing it. Stage 1 is a bidirectional GRU (Cho et al., 2014; Schuster & Paliwal, 1997) that predicts a base curve at every position from the observed values and the mask. Stage 2 is a Transformer encoder (Vaswani et al., 2017) that consumes this base curve (both as conditioning and as a residual), together with any other modalities the dataset provides, and corrects it over three successive refinement passes, each conditioned on the previous estimate.

To address (C1), CAIR is trained under a signal-agnostic random-gap curriculum that mixes scattered dropouts with contiguous blocks, rather than under the single masking pattern used at evaluation time. The training gap distribution proves as important as the architecture, which the name of the method reflects: the *same* network attains higher error than linear interpolation when trained on masks tailored to the physiology of a source domain, and lower error under all three mechanisms when trained on the broad curriculum (Sec. 4.6). To address (C2), we evaluate every method stratified by gap length and by missingness mechanism instead of reporting one average, and we show that the ranking of methods depends on the length regime, so a single averaged score cannot express it (Sec. 4.3).

Contributions. Our contributions are threefold:

- **Methodologically**, we propose CAIR, a two-stage imputer that replaces the deterministic interpolant with a jointly trained learned interpolator, and refines it with an iteratively unrolled Transformer. The design targets the two properties of physiological missingness above (C1, C2) rather than generic sequence modeling.
- **Empirically**, across two clinical domains, CGM from AI-READI (AI-READI Consortium, 2024) and arterial pressure from MIMIC-III (Johnson et al., 2016), CAIR attains the lowest reconstruction error of every method we evaluate under all three mechanisms, and its margin over the strongest baseline *grows* with difficulty: 9% under MCAR, 16% under MAR, and 19% under NMAR. The generic learned imputers that Toye et al. (2025) report as failures fail here too, which isolates how the model is designed and trained, rather than neural capacity, as the cause of the gain.
- **Analytically**, we show that low reconstruction error and faithful downstream clinical metrics are distinct axes. Interpolants that match the reconstruction error of CAIR fail to preserve the burden metrics clinicians read (linear interpolation recovers 0.14 of the recoverable burden on MIMIC-III), while tabular and neural imputers that do recover the burden incur 10–60% higher error. CAIR is the only method that ranks among the best on both axes.

2 Related Work

Classical and statistical imputation. Deployed clinical pipelines still rely on deterministic interpolants: linear fills, shape-preserving cubics such as PCHIP (Fritsch & Carlson, 1980) and akima (Akima, 1970), and smoothers such as Savitzky–Golay (Savitzky & Golay, 1964). They carry no fitting cost and remain accurate on short gaps, which is why they are still the default for CGM metric computation (Cichosz et al., 2025). Statistical imputers (MICE (Van Buuren & Groothuis-Oudshoorn, 2011), missForest (Stekhoven & Bühlmann, 2012), kNN and hot-deck) instead treat the series as a table of features, recovering distributional structure that interpolation discards at the cost of temporal smoothness. Both families share the property CAIR targets: what they place inside a gap is determined in advance, either by the two endpoints or by a marginal distribution, and cannot be adapted to the model that consumes it. CAIR keeps the interpolate-then-refine structure that motivates the classical prior, but makes the first stage learned and trainable end-to-end, so the base curve adapts to the refiner.

Learned time-series imputation. Recurrent imputers exploit informative missingness directly: GRU-D uses decay toward the empirical mean (Che et al., 2018), BRITS imputes bidirectionally with consistency between directions (Cao et al., 2018), and M-RNN combines intra- and inter-stream recurrence (Yoon et al., 2018). Attention-based and generative approaches followed. Closest to CAIR, SAITS (Du et al., 2022) estimates with one diagonally-masked self-attention block, writes those estimates back into the input, and re-estimates with a second, supervising both; we retain that structure but make the first stage an information-restricted recurrent interpolator, unroll a single tied-weight refiner over it, and mark the model's own fills with a provenance code at every pass. A separate line makes the interpolation step itself learned: interpolation prediction networks (Shukla & Marlin, 2019) and mTAN (Shukla & Marlin, 2021) attach a learned interpolation layer to a downstream network and train the pair end-to-end. Those layers are kernel smoothers designed to place irregular samples on a regular grid, and they feed a classifier; CAIR's Stage 1 is a recurrent sequence model supervised directly against held-out values, and what it feeds is a refiner that corrects it. GP-VAE (Fortuin et al., 2020) places a Gaussian-process prior in a VAE latent space, CSDI (Tashiro et al., 2021) runs conditional score-based diffusion, and backbones such as TimesNet (Wu et al., 2022) treat imputation as one of several tasks. All are domain-agnostic, and our results suggest physiological imputation needs a domain-specific model: M-RNN and GP-VAE fall below every non-constant baseline on both CGM and ICU vitals.

Evaluating under realistic missingness. Rubin's MCAR/MAR/NMAR taxonomy (Rubin, 1976) is standard in statistics but is rarely used to structure machine-learning imputation benchmarks, which typically delete values completely at random. Recent work pushes back: Qian et al. (2024) and Poette et al. (2026) both show that clinically plausible missingness patterns change method rankings. Closest to our work, Toye et al. (2025) evaluate eleven imputers on real-world clinical signals under mechanism-driven deletion and find that linear interpolation outperforms all of them. We adopt their protocol verbatim as our hardest evaluation, and show that their result is a statement about generic imputers rather than about learned imputation in general. Detailed discussion is in the supplementary material.

3 Method

3.1 Problem Setup

A physiological trace is a uniformly sampled series $x \in \mathbb{R}^T$ with observation mask $m \in \{0,1\}^T$ (m_t=1 if x_t is observed; 5-minute grid throughout). We write $\odot$ for the elementwise product, $e \in \{0,1\}^T$ for the held-out evaluation mask, and $\widehat{x}$ for the imputed trace; a *gap* is a maximal run of consecutive missing positions and its *length* is that run's size. Imputation predicts $\widehat{x}_t$ at every missing position (m_t=0) from the observed ones. Values are z-normalized with training statistics and rescaled to native units for reporting. When auxiliary channels are available (co-recorded vitals, activity, sleep state), they are appended to the per-position conditioning and the model is otherwise unchanged.

3.2 Interpolate-then-Refine Architecture

CAIR has two neural stages, illustrated in Fig. 1d.

Stage 1: Learned interpolator. A bidirectional GRU (Cho et al., 2014; Schuster & Paliwal, 1997) reads the masked signal together with its mask, $f_t = [\, x_t m_t,\ m_t \,]$, and predicts a base value at *every* position,

$$y^0 = \mathrm{Head}\big(\mathrm{BiGRU}(f)\big) \in \mathbb{R}^T, \tag{3.1}$$

with hidden size 128 and 4 layers, where Head is a single linear map from the 256-dimensional concatenated forward and backward hidden state to one value per position. This is the component that replaces the deterministic interpolant: where a classical pipeline computes a PCHIP or akima curve from the two gap endpoints, y^0 is trained. The output head is zero-initialized, so at the start of training the model reduces to its refiner and learns the base curve as a residual correction.

Stage 2: Transformer refiner. A bidirectional pre-norm Transformer encoder (Vaswani et al., 2017) (d_{model}=128, 8 layers, 8 heads, FFN width 512) corrects the base curve. The interpolator output enters

twice: once inside the per-position conditioning, occupying the slots a classical pipeline reserves for its interpolant, and once as an additive residual base,

$$\widehat{x} = y^0 + \mathrm{Transformer}\big(\phi(x \odot m,\, m,\, y^0,\, \tau)\big), \tag{3.2}$$

where ϕ is the per-position conditioning built from the observed samples, the mask, learned time-of-day embeddings τ, and y^0. The two stages are assigned complementary roles: the GRU produces a smooth base curve, and the Transformer adds the data-driven physiological shape (post-prandial rises, nocturnal dips, pressure excursions) that a smooth base cannot express.

3.3 Training Objective

CAIR is trained from scratch under a *signal-agnostic random-gap curriculum*: at each step a mixture of scattered points and contiguous blocks ($\sim$20% of observed samples) is held out and reconstructed. The curriculum is deliberately not specialized for the typical failure modes of any single signal type, which keeps the method general (C1). Indeed, Sec. 4.6 shows that using masks tailored to the physiology of a *source* domain results in worse generalization to new signals.

The refiner is unrolled for three passes, each conditioned on the previous prediction, under an increasing weight schedule that emphasizes the final pass. An auxiliary term supervises the Stage-1 output directly so it learns a smooth base rather than collapsing into the refiner:

$$\mathcal{L} = \sum_{k=1}^{3} w_k \big\| (\widehat{x}^{(k)} - x) \odot e \big\|^2 + \lambda_{\text{aux}} \big\| (y^0 - x) \odot e \big\|^2, \tag{3.3}$$

with $w = (0.15,\, 0.35,\, 0.50)$ and λ_{aux}=0.7. Optimization uses AdamW (Loshchilov & Hutter, 2017) at learning rate 3×10^{-4} with cosine decay, and we keep an exponential moving average of the weights (decay 0.999) for inference (Izmailov et al., 2018).

3.4 Inference

At test time CAIR slides a 576-step window over the full trace (stride 144) with cosine-window averaging and the same three refiner passes used in training (a base pass plus two conditioned on the previous estimate); predictions at observed positions are left unchanged.

4 Experiments

We ask five questions in turn. Does CAIR outperform linear interpolation under *realistic* missingness (Sec. 4.2)? Does the answer depend on gap length, and is that dependence visible in a single averaged score (Sec. 4.3–4.4)? Does the advantage survive a change of signal and clinical domain (Sec. 4.5–4.6)? Does lower reconstruction error recover the metrics clinicians read (Sec. 4.7)? And which components of the model are responsible (Sec. 4.9)? Full preprocessing, hyperparameters and protocol details are in the supplementary material.

4.1 Setup

Data. **AI-READI** (AI-READI Consortium, 2024; 2025) provides CGM, heart rate and respiration for 2,280 participants, split 1,576 train / 352 validation / 352 test. The imputation target is day two of each trace, a 24 h window on the 5-min grid. **MIMIC-III** (Johnson et al., 2016) provides intensive-care vitals; we assemble 22,156 twenty-four-hour windows on the same grid with the observed sensor masks, split subject-disjoint, and impute arterial blood pressure (ABP) and heart rate (HR).

Missingness protocols. Following Toye et al. (2025) we simulate three mechanisms (Rubin, 1976) on the target window, each at six rates from 5% to 30%: *MCAR* (independent deletion), *MAR* (contiguous windows triggered by a co-recorded covariate (time-aligned wearable activity on AI-READI, a covariate vital on MIMIC-III), and *NMAR* (windows triggered by clinically extreme values of the target itself: < 70 or > 150 mg/dL for glucose). The *gap-length protocol* carves a single contiguous gap of fixed length $L \in \{3, 6, 9, 12\}$ samples (15–60 min) at observed positions, 10 placements per participant and length, and scores only the held-out points. The *physiological protocol* masks 20% of samples in event-aligned blocks under five strategies (meal-post, sleep, ascending, dipping, combined).

Method	MCAR ↓	MAR ↓	NMAR ↓
CAIR (Ours)	**2.66**	**10.33**	**23.50**
linear interp	<u>2.91</u>	<u>12.34</u>	<u>28.94</u>
MICE	3.26	19.83	49.35
missForest	3.33	16.54	40.90
hot-deck	5.04	16.06	43.44
kNN	5.19	15.45	43.54
LOCF	6.22	19.84	34.44
Fourier	9.91	18.47	32.72
mean	28.59	28.81	54.86
GP-VAE	26.46	40.49	70.39
M-RNN	44.39	44.34	64.63

Table 1: **Realistic missingness imputation on AI-READI CGM** (all 352 test participants; RMSE mg/dL averaged over six missingness rates, lower is better). CAIR is best under all three mechanisms and its margin grows with difficulty. The two generic learned imputers fall below every non-constant baseline throughout. **Bold** = best, <u>underline</u> = second best, per column.

Baselines. We compare against twenty baselines in four families: constant and interpolation fills (linear, LOCF (Lachin, 2016), mean, mode); shape-preserving and smoothing methods (PCHIP (Fritsch & Carlson, 1980), akima (Akima, 1970), cubic spline, Savitzky–Golay (Savitzky & Golay, 1964), EWMA, a Kalman smoother (Kalman, 1960), truncated-Fourier reconstruction, bidirectional AR); tabular imputers (MICE (Van Buuren & Groothuis-Oudshoorn, 2011), missForest (Stekhoven & Bühlmann, 2012), kNN (Troyanskaya et al., 2001), hot-deck (Andridge & Little, 2010)); and learned sequence imputers (SAITS (Du et al., 2022), BRITS (Cao et al., 2018), M-RNN (Yoon et al., 2018), GP-VAE (Fortuin et al., 2020)), the last group retrained on the same windows as CAIR. All see identical masks and seeds; per-protocol subsets and settings are in the supplement.

Metrics. We report RMSE in native units (mg/dL, mmHg, bpm, breaths/min). For downstream quality we report the *metric-recovery ratio* (MRR): one minus a method's error on a clinical metric divided by the mean-fill error on that metric, so 1 is perfect recovery, 0 is no better than mean imputation, and negative values are worse. We restrict MRR to the shape and variability metrics that mean imputation fails to preserve (time in, above and below range (Battelino et al., 2019), MAGE (Service et al., 1970), coefficient of variation); on mean-preserving metrics the denominator degenerates.

4.2 Main Results

Table 1 is the main result. Under all three mechanisms CAIR attains the lowest RMSE, and its margin over linear interpolation, the method Toye et al. (2025) found strongest, *grows* with difficulty: 9% under MCAR, 16% under MAR, and 19% under NMAR, the hardest and most clinically loaded regime.

One of the most striking results is the contrast with the neural baselines. M-RNN and GP-VAE fall below every non-constant baseline under all three mechanisms, reproducing the failure Toye et al. (2025) report. Neural capacity therefore cannot be what separates CAIR from them; the difference is that CAIR is trained under a gap distribution broad enough to cover the failure modes that occur in practice (C1). Sec. 4.6 tests this claim directly by retraining the identical network on masks designed for a different signal's physiology.

4.3 Gap Length

Imputation beyond roughly one hour is not a realistic clinical target: once a gap spans an excursion, the in-gap information is not present in the endpoints. Standard CGM pipelines accordingly interpolate only gaps shorter than 30–45 min and segment the trace beyond that (Sergazinov et al., 2024). We therefore re-score every method on single contiguous gaps of 15–60 min (Table 2).

Method	15 min	30 min	45 min	60 min	**mean ↓**
CAIR (Ours)	2.84	5.03	**6.66**	**8.23**	**5.69**
akima	**2.76**	**5.01**	6.87	8.61	5.81
PCHIP	2.84	5.14	6.92	8.58	5.87
linear	3.20	5.64	7.45	9.09	6.34
SAITS	4.60	7.54	9.76	11.79	8.42
BRITS	6.25	9.42	11.67	13.33	10.17

Table 2: **Imputation performance vs. gap length** on AI-READI CGM (all 352 test participants; RMSE mg/dL). CAIR is best on the mean and at 45–60 min, and second by a small margin at 15–30 min. **Bold** = best, underline = second, per column.

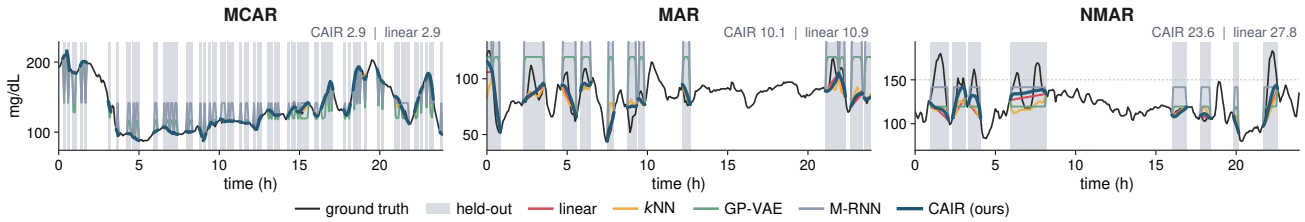

Figure 2: **Missingness mechanism examples.** (AI-READI CGM at 30% missing; five representative methods, drawn at held-out positions only). Each panel is its mechanism's median-difficulty window; insets give in-gap RMSE (mg/dL). MCAR gaps are near-trivial; NMAR puts contiguous gaps on the excursions crossing the dashed 150 mg/dL line.

CAIR is superior to the other methods on average across these gap lengths, being best at the longer intervals (45 and 60 min) and a very close runner-up at the shorter ones (15 and 30 min).

4.4 Physiological Masking

Table 3 reports the physiological five-strategy protocol over the full baseline suite. CAIR attains the best average, improving on the strongest classical baseline by 1.33 mg/dL, and is the most accurate method on the *meal*, *sleep* and *combined* strategies. The gain concentrates where a data-driven prior helps most: the long *sleep* blocks, which cost 24.9 mg/dL for CAIR against 29.3 for the strongest classical baseline.

4.5 Heart Rate and Respiration

CAIR is not bespoke to any one type of time series such as blood glucose, and is designed to be general. Accordingly, in this section and in Sec. 4.6 we evaluate it in two further physiological domains. We apply it unchanged to the heart-rate and respiration channels of AI-READI under the gap-length protocol, in two variants: *univariate*, seeing only the target channel, and *multivariate*, also conditioned on the co-recorded context the dataset provides (steps, energy expenditure, sleep state, stress, and the complementary cardiorespiratory channel).

The two variants separate sharply (Table 5). Univariate CAIR is *worse than linear* at every gap length on both signals; the multivariate variant reverses this, attaining the lowest RMSE everywhere and improving on the strongest classical baseline by 20–34% on heart rate and 15–28% on respiration. By contrast, the *same* channels leave CGM accuracy unchanged (a nine-rung ladder moves its own average only $13.11 \rightarrow 12.98$ mg/dL, inside seed noise; supplement). Glucose is autocorrelated and endogenously driven, so its own history already carries what a covariate could add; heart rate and respiration are driven by exogenous activity, so the covariate adds what the target's history lacks. This yields a general rule: *cross-modal conditioning helps exactly when the target is exogenously driven*, and requires no change to the architecture.

Method	meal	sleep	asc	dip	comb	**AVG ↓**
CAIR (Ours)	**15.25**	**24.94**	9.01	7.49	**16.37**	**14.61**
PCHIP	16.56	29.26	**7.86**	7.36	18.65	15.94
linear	16.97	29.27	8.42	7.31	18.30	16.05
akima	17.82	36.73	8.02	**7.24**	19.64	17.89
Kalman	17.97	34.70	8.88	8.09	20.29	17.99
Savitzky–Golay	17.27	29.45	13.41	11.22	19.43	18.15
SAITS	21.27	32.02	11.22	10.22	19.70	18.89
BRITS	32.72	43.61	24.25	20.81	31.45	30.57

Table 3: **Imputation of physiologically-significant missingness** on AI-READI (all 352 test participants; RMSE mg/dL, lower is better). CAIR attains the best average over the full baseline suite, and the gain concentrates in the long *sleep* blocks. The four weakest classical fills are in the supplement. **Bold** = best, underline = second best, per column.

(a) Physiological vs. random-gap training masks

Training masks	ABP RMSE ↓		
	MCAR	MAR	NMAR
linear	4.40	9.23	11.85
CAIR, CGM masks	5.47	8.74	12.10
CAIR, random-gap	**4.30**	**8.60**	**11.42**

(b) Clinical-burden recovery

Method	MRR ↑		
	MCAR	MAR	NMAR
linear	-0.31	-0.30	$+0.14$
kNN	$\mathbf{+0.79}$	$\mathbf{+0.74}$	$\mathbf{+0.65}$
CAIR (ours)	$+0.71$	$+0.64$	$+0.64$

Table 4: **Two analyses behind the MIMIC-III results** (ABP). **(a)** The *identical* architecture, trained on the same ABP data, is less accurate than linear under MCAR and NMAR when trained on masks carried over from glucose physiology, and more accurate under all three under the random-gap curriculum. **(b)** Linear is *worse than mean-fill* ($\mathrm{MRR} < 0$) under MCAR/MAR; CAIR and kNN recover most of the burden. **Bold** = best, underline = second, per column. A third analysis, in which no imputer separates from any other when predicting mortality from arterial pressure alone, is reported in Sec. 4.7.

4.6 MIMIC-III

AI-READI provides a single signal from a single sensor class. We therefore repeat the three-mechanism protocol on MIMIC-III intensive-care vitals, adapting the triggers to ICU physiology (MAR from a co-recorded covariate vital, NMAR from clinically extreme target values). These signals are smoother and more locally linear than glucose, which sets a demanding standard for linear interpolation.

However, on arterial pressure CAIR is the most accurate of every method we evaluate under every mechanism (Table 6); a paired Wilcoxon signed-rank test against linear interpolation is significant in its favor under all three ($p < 3 \times 10^{-3}$). Heart rate is a boundary case: CAIR attains the lowest mean RMSE under MAR and linear interpolation the lowest under MCAR and NMAR, but only the MCAR difference is statistically significant, the MAR and NMAR differences being ties ($p = 0.83$ and $p = 0.37$). This is consistent with a smooth signal on which interpolation is already near-optimal. Per-mechanism tests for both signals, including the heart-rate case where linear interpolation is significantly ahead, are reported in the supplement. As on CGM, M-RNN and GP-VAE fall far below every interpolant on both signals.

	Heart rate (bpm) ↓				Resp. (br/min) ↓			
Method	15	30	45	60	15	30	45	60
Akima	4.54	5.93	6.86	7.87	2.73	3.23	3.72	4.30
AR (bidir.)	4.71	5.62	6.09	6.72	2.77	3.09	3.33	3.57
PCHIP	4.45	5.62	6.27	7.05	2.63	3.07	3.43	3.76
Linear	4.40	5.52	6.12	6.86	2.61	3.00	3.34	3.65
CAIR (ours)								
univariate	4.96	6.49	6.74	7.51	2.90	3.92	4.50	5.21
multivariate	**3.50**	**4.19**	**4.29**	**4.43**	**1.88**	**2.54**	**2.62**	**2.72**

Table 5: **Cross-modal conditioning is what transfers** (AI-READI, gap-length protocol, RMSE in native units, lower is better). Column headings are gap lengths in minutes; "Resp." is respiration and "AR (bidir.)" is bidirectional AR. Univariate CAIR is less accurate than linear at every gap length on both signals; adding the co-recorded context makes it best everywhere. On glucose the same context adds nothing (Sec. 4.5). **Bold** = best, underline = second, per column.

	ABP-mean (mmHg) ↓			HR (bpm) ↓		
Method	MCAR	MAR	NMAR	MCAR	MAR	NMAR
CAIR (Ours)	**4.30**	**8.60**	**11.42**	2.26	**5.28**	6.48
linear interp	4.40	9.23	11.85	**2.22**	5.50	**6.42**
GP-VAE	11.03	12.82	17.94	9.17	14.01	18.00
M-RNN	13.45	13.19	18.35	15.28	14.53	18.60

Table 6: **Cross-domain transfer to MIMIC-III ICU vitals** (reconstruction RMSE, mean over six missingness rates, lower is better). CAIR is the most accurate method on arterial pressure under every mechanism; heart rate is a boundary case. The CAIR row is the multivariate variant; both variants, the remaining baselines and per-mechanism Wilcoxon tests are in the supplement. **Bold** = best, underline = second, per column.

The training gap distribution is what drives the result. Both rows are trained on the same ABP data with the same architecture and budget; only the training-time masking differs. Under the hand-designed glucose-physiological masks the *identical* model is *less accurate* than linear (MCAR 5.47, NMAR 12.10 mmHg); under the signal-agnostic random-gap curriculum of Sec. 3.3, specialized to no signal in particular, it is more accurate under all three (Table 4a). The masks encoding one domain's failure modes are thus what does *not* carry over, and a deliberately broad gap distribution is what lets the architecture transfer.

4.7 Downstream Clinical Metrics

Low reconstruction error does not by itself recover the metrics clinicians act on. We score clinical-burden recovery as MRR on the threshold metrics read at the bedside (time in the normal band and time above and below it) under the hardest NMAR mechanism.

These two axes split the methods into two groups, and CAIR is the only method that ranks among the best on both (full table in the supplement). Interpolants that match its RMSE *fail to preserve* the burden: linear recovers only 0.14 of it and under MCAR is *worse than mean-fill* (-0.31, Table 4b), smoothing away the excursions the threshold metrics count. Conversely, the tabular and neural imputers that match CAIR's burden recovery (kNN at 0.65, then MICE, M-RNN, GP-VAE) carry 10–60% higher RMSE, so only CAIR attains both the lowest RMSE (11.4 mmHg) and burden recovery among the best (0.64). The same ordering holds on CGM (time-in-range recovery 0.44 vs. 0.36 for linear under NMAR (Battelino et al., 2019; Service et al., 1970)) and in all four diabetes study groups (supplement), which rules out differences in cohort composition as the explanation.

Hard outcomes are insensitive to reconstruction fidelity. On stay-level outcomes the effect is mediated by the target vital, so we isolate it with a single-vital classifier. Predicting mortality from arterial pressure alone, any imputer improves on mean-fill and tracks the oracle ceiling (supplement), but the fills do not separate: with a shape-sensitive 1D-CNN readout every fill matches or *exceeds* the oracle trace (linear 0.692 vs. oracle 0.679 at 30% MCAR), because interpolation denoises the vital and a weak-signal classifier rewards smoothing. The fidelity CAIR optimizes is thus the opposite of what single-vital outcome prediction rewards, which argues that physiological imputation should be judged by reconstruction and burden recovery, not hard-outcome AUROC.

4.8 Qualitative Results

Aggregate error does not show *how* methods fail; Fig. 2 does (both rates, MIMIC-III and a difficulty sweep are in the supplement). Scattered masking is near-trivial (1–4 mg/dL), long contiguous blocks far harder (30–100 mg/dL): once a block spans an excursion, every method reverts to a near-flat fill. The methods differ in *what* they revert to: GP-VAE and M-RNN to a training-set constant, linear to the chord its endpoints imply, and CAIR toward the excursion without inventing one. CAIR therefore degrades gracefully rather than hallucinating, which is the visual basis for restricting clinical claims to short gaps.

4.9 Ablation Studies

Two choices separate CAIR from a generic masked autoencoder: the base curve is learned *and supervised on the interpolation task*, and it reaches the refiner as an additive residual rather than as conditioning alone. We remove each in turn, retrain from scratch, and score on the protocol of Table 1, so the full model and linear interpolation carry over unchanged (Table 7).

All three ablations reduce accuracy, and their ordering identifies the component that contributes most. Removing the auxiliary interpolation loss is the most expensive ($+2.74$ mg/dL on the three-mechanism mean, $+22.5\%$), the residual path next ($+1.90$), and replacing the bidirectional GRU with a self-attention block, the estimate-complete-re-estimate structure of SAITS (Du et al., 2022), the least ($+1.68$). What separates CAIR from that family is thus not that its first stage is neural, nor which sequence model implements it, but that the stage is supervised on the interpolation task itself, the property Sec. 2 identifies as absent from generic imputers: without that term the model is worse on the mean than linear (14.90 vs. 14.73).

The effect is not uniform. Under MAR the four model rows lie within 0.53 mg/dL and the attention variant is nominally ahead, so there the components are near-interchangeable; the mechanisms that separate them are MCAR, where the base curve is nearly sufficient and every ablated variant is less accurate than linear interpolation, and NMAR, where they give up 3.6–4.1 mg/dL but all remain more accurate than it. Cells are single runs, so the full row is the reference and we do not read the MAR spread as an effect. The third component reflected in the name of the method, the training curriculum, is ablated separately in Sec. 4.6: replacing the broad curriculum with masks tailored to the physiology of a source domain, architecture unchanged, makes the model less accurate than linear interpolation.

Variant	MCAR	MAR	NMAR	**ALL** ↓	Δ
CAIR (full)	**2.66**	<u>10.33</u>	**23.50**	**12.16**	n/a
w/o aux. loss	6.36	10.75	27.59	14.90	+2.74
w/o residual	4.45	10.34	27.38	14.06	+1.90
GRU → self-attn.	4.15	**10.22**	<u>27.14</u>	<u>13.84</u>	+1.68
linear interp	<u>2.91</u>	12.34	28.94	14.73	+2.57

Table 7: **Component ablation** on AI-READI CGM, on the protocol of Table 1 (all 352 test participants; RMSE mg/dL over six missingness rates). **ALL** is the mean of the three mechanisms, Δ the change against the full model. Rows are single training runs under the published recipe, differing only in the ablated component; linear is repeated from Table 1 for scale. **Bold** = best, <u>underline</u> = second best, per metric column.

5 Conclusion

We address the gap highlighted by Toye et al. (2025), where linear interpolation outperforms every learned imputer on clinical time series with realistic gaps. CAIR is a two-stage *interpolate-then-refine* model: a learned interpolator predicts a base curve inside each gap, and an iteratively unrolled Transformer refines it, with both stages trained jointly under a broad gap curriculum. The design targets the two properties of physiological missingness that generic imputers ignore: mechanism-driven gaps (C1) and gap lengths spanning orders of magnitude (C2).

CAIR is the most accurate method under every missingness mechanism on both glucose and arterial pressure, its margin growing with difficulty to 19% under value-dependent dropout. The generic learned imputers that fail in Toye et al. (2025) fail here too, placing the cause in the training curriculum rather than neural capacity. We verify this by changing only the gap distribution, which on its own removes the advantage over linear interpolation. Reconstruction accuracy and clinical-metric fidelity are distinct axes, and CAIR alone ranks among the best on both.

Acknowledgment

This research was partially funded by the National Institutes of Health (NIH) under award 1OT2OD038051. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.

References

AI-READI Consortium. AI-READI: rethinking AI data collection, preparation and sharing in diabetes research and beyond. *Nature Metabolism*, 6(12):2210–2212, 2024. doi: 10.1038/s42255-024-01165-x. https://doi.org/10.1038/s42255-024-01165-x. 3, 5, 17

AI-READI Consortium. Flagship dataset of type 2 diabetes from the AI-READI project (version 3.0.0). [Data set]. FAIRhub, 2025. https://doi.org/10.60775/fairhub.3. 5, 17

Akima, H. A new method of interpolation and smooth curve fitting based on local procedures. *Journal of the ACM (JACM)*, 17(4):589–602, 1970. 3, 6, 16

Andridge, R. R. and Little, R. J. A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64, 2010. 6, 16

Battelino, T., Danne, T., Bergenstal, R. M., Amiel, S. A., Beck, R., Biester, T., Bosi, E., Buckingham, B. A., Cefalu, W. T., Close, K. L., et al. Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. *Diabetes care*, 42(8):1593–1603, 2019. 6, 9, 19

Bent, B., Goldstein, B. A., Kibbe, W. A., and Dunn, J. P. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine*, 3(1):18, 2020. 1

Braem, C. I., Yavuz, U. S., Hermens, H. J., and Veltink, P. H. Missing data statistics provide causal insights into data loss in diabetes health monitoring by wearable sensors. *Sensors*, 24(5):1526, 2024. 1

Cao, W., Wang, D., Li, J., Zhou, H., Li, L., and Li, Y. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018. 1, 4, 6, 16

Che, Z., Purushotham, S., Cho, K., Sontag, D., and Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018. 4, 16

Cho, K., Van Merriënboer, B., Gulçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1724–1734, 2014. 3, 4

Cichosz, S. L., Kronborg, T., Hangaard, S., Vestergaard, P., and Jensen, M. H. Assessing the accuracy of continuous glucose monitoring metrics: The role of missing data and imputation strategies. *Diabetes Technology & Therapeutics*, 27(10):790–800, 2025. 1, 3, 17

De Andrade, J. B. C., de Medeiros Cavalcante, M. A. O., Lopes, T. L. M., Treigher, J. M. S., Balsells, M. D., Vasconcelos, J. L., Monteiro, L. C., and da Silveira Mota, D. D. Discovery of data quality issues in electronic health records: profound consequences for critical care medicine applications–a systematized review. *Critical Care*, 30(1):19, 2026. 1

De Boor, C. and De Boor, C. *A practical guide to splines*, volume 27. springer New York, 1978. 16

Du, W., Côté, D., and Liu, Y. Saits: Self-attention-based imputation for time series. *arXiv preprint arXiv:2202.08516*, 2022. 1, 4, 6, 10, 16

Fortuin, V., Baranchuk, D., Rätsch, G., and Mandt, S. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*, pp. 1651–1661. PMLR, 2020. 4, 6, 17

Fritsch, F. N. and Carlson, R. E. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246, 1980. 3, 6, 16

Hidalgo, J. I., Alvarado, J., Botella, M., Aramendi, A., Velasco, J. M., and Garnica, O. Hupa-ucm diabetes dataset. *Data in Brief*, 55:110559, 2024. 19

Hurst, M., O'Neill, M., Pagalan, L., Diemert, L. M., and Rosella, L. C. The impact of different imputation methods on estimates and model performance: an example using a risk prediction model for premature mortality. *Population Health Metrics*, 22(1):13, 2024. 1

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 5, 18

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. 3, 5, 17

Kalman, R. E. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960. 6, 16

Kok, N., Williamson, W. T., Lee, J. M., and Gaynanova, I. Impact of missing data and monitoring duration on downstream analyses in continuous glucose monitoring. *Diabetes Care*, 49(6):1031–1039, 2026. 1

Lachin, J. M. Fallacies of last observation carried forward analyses. *Clinical trials*, 13(2):161–168, 2016. 6, 16

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 18

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5, 18

Marling, C. and Bunescu, R. The ohiot1dm dataset for blood glucose level prediction: Update 2020. In *CEUR workshop proceedings*, volume 2675, pp. 71, 2020. 19

Poette, M., Mouysset, S., Ruiz, D., Pey, V., Alliot, J.-M., and Minville, V. Benchmarking imputation strategies for missing time-series data in critical care using real-world-inspired scenarios. *Scientific Reports*, 16(1):8116, 2026. 1, 4, 17

Pratap, A., Neto, E. C., Snyder, P., Stepnowsky, C., Elhadad, N., Grant, D., Mohebbi, M. H., Mooney, S., Suver, C., Wilbanks, J., et al. Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. *NPJ digital medicine*, 3(1):21, 2020. 1

Qian, L., Yang, Y., Du, W., Wang, J., Dobsoni, R., and Ibrahim, Z. Beyond random missingness: Clinically rethinking for healthcare time series imputation. *arXiv preprint arXiv:2405.17508*, 2024. 4, 17

Roberts, S. W. Control chart tests based on geometric moving averages. *Technometrics*, 42(1):97–101, 2000. 16

Rubin, D. B. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. 2, 4, 5, 17

Savitzky, A. and Golay, M. J. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964. 3, 6, 16

Schuster, M. and Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 3, 4

Sergazinov, R., Chun, E., Rogovchenko, V., Fernandes, N., Kasman, N., and Gaynanova, I. Glucobench: Curated list of continuous glucose monitoring datasets with prediction benchmarks. *arXiv preprint arXiv:2410.05780*, 2024. 6

Service, F. J., Molnar, G. D., Rosevear, J. W., Ackerman, E., Gatewood, L. C., and Taylor, W. F. Mean amplitude of glycemic excursions, a measure of diabetic instability. *Diabetes*, 19(9):644–655, 1970. 6, 9, 19

Shukla, S. N. and Marlin, B. Interpolation-prediction networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2019. 4

Shukla, S. N. and Marlin, B. M. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021. 4

Stekhoven, D. J. and Bühlmann, P. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012. 3, 6, 16

Tashiro, Y., Song, J., Song, Y., and Ermon, S. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021. 1, 4, 17

Toye, A. A., Celik, A., and Kleinberg, S. Benchmarking missing data imputation methods for time series using real-world test cases. *Proceedings of machine learning research*, 287:480, 2025. 2, 3, 4, 5, 6, 11, 17

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001. 6, 16

Van Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011. 3, 6, 16

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022. 1, 4, 17

Yoon, J., Zame, W. R., and Van Der Schaar, M. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 66(5): 1477–1490, 2018. 4, 6, 16

Zhao, Q., Zhu, J., Shen, X., Lin, C., Zhang, Y., Liang, Y., Cao, B., Li, J., Liu, X., Rao, W., et al. Chinese diabetes datasets for data-driven machine learning. *Scientific Data*, 10(1):35, 2023. 19

Appendix

A Extended Related Work **16**
A.1 Interpolation and smoothing . . . 16
A.2 Statistical and tabular imputation . . . 16
A.3 Learned sequence imputation . . . 16
A.4 Evaluating under realistic missingness . . . 17

B Detailed Experimental Setup **17**
B.1 Data preparation . . . 17
B.2 Missingness mechanisms . . . 17
B.3 CAIR configuration . . . 17
B.4 Baseline configuration . . . 18
B.5 Which baselines run on which protocol . . . 18
B.6 Metrics . . . 19

C External-Cohort Pretraining Does Not Improve Accuracy **19**

D Additional Results **20**
D.1 Remaining baselines on the physiological protocol . . . 20
D.2 Conditioning-modality ladder on CGM . . . 20
D.3 Complete MIMIC-III baselines, both signals . . . 20
D.4 Significance tests on MIMIC-III . . . 21
D.5 Reconstruction vs. burden on MIMIC-III ABP . . . 21
D.6 Single-vital outcome prediction . . . 22
D.7 Per-cohort stratification on CGM . . . 23

E Qualitative Galleries **23**

A Extended Related Work

A.1 Interpolation and smoothing

Deployed clinical pipelines overwhelmingly use deterministic interpolants. Linear interpolation draws a chord between the observations bracketing a gap; it is exactly reconstructive when the underlying signal is locally affine and degrades gracefully otherwise, which explains its persistent strength on smooth vitals. Shape-preserving cubics improve on it by constraining the interpolant's derivatives: PCHIP (Fritsch & Carlson, 1980) enforces monotonicity on monotone data, so it does not introduce spurious overshoot at the edges of a gap, and the Akima spline (Akima, 1970) computes slopes from a local five-point stencil, which makes it robust to outliers at the cost of second-derivative continuity. Unconstrained cubic splines (De Boor & De Boor, 1978) are smoother still but overshoot badly across long gaps, which is visible in our physiological protocol where cubic spline is the second-worst method overall. Savitzky–Golay filtering (Savitzky & Golay, 1964) fits a low-order polynomial in a sliding window by least squares and is a smoother rather than an interpolator; it is standard in CGM preprocessing. Exponentially weighted moving averages (Roberts, 2000) and Kalman smoothing (Kalman, 1960) bring an explicit state model, but both assume a stationarity that physiological signals violate across meals and sleep.

A.2 Statistical and tabular imputation

A second family treats the series as a table of correlated features. Multiple imputation by chained equations (Van Buuren & Groothuis-Oudshoorn, 2011) iteratively regresses each variable on the others and produces proper multiple imputations, so it carries uncertainty correctly under MAR. missForest (Stekhoven & Bühlmann, 2012) replaces those conditional models with random forests, which captures interactions without a parametric specification. k-nearest-neighbour imputation (Troyanskaya et al., 2001) fills a value from the most similar complete records, and hot-deck imputation (Andridge & Little, 2010) donates observed values from a matched donor rather than synthesizing them. All four recover distributional structure that interpolation discards. Their weakness on time series is the mirror image: because they do not model temporal order, their reconstructions are not smooth, which is why they score well on threshold-based burden metrics and poorly on RMSE in Table 4 of the main paper. Last-observation-carried-forward is the degenerate member of this family and its biases are well documented (Lachin, 2016).

A.3 Learned sequence imputation

Recurrent imputers exploit informative missingness directly. GRU-D (Che et al., 2018) adds a learned decay that pulls the hidden state toward the empirical mean as the time since the last observation grows. BRITS (Cao et al., 2018) imputes in both directions and penalizes disagreement between them, treating missing values as trainable variables. M-RNN (Yoon et al., 2018) combines within-stream interpolation and across-stream imputation in a multi-directional architecture. Attention-based and generative models followed. SAITS (Du et al., 2022) is the closest published architecture to CAIR. It runs a diagonally-masked self-attention (DMSA) block to obtain a first estimate, replaces the missing entries of the input with that estimate to form a completed series, passes the completed series to a second DMSA block, and blends the two blocks' outputs through a gate computed from the attention map and the missingness mask; its reconstruction loss is accumulated over the first block's output, the second block's output, and the blend, so the intermediate estimate is supervised. CAIR shares this estimate-complete-re-estimate structure and differs in every stage of it. Its first stage is a bidirectional GRU restricted to the observed values and the mask rather than a second attention block, and it is supervised at the *held-out* positions, so it is trained on the interpolation task itself rather than on reconstructing values it can already see. Its second stage receives the base curve as an additive residual as well as through the conditioning, which makes the refiner's regression target the residual by construction. And a single refiner is applied with tied weights rather than two distinct blocks, so the number of passes is a deployment choice rather than an architectural constant (we use three at training and at inference), and each pass is told, through a three-valued provenance code, which entries are observations and which are its own earlier fills. GP-VAE

(Fortuin et al., 2020) places a Gaussian-process prior over the latent trajectory of a VAE, giving calibrated uncertainty. CSDI (Tashiro et al., 2021) formulates imputation as conditional score-based diffusion, and general-purpose backbones such as TimesNet (Wu et al., 2022) treat imputation as one task among several.

These models are designed to be domain-agnostic. Both M-RNN and GP-VAE fall below every non-constant baseline on *both* of our domains, reproducing the result of Toye et al. (2025). They differ from CAIR in the gap distribution they are trained under rather than in capacity.

A.4 Evaluating under realistic missingness

Rubin's MCAR/MAR/NMAR taxonomy (Rubin, 1976) is standard in statistics but is rarely used to structure machine-learning imputation benchmarks, which typically delete values completely at random and report one averaged error. Qian et al. (2024) argue that clinically plausible missingness patterns change method rankings in healthcare time series, and Poette et al. (2026) reach a similar conclusion for critical care. Cichosz et al. (2025) show specifically that the choice of imputation strategy changes the CGM metrics clinicians read. Closest to our work, Toye et al. (2025) evaluate eleven imputers on real-world clinical signals under mechanism-driven deletion and find that linear interpolation outperforms all of them. We adopt their protocol verbatim as our hardest evaluation.

B Detailed Experimental Setup

B.1 Data preparation

AI-READI. We use the flagship type-2 diabetes release (AI-READI Consortium, 2024; 2025). CGM traces are resampled onto a uniform 5-minute grid; heart rate and respiration are resampled onto the same grid from the wearable's native sampling. Participants are split $1{,}576$ / 352 / 352 into train, validation and test by participant identifier, so no participant appears in two splits. The evaluation target is day two of each trace, indices $[288, 576)$, a 24 h window. Values are z-normalized with the training mean and standard deviation ($\mu = 132.05$, $\sigma = 42.33$ mg/dL for CGM) and all errors are rescaled to native units for reporting.

MIMIC-III. We extract $22{,}156$ twenty-four-hour windows on a 5-minute grid from the numerics of MIMIC-III (Johnson et al., 2016), retaining the observed sensor masks rather than imposing complete data. Splits are subject-disjoint. Targets are mean arterial blood pressure and heart rate; the covariate channel used for the MAR trigger is a co-recorded vital distinct from the target.

B.2 Missingness mechanisms

Each mechanism is applied to the target window at six missingness rates ($5, 10, 15, 20, 25, 30\%$), with five seeded masks per rate. *MCAR* deletes positions independently at the target rate. *MAR* deletes contiguous windows whose placement is drawn from a distribution over a co-recorded covariate: on AI-READI the time-aligned wearable activity signal, available for 303 of the 352 test participants; on MIMIC-III a co-recorded covariate vital. The target value itself never enters the trigger. *NMAR* deletes contiguous windows triggered by the target's own value crossing a clinically extreme threshold (< 70 or > 150 mg/dL for glucose; the analogous clinical bands for ABP and HR).

B.3 CAIR configuration

Stage 1 is a 4-layer bidirectional GRU with hidden size 128 and a zero-initialized linear output head. Stage 2 is an 8-layer pre-norm Transformer encoder with $d_{\text{model}} = 128$, 8 attention heads and FFN width 512, operating bidirectionally with no causal mask. Each position is encoded as the sum of a linear embedding of the observed value (a learned mask token at missing positions), learned day and time-of-day embeddings over the 288 five-minute bins of a day, an embedding of the observation state, and a linear projection of the 42-dimensional feature vector ϕ. The layout of ϕ is: diurnal history (14), window-level summary statistics (5), gap geometry (2), the two interpolant value slots and their validity

flags (4), boundary first differences (4), boundary second differences (4), least-squares boundary slopes (4), and gap context (5). Four of those dimensions form an interpolant interface: a conventional pipeline fills them with a linear and a PCHIP estimate of the missing value, each with a flag marking it present. CAIR computes neither. It writes the Stage-1 prediction y^0 into both value slots and sets both flags, so the refiner reads one learned base curve where it would otherwise read two deterministic ones.

Training uses AdamW (Loshchilov & Hutter, 2017) at learning rate 3×10^{-4} with cosine decay, batch size 64, and an exponential moving average of the weights with decay 0.999 used for inference (Izmailov et al., 2018). The refiner is unrolled for three passes with loss weights $(0.15, 0.35, 0.50)$ and the Stage-1 auxiliary loss is weighted $\lambda_{\text{aux}} = 0.7$. The random-gap curriculum holds out approximately 20% of observed samples per step as a mixture of scattered points and contiguous blocks. The block-length distribution is signal-agnostic: it is specialized to no domain in particular, and it never coincides with an evaluation mask.

At inference we slide a 576-step window with stride 144, average overlapping predictions under a cosine window, and run the same three refiner passes used in training, a base pass plus two conditioned on the previous estimate. Predictions at observed positions are left unchanged. Five seeds are averaged into a deep ensemble (Lakshminarayanan et al., 2017). Each member trains on a single NVIDIA RTX 6000 Ada GPU.

B.4 Baseline configuration

All baselines are evaluated on the identical masks and seeds as CAIR. Interpolants and smoothers use the SciPy implementations (`interp1d`, `PchipInterpolator`, `Akima1DInterpolator`, `CubicSpline`, `savgol_filter`) at their default settings; Savitzky–Golay uses a window of 31 samples and polynomial order 3; EWMA uses a smoothing factor $\alpha = 0.3$, applied in both directions and averaged where the forward and backward estimates overlap; the Kalman smoother uses a local-linear-trend state model fitted per trace. The *mode* baseline fills each gap with the most frequent observed value in the window. On the physiological protocol the *mean* fill is applied locally, over a centered 48-sample (4 h) neighborhood rather than the whole trace, falling back to the window-level mean where that neighborhood contains no observation; we write it *local mean* in Table 9 to distinguish it from the trace-level mean fill of Table 1 of the main paper. The bidirectional AR baseline fits an order-12 autoregressive model forward and backward and blends the two predictions linearly across the gap. Tabular imputers (MICE, missForest, kNN, hot-deck) treat each window as a feature vector; kNN uses $k = 5$ with a masked Euclidean metric and hot-deck is its $k = 1$ donor limit. SAITS, BRITS, M-RNN and GP-VAE are trained on the same windows and the same curriculum budget as CAIR, using the authors' published hyperparameters where available.

B.5 Which baselines run on which protocol

The four protocols score overlapping but not identical baseline subsets, and we state the mapping here. The three-mechanism protocol is the broadest and is what establishes the ordering between families: it scores the constant and interpolation fills (linear, LOCF, mean), the spectral reconstruction (Fourier), all four tabular imputers (MICE, missForest, kNN, hot-deck) and both generic learned imputers (M-RNN, GP-VAE). The MIMIC-III transfer repeats that same set on arterial pressure and heart rate (Sec. D.3). The gap-length and physiological protocols instead target the short- and structured-gap regimes, where the ordering established above makes the constant fills and tabular imputers uninformative: they score the interpolants and smoothers that are competitive there (PCHIP, akima, cubic spline, Savitzky–Golay, EWMA, Kalman, and on heart rate and respiration a bidirectional AR), together with SAITS and BRITS, the two learned imputers architecturally closest to CAIR. The rows the main paper abbreviates are reported in Sec. D.1 and Sec. D.3.

Pretrain, then fine-tune	meal	sleep	asc	dip	comb	AVG
control (AI-READI only)	**15.60**	27.52	**5.96**	**5.14**	**14.74**	**13.79**
+ HUPA-UCM (T1D)	15.99	**26.32**	6.21	5.36	15.66	13.91
+ OhioT1DM (T1D)	15.77	<u>27.14</u>	<u>6.08</u>	5.21	15.25	<u>13.89</u>
+ Shanghai (T2D)	<u>15.71</u>	27.58	6.13	<u>5.20</u>	<u>15.18</u>	13.96

Table 8: **External-cohort pretraining does not improve accuracy** (physiological protocol, 352 participants, single seed; RMSE mg/dL). Every external cell is within 0.2 mg/dL of the control and none improves on it. **Bold** = best, <u>underline</u> = second best, per column.

B.6 Metrics

RMSE is computed over held-out positions only, in native units. The metric-recovery ratio for a clinical metric g is

$$\mathrm{MRR}(g) = 1 - \frac{|g(\widehat{x}) - g(x)|}{|g(x_{\text{mean}}) - g(x)|}, \tag{B.1}$$

where x_{mean} is the mean-filled trace, so $\mathrm{MRR} = 1$ is exact recovery, 0 is no better than mean-fill, and negative values are worse than mean-fill. Clinical metrics are time-in-range (Battelino et al., 2019), MAGE (Service et al., 1970), coefficient of variation, and for MIMIC-III the fractions of time in, above and below the normal band.

C External-Cohort Pretraining Does Not Improve Accuracy

Pretraining on external CGM cohorts does not improve accuracy on the target domain. Three pools spanning both type-1 and type-2 physiology leave the physiological average within 0.2 mg/dL of a model trained on AI-READI alone, and none improves on it.

Procedure. External datasets are standardized to the AI-READI schema through a dataset-adapter registry and pooled with a domain-balanced sampler. A zero-initialized per-dataset *domain embedding* e_d is added to the Transformer's per-position conditioning. We pretrain on AI-READI ∪ external, then fine-tune on AI-READI alone with e_0 frozen at zero, so that inference is bit-identical to a model that never saw external data. Any gain must therefore survive in-domain fine-tuning.

Result. Table 8 reports the physiological protocol for three pretraining pools: HUPA-UCM (Hidalgo et al., 2024) (18 type-1 adults), OhioT1DM (Marling & Bunescu, 2020) (6 type-1 adults) and Shanghai T2DM (Zhao et al., 2023) (109 type-2 adults). Every external cell lands within 0.2 mg/dL of the in-domain control and none below it. The distribution the pool is drawn from does not change this: off-distribution (type-1) and on-distribution (type-2) pretraining behave alike, and fine-tuning on a 1,576-participant target recovers the same minimizer to within 0.2 mg/dL regardless of which pool preceded it.

The same conclusion holds on the gap-length protocol: the Shanghai-pretrained checkpoint scores 7.73 mg/dL on the short-gap mean against 7.99 for its own control, both far worse than the 5.69 of the main model, so the pretraining delta is negligible next to the effect of the training curriculum.

Why this control's average is lower than the published model's. The control row of Table 8 attains 13.79 mg/dL on the physiological average, lower than the 14.61 of the published model in Table 3 of the main paper. The two use different recipes: this experiment uses a two-stage recipe selected against the physiological average, and that average is governed by the multi-hour *sleep* blocks (26–28 mg/dL), which dwarf the short *ascending* and *dipping* segments (5–6 mg/dL). Selecting a checkpoint against it therefore rewards handling large masked fractions and over-smoothing the short gaps: the same recipe is more than 2 mg/dL worse at *every* gap length of Table 2 of the main paper (7.99 against 5.69 on the short-gap mean, above). All rows of Table 8 use this recipe, so the comparison across pretraining pools is

Method	meal	sleep	asc	dip	comb	**AVG ↓**
CAIR (Ours)	**15.25**	**24.94**	**9.01**	**7.49**	**16.37**	**14.61**
Savitzky–Golay	17.27	29.45	13.41	11.22	19.43	18.15
EWMA	20.59	31.69	21.45	18.32	23.11	23.03
local mean	22.66	32.23	24.00	20.65	24.40	24.79
cubic spline	22.64	45.22	21.90	18.62	23.85	26.44
LOCF	25.91	37.33	34.19	29.36	31.81	31.72

Table 9: **The four baselines omitted from Table 3 of the main paper** (AI-READI, physiological five-strategy protocol, all 352 test participants; RMSE mg/dL, lower is better). The lower block is the omitted group; CAIR and Savitzky–Golay are repeated from Table 3 as the best method and the weakest one carried there. Every omitted method is at least 4.9 mg/dL behind that floor on the average. **Bold** = best, underline = second best, per column.

unaffected. This is the same effect (C2) describes in the main paper: a single averaged score, dominated by an unrealistically long-gap regime, can rank a model first while it is worse in the regime clinical practice requires.

D Additional Results

D.1 Remaining baselines on the physiological protocol

Table 3 of the main paper reports the physiological five-strategy protocol over the baselines that are competitive on it. Table 9 completes that table with the four weakest, which were omitted there for space: EWMA, a local mean, an unconstrained cubic spline and LOCF. The weakest method carried in the main table, Savitzky–Golay, averages 18.15 mg/dL; the best of the four here averages 23.03 and the worst 31.72, so the gap from the main table's floor to this group (4.9 mg/dL) is larger than the gap from CAIR to that floor (3.5 mg/dL). Two failure modes separate them. The cubic spline is accurate on the short *ascending* and *dipping* segments but overshoots badly across the long *sleep* blocks (45.22 mg/dL, the worst cell in the table), which is the known cost of an unconstrained third-order fit over a wide gap and the reason shape-preserving cubics such as PCHIP and akima are preferred in deployment. LOCF is the reverse: holding the last observation is uniformly poor and degrades most on the short segments (34.19 and 29.36 mg/dL), where a fill is expected to track a rising or falling trend rather than freeze it.

D.2 Conditioning-modality ladder on CGM

Table 10 sweeps the conditioning set for CAIR on AI-READI CGM, from the target channel alone up to ten co-recorded modalities, five seeds per rung. The physiological average moves from 13.11 to 12.98 mg/dL across the entire ladder and is non-monotone, so no rung is distinguishable from the control at this seed count. This is what makes the heart-rate and respiration result in the main paper informative: the identical conditioning mechanism adds nothing on glucose and 20–34% on the exogenously driven signals, so what transfers is the mechanism's dependence on the target, not the mechanism itself.

D.3 Complete MIMIC-III baselines, both signals

Table 6 of the main paper reports CAIR, linear interpolation and the two generic learned imputers on arterial pressure and heart rate. Tables 11 and 12 complete it with both CAIR variants and the constant-fill and tabular baselines omitted there for space. On arterial pressure both CAIR variants are more accurate than every baseline under every mechanism. On heart rate CAIR and linear interpolation are separated by at most 0.22 bpm under any mechanism, so the main paper reports the signal as a boundary case. That boundary is between those two methods alone: under MAR and NMAR the closest remaining baseline, LOCF, still trails linear interpolation by 1.6 and 1.4 bpm. Under MCAR the whole

Conditioning set	meal	sleep	asc	dip	**AVG**
target only (control)	14.94	25.17	6.07	4.93	13.11
+ heart rate	14.93	25.16	6.03	4.92	13.10
+ steps, calories	14.89	25.25	6.10	4.97	13.14
+ sleep state	14.89	25.27	5.98	4.92	13.09
+ respiration, stress	14.79	24.96	6.00	4.89	**12.98**
+ environment	14.82	24.83	6.24	5.06	13.10
+ clinical	14.79	24.85	6.18	5.01	<u>13.06</u>
+ ECG	14.86	24.87	6.26	5.09	13.14
+ retinal	14.86	24.90	6.12	4.98	13.07

Table 10: Conditioning-modality ladder on AI-READI CGM (five-seed ensemble per rung; RMSE mg/dL). Adding modalities does not improve glucose imputation; the AVG spread (12.98–13.14) is within seed noise. AVG is over all five physiological strategies; the combined column is omitted for space. Absolute values are not comparable to Table 3 of the main paper. **Bold**/<u>underline</u> mark the best/second AVG.

Method	MCAR	MAR	NMAR
CAIR (univariate)	**4.26**	**8.58**	<u>11.44</u>
CAIR (multivariate)	<u>4.30</u>	<u>8.60</u>	**11.42**
linear interp	4.40	9.23	11.85
missForest	4.62	9.60	12.91
MICE	4.88	9.02	12.65
kNN	5.27	8.91	12.69
hot-deck	5.82	9.11	12.81
LOCF	5.89	10.43	12.79
Fourier	5.94	11.28	13.56
mean	9.33	9.93	14.01
mode	11.42	11.82	15.11
GP-VAE	11.03	12.82	17.94
M-RNN	13.45	13.19	18.35

Table 11: Complete MIMIC-III arterial-pressure reconstruction RMSE (mmHg, mean over six missingness rates). **Bold** = best, <u>underline</u> = second best, per column.

field is tight (missForest is within 0.15 bpm of linear), consistent with scattered single-sample deletion on a smooth signal.

D.4 Significance tests on MIMIC-III

Table 13 reports the paired Wilcoxon signed-rank test of CAIR (multivariate) against linear interpolation, computed per (window, rate) pair and reported separately for each signal and mechanism. On arterial pressure all three tests favor CAIR, the weakest at $p = 2.9 \times 10^{-3}$, which is the bound quoted in Sec. 4.6 of the main paper. On heart rate the picture is mixed and we report it in full: linear interpolation is significantly more accurate under MCAR, while the MAR and NMAR differences are not significant at any conventional level, so on those two mechanisms the two methods are statistically tied.

D.5 Reconstruction vs. burden on MIMIC-III ABP

Table 14 gives the full method-by-method breakdown of the two axes summarized in the main paper: interpolants attain low RMSE but fail to preserve the clinical burden, tabular and neural imputers recover the burden at much higher RMSE, and only CAIR occupies both corners.

Method	MCAR	MAR	NMAR
CAIR (univariate)	<u>2.26</u>	<u>5.36</u>	6.72
CAIR (multivariate)	<u>2.26</u>	**5.28**	<u>6.48</u>
linear interp	**2.22**	5.50	**6.42**
missForest	2.37	7.40	9.25
MICE	2.49	7.55	9.63
kNN	2.69	7.09	9.21
hot-deck	2.90	7.19	9.28
LOCF	3.07	7.06	7.79
Fourier	3.31	7.17	7.86
mean	8.41	9.22	11.69
mode	10.88	10.39	14.10
GP-VAE	9.17	14.01	18.00
M-RNN	15.28	14.53	18.60

Table 12: Complete MIMIC-III heart-rate reconstruction RMSE (bpm, mean over six missingness rates). The CAIR (multivariate) and linear rows are those of Table 6 of the main paper. CAIR is the most accurate method under MAR and linear interpolation under MCAR and NMAR, the two separated by at most 0.22 bpm; the per-mechanism significance tests for those differences are in Table 13. **Bold** = best, <u>underline</u> = second best, per column.

Signal	Mechanism	Δ	p
ABP-mean (mmHg)	MCAR	-0.105	7.9×10^{-9}
	MAR	-0.625	4.1×10^{-21}
	NMAR	-0.375	2.9×10^{-3}
Heart rate (bpm)	MCAR	$+0.040$	7.0×10^{-40}
	MAR	-0.220	0.83
	NMAR	$+0.060$	0.37

Table 13: **Paired Wilcoxon signed-rank tests on MIMIC-III**, CAIR (multivariate) versus linear interpolation, per (window, rate) pair. Δ is the mean of the per-pair RMSE differences in native units, which need not equal the difference of the aggregate RMSEs in Table 6 of the main paper because RMSE does not aggregate linearly; negative favors CAIR. All three arterial-pressure tests favor CAIR; on heart rate, linear interpolation is significantly better under MCAR and the remaining two differences are not significant ($\alpha = 0.05$).

D.6 Single-vital outcome prediction

Sec. 4.7 of the main paper reports that hard stay-level outcomes are insensitive to reconstruction fidelity. Table 15 reports that experiment. We isolate the effect of the fill by predicting in-hospital mortality from arterial pressure *alone*, so that the imputed channel is the classifier's only input and cannot be compensated by co-recorded vitals. The readout is a 1D-CNN over the completed trace, which is sensitive to the shape of the reconstruction rather than to summary statistics of it. The oracle row is the recorded trace with no deletion applied, and is therefore the ceiling this task can reach.

No fill is distinguishable from the oracle, and four of the five exceed it. Under MCAR every imputer improves on mean-fill, as expected, but linear interpolation and PCHIP then score 0.692 against the oracle's 0.679: a reconstruction further from the recorded signal yields a *better* classifier than the recorded signal itself. The mechanism is that interpolation denoises the vital, and a weak-signal classifier rewards smoothing; under NMAR, where the deletions sit on the clinically extreme excursions, the inversion is no smaller and becomes uniform, with every fill including mean-fill above the oracle. The fidelity CAIR optimizes is therefore orthogonal to what this task rewards. We report reconstruction error and burden

Method	RMSE (mmHg) ↓	Burden MRR ↑
Low RMSE, burden not preserved		
linear interp	11.85	+0.14
LOCF	12.79	+0.07
Fourier	13.56	+0.05
Recovers burden, high RMSE		
MICE	12.65	+0.63
kNN	12.69	**+0.65**
hot-deck	12.81	**+0.65**
GP-VAE	17.94	+0.64
M-RNN	18.35	+0.64
Both		
CAIR (univariate)	<u>11.44</u>	**+0.65**
CAIR (multivar.)	**11.42**	+0.64

Table 14: **Only CAIR ranks among the best on both axes** (MIMIC-III ABP, NMAR, mean over six rates). Reconstruction RMSE (mmHg, lower better) and clinical-burden recovery (MRR on time in/above/below range, higher better). Interpolants match the RMSE of CAIR but fail to preserve the burden; tabular and neural imputers recover the burden at much higher RMSE, with kNN and hot-deck matching its recovery at $\sim$11% higher RMSE. **Bold** = best, <u>underline</u> = second best, per column.

Fill	MCAR AUROC	NMAR AUROC
oracle (recorded trace)	0.679	0.678
mean-fill	0.677	0.683
kNN	0.680	0.685
CAIR (multivar.)	0.688	0.688
linear interp	0.692	0.688
PCHIP	0.692	0.689

Table 15: **Hard outcomes do not reward reconstruction fidelity** (MIMIC-III, in-hospital mortality predicted from arterial pressure alone, 30% missingness, 1D-CNN readout; $n = 1200$ under MCAR and 1181 under NMAR). The oracle is the recorded trace and is the ceiling for this task, yet four fills exceed it under MCAR and all five do under NMAR, because interpolation denoises the vital. No column is ordered as reconstruction accuracy would predict, and no best value is marked.

recovery (Table 14) rather than hard-outcome AUROC. MAR is omitted because the outcome experiment was run under MCAR and NMAR only.

D.7 Per-cohort stratification on CGM

Under NMAR, CAIR attains both lower RMSE and higher time-in-range recovery than linear interpolation in all four AI-READI diabetes study groups (Table 16). The advantage is present in the healthy group, where absolute error is lowest, and largest in the insulin-dependent group, where the excursions that NMAR deletes are most frequent. AI-READI de-identifies gender, so study group is the available and more physiologically relevant stratification axis.

E Qualitative Galleries

This section extends Fig. 2 of the main paper. Every trace is read from the committed evaluation caches; no panel re-runs a model. Predictions are drawn at held-out positions only and joined to the two observed samples bracketing each gap, so a method's curve inside a band is exactly what it contributed to that

	RMSE (mg/dL) ↓		TIR recovery ↑	
Study group	CAIR	linear	CAIR	linear
healthy	**18.29**	21.49	**0.165**	0.097
pre-diabetes	**15.90**	21.15	**0.467**	0.382
oral medication	**33.96**	40.19	**0.601**	0.556
insulin-dependent	**27.23**	35.17	**0.608**	0.489

Table 16: **The CGM result holds in every diabetes study group** (AI-READI, NMAR, all 352 test participants). CAIR attains lower reconstruction error and higher time-in-range recovery than linear interpolation in all four groups. **Bold** = better of the two, per group and metric.

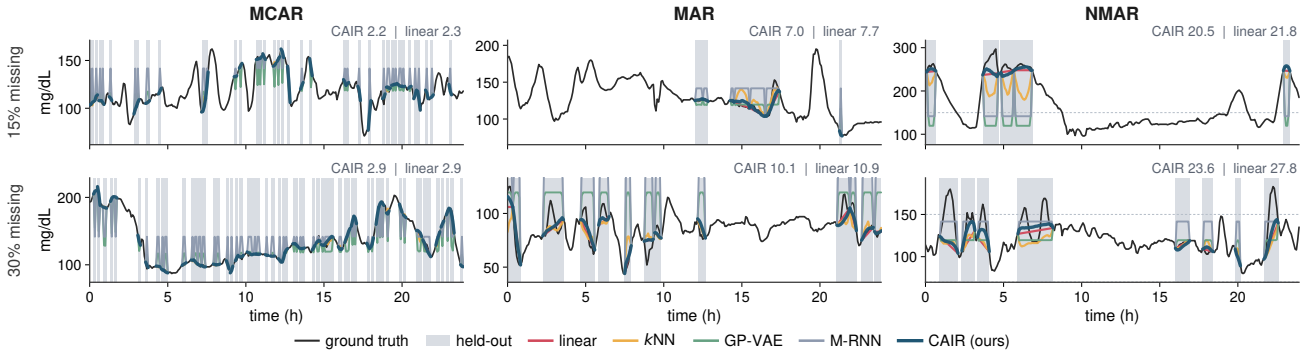

Figure 3: **AI-READI CGM, both missingness rates** (columns: mechanism; rows: rate). The main paper shows the 30% row. Raising the rate from 15% to 30% lengthens and multiplies the blocks but does not change the ordering: CAIR and linear track the trace under MCAR, separate under MAR, and separate furthest under NMAR, where the dashed 70/150 mg/dL thresholds mark the excursions that trigger the deletion. GP-VAE and M-RNN revert to a near-constant fill under every mechanism, which is what puts them below every non-constant baseline in Table 1 of the main paper.

window's in-gap RMSE.

How the windows are chosen. Panels are selected by a fixed rule rather than by eye. Within a (mechanism, rate) cell we first discard windows whose gap count falls outside $[0.5\times, 2\times]$ the cell's median gap count, which removes masks that are structurally atypical for that mechanism; among the rest we take the window at the median of CAIR's in-gap RMSE. The selection does not consult the baselines. The resulting panels have CAIR-to-linear error ratios of 0.85–1.01, against 0.81–0.91 for the mechanism-level aggregates in Table 1 of the main paper. Fig. 5 is the one deliberate exception, sweeping the same cell from its easiest to its hardest window.

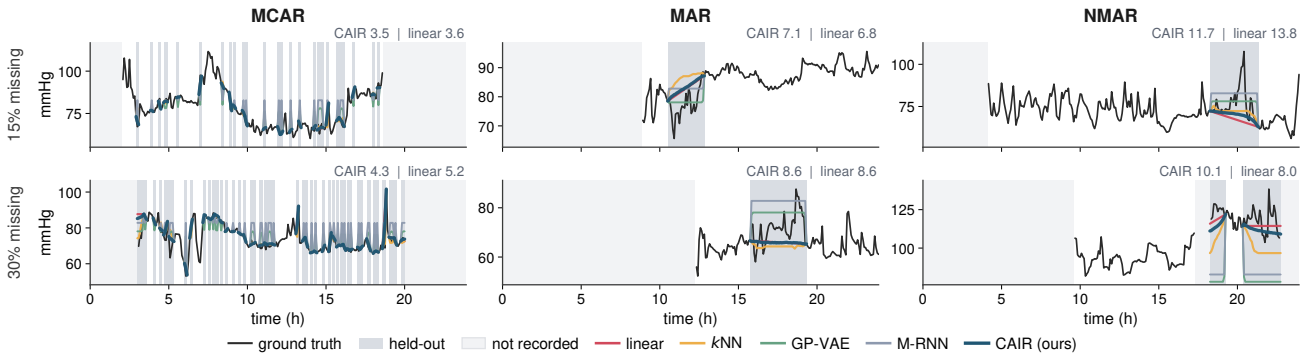

Figure 4: **MIMIC-III arterial pressure, same protocol.** Pale regions marked *not recorded* are positions the ICU monitor never sampled (23% of ABP positions); they are neither observed nor scored, and no ground truth is drawn there. ABP is smoother and more locally linear than glucose, so the per-window gap between CAIR and linear is small and its sign varies from window to window (three of these six median-difficulty panels favour CAIR, two favour linear and one is a tie), while the six-rate aggregate in Table 6 of the main paper favours CAIR under all three mechanisms. The neural baselines fail here exactly as they do on CGM.

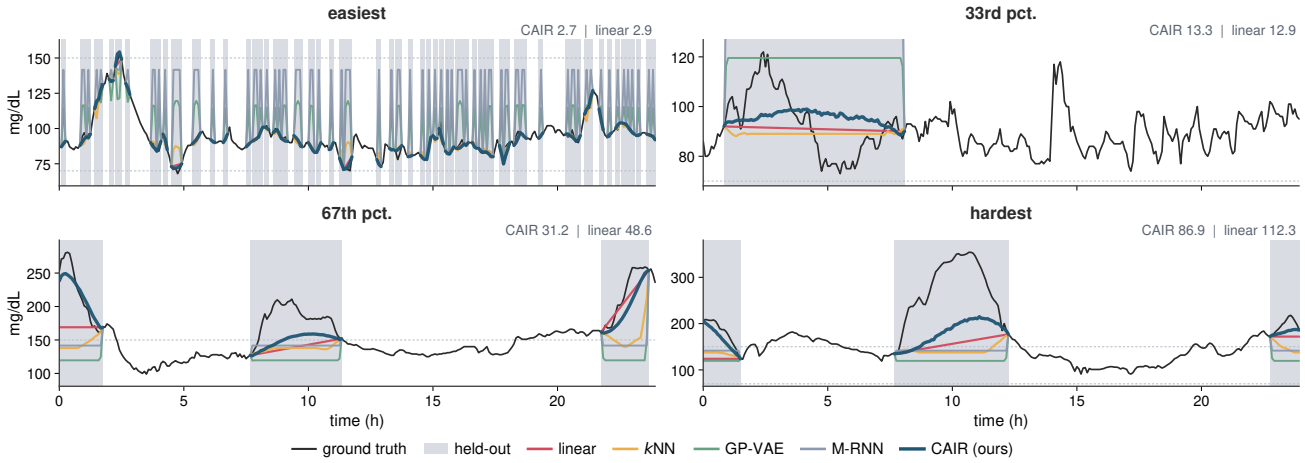

Figure 5: **Easiest to hardest, at fixed mechanism and rate** (AI-READI CGM, NMAR, 30%; panels at the 0th, 33rd, 67th and 100th percentile of CAIR's in-gap RMSE). These panels support the claim in Sec. 4.8 of the main paper that CAIR degrades gracefully. As the blocks lengthen and swallow whole excursions, no method recovers the excursion, but the failures differ in kind: GP-VAE and M-RNN sit at a constant fixed by the training distribution, linear draws the chord its endpoints imply, and CAIR bends part of the way toward the excursion and stops. In the hardest panel the recorded in-gap peak is 354 mg/dL; CAIR reaches 215 and linear 176. Under-shooting is the safe direction of error for a clinical burden metric, since an invented excursion would create a treatment signal that never occurred.