DISTRIBUTION RECOVERY WITH DOMAIN CONSTRAINTS: APPLICATIONS TO MEDICAL TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering the occurrence and severity of physiological abnormalities from intermittent clinical measurements requires a distribution over possible trajectories, yet sparse observations may not uniquely determine this distribution. SFBD-OMNI guides recovery with clean reference sample anchors, which are generally unavailable in retrospective medical records. In this work, we introduce Projected SFBD, which directly incorporates domain knowledge as constraints on recovered trajectories through a projection-based update. We provide theoretical convergence guarantees and show how physiological priors narrow the space of admissible distributions and can improve identifiability without clean reference samples. On irregular MIMIC-III charts, evaluation against bedside-monitor references shows better trajectory-distribution scores and more accurate recovery of hypotension occurrence, severity, recurrence, and within-stay variation in vital signs than strong baselines. Ablations on MIMIC-III and the Pulse simulator show that projection better preserves clinically relevant variation and yields less biased clinical summaries than constructed clean anchors.

1 INTRODUCTION

Clinical records contain intermittent measurements of vital signs and laboratory values, leaving much of a patient's physiological trajectory unobserved (Che et al., 2018). Measurement times differ across channels and may depend on the patient's condition (Agniel et al., 2018). These gaps obscure whether an abnormal event occurred, how severe it became, and how often it recurred; intermittent blood-pressure monitoring, for example, can miss hypotensive episodes (Wijnberge et al., 2022). A single smooth completion can also fail to preserve clinically informative fluctuations (Xu et al., 2022). We therefore need to seek a distribution over trajectories consistent with the observations, allowing us to estimate whether an event occurred and how severe it was while accounting for uncertainty in the missing values (Fortuin et al., 2020).

The central challenge is that sparse observations do not uniquely determine the distribution of complete trajectories. Two trajectory distributions may agree on the recorded measurements while differing substantially between them, for example, in the frequency and severity of unobserved physiological events. SFBD-OMNI resolves this ambiguity using clean reference samples (Lu et al., 2026b). In retrospective clinical data, however, complete trajectories are rarely available. Restricting the reference set to unusually complete records can change the patient population (Weber et al., 2017), while filling the gaps by interpolation turns an assumed smooth trajectory into a training target and can suppress the fluctuations we aim to recover (Xu et al., 2022).

In this work, we propose to tackle this dilemma by introducing Projected SFBD, an extension of SFBD-OMNI that incorporates domain knowledge through constraints on recovered trajectories. Physiological bounds restrict admissible values, while cross-channel relationships enforce conditions such as systolic pressure remaining no lower than diastolic pressure at the same time. At each update, the method projects proposed completions onto trajectories that satisfy these constraints while preserving the recorded measurements. This replaces clean-anchor regularization with an explicit restriction on admissible trajectories, a capability absent from the original SFBD-OMNI update. Under weak assumptions, we prove convergence of the adaptive projected updates and establish an $O(1/K)$ convergence rate for the observed-data KL objective. By narrowing the family of distributions consistent with the observations, domain constraints can also improve identifiability:

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.

when the target is the only admissible distribution compatible with the observations, it can be recovered. Although motivated by medical time series, the framework and guarantees apply to general observation and corruption processes, including irregular sampling and masking.

We evaluate Projected SFBD on irregular MIMIC-III charts paired with separately recorded bedside-monitor references (Johnson et al., 2016; Moody et al., 2020), which allow recovery to be assessed at times without a recorded chart measurement. In addition to evaluating the recovered trajectory distribution, we examine downstream clinical tasks that depend on its temporal structure. The results show that Projected SFBD improves both distributional scores and performance on these clinical tasks over strong baselines. Comparisons with SFBD-OMNI's clean-anchor regularization on MIMIC-III and controlled Pulse simulator experiments (Bray et al., 2019) show that constraint projection better preserves clinically relevant variation and reduces bias in clinical summaries.

2 Related Work

Imputation is a common first step in clinical time-series analysis, and existing imputers fall into four broad families. The simplest are heuristic rules such as mean or median substitution, forward fill, and linear interpolation, which are common baselines in clinical benchmarks. Classical statistical imputers rely on classical learning methods: MICE imputes each incomplete variable by regressing it on the others, cycling through the variables over several iterations (van Buuren & Groothuis-Oudshoorn, 2011). MissForest applies the same iterative scheme with random forests in place of the regressions (Stekhoven & Bühlmann, 2012), while k-nearest-neighbour imputation fills each gap by averaging values from the most similar records (Troyanskaya et al., 2001). Deterministic deep-learning imputers learn a single completion from data: BRITS uses a bidirectional recurrent network (Cao et al., 2018), while SAITS uses self-attention for time-series imputation (Du et al., 2023). Finally, generative imputers return samples from a distribution rather than a point estimate: GP-VAE uses a variational autoencoder to encode incomplete trajectories into a latent space with a Gaussian-process prior and decode them into complete ones (Fortuin et al., 2020), and CSDI imputes missing values by reversing a diffusion process conditioned on the observed entries (Tashiro et al., 2021).

A separate line of work recovers a clean data distribution mainly from corrupted samples. SFBD alternates between denoising Gaussian-corrupted samples and retraining on the denoised outputs, and shows that pretraining on a small clean set lets it recover the underlying clean data distribution (Lu et al., 2025). SFBD-OMNI extends the framework to arbitrary corruptions with bridge models and shows that a few clean samples can resolve cases where the corruption is not identifiable (Lu et al., 2026b). We build on SFBD-OMNI, interpreting missing data as a form of corruption, but replace these clean samples, which are impossible to obtain in a medical setting, with constraints on admissible trajectories. A fuller comparison with ambient diffusion, EM-style learning from corrupted observations, and diffusion-based inverse-problem solvers is provided in Section A.

3 Preliminaries

Consider a problem that recovers the distribution p_{data} from corrupted observations. For a known observation kernel $r(dy \mid f)$, operator T_r maps a candidate clean law p to its observation distribution,

$$(T_r p)(dy) = \int_{\mathcal{H}} r(dy \mid f)\, p(df). \tag{1}$$

A natural recovery objective is

$$\min_{p \in \mathfrak{P}} D_{\mathrm{KL}}\left(q \parallel T_r p\right), \tag{2}$$

where $q = T_r p_{\mathrm{data}}$ is the observed distribution and $\mathfrak{P}$ is a class of probability laws on the clean-sample space $\mathcal{H}$ containing p_{data}.

SFBD-OMNI. Following Lu et al. (2026b), when the observation kernel admits a likelihood density, we can rewrite Eq. (2), up to an additive constant, as

$$\min_{p \in \mathfrak{P}} \min_{\{u_y\}} \mathcal{F}(p, u_y), \qquad \mathcal{F}(p, u_y) := \mathbb{E}_{y \sim q}\Big[D_{\mathrm{KL}}\left(u_y \parallel p\right) - \mathbb{E}_{f \sim u_y}[\log r(y \mid f)]\Big]. \tag{3}$$

Here u_y is a candidate conditional distribution of the clean sample f given y, and the inner minimization ranges over these distributions for all y. Alternating minimization gives two updates. With

p^k fixed, minimizing over the conditional distributions gives the posterior u_y^k of the clean sample given y under p^k: $u_y^k(df) = \frac{r(y|f)\,p^k(df)}{(T_r p^k)(y)}$. Using the diffusion training pipeline, it is easy to learn a neural-based bridge approximation $\hat{u}_y^k$ to u_y^k from pairs generated using p^k and r. With u_y^k fixed, minimizing over p aggregates these posteriors over the observed data, provided the aggregate belongs to $\mathfrak{P}$:

$$p^{k+1} = \arg\min_{p\in\mathfrak{P}} \mathcal{F}(p, u_y^k) = \rho^k, \qquad \rho^k(df) = \int u_y^k(df)\, q(dy). \tag{4}$$

Sampling this model at observed records supplies the next clean-sample distribution. Its online variant replaces only a fraction $\gamma \in (0, 1]$ of the stored completions at each update:

$$p^{k+1} = (1-\gamma)p^k + \gamma\rho^k. \tag{5}$$

Replacing a smaller fraction lets the same network track the evolving distribution; $\gamma = 1$ recovers the full update. Matching q identifies p_{data} only when no other candidate induces the same observed distribution. Under their stated assumptions and with exact posterior updates, Lu et al. (2026b) prove convergence to an optimum for every fixed $\gamma \in (0, 1]$, yielding recovery of p_{data} when it is identifiable.

SFBD Flow. For Gaussian corruption, Lu et al. (2026a) interpret the relaxed update $p^{k+1} = (1-\gamma)p^k + \gamma\rho^k$ as a discretization of a continuous flow as $\gamma \to 0$. Their analysis also determines how the fraction of stored completions replaced at each update can adapt to errors in the learned reverse process:

$$v_{k+1} = \beta v_k + (1-\beta)\delta_k^2, \qquad \gamma_k = \min\left\{\frac{\gamma_{\text{base}}}{\sqrt{v_{k+1}}}, \gamma_{\max}\right\}, \tag{6}$$

with $\beta, \gamma_{\max} \in (0, 1)$ and $\gamma_{\text{base}}, v_0 > 0$. Let u_y^{data} denote the true posterior. They use a nonnegative bound δ_k on the learned posterior's excess KL error relative to the exact posterior u_y^k, defined by

$$\int \left[D_{\text{KL}}\left(u_y^{\text{data}} \,\|\, \hat{u}_y^k\right) - D_{\text{KL}}\left(u_y^{\text{data}} \,\|\, u_y^k\right)\right] q(dy) \le \delta_k, \qquad \delta_k \ge 0. \tag{7}$$

Larger error bounds reduce γ_k in Eq. (6); as the smoothed error decreases, γ_k increases up to $\gamma_{\max}$. In this work, we adapt this error control to trajectory posteriors and prove convergence guarantees for general corruption processes.

4 Prior-Constrained SFBD-OMNI

When multiple clean distributions explain the same observations, SFBD-OMNI uses a clean anchor to guide recovery. In medical records, however, complete trajectories are rarely available, and selecting densely observed patients can introduce cohort bias (Agniel et al., 2018; Weber et al., 2017). Such an anchor may therefore steer recovery toward an unrepresentative population. Alternatively, medical knowledge offers another source of information: physiological bounds and relations between measurements restrict plausible trajectories without requiring clean examples. In this section, we present an alternative method that uses these constraints directly.

4.1 Recovery with Structured Priors

For medical time series, we specialize $\mathcal{H}$ to a separable Hilbert space of continuous trajectories with continuous point evaluations. A clean sample is a patient's multichannel trajectory $f = (f_1, \ldots, f_D) : [0, \tau] \to \mathbb{R}^D$. An observed record contains only sampled values:

$$y = (y_c)_{c=1}^D, \qquad y_c = ((t_{c,i}, v_{c,i}))_{i=1}^{n_c}, \qquad v_{c,i} = f_c(t_{c,i}). \tag{8}$$

For each channel c, $0 \le t_{c,1} < \cdots < t_{c,n_c} \le \tau$. Measurement times may depend on the patient, channel, and underlying trajectory; channels need not be observed at the same times. The observation kernel $r(dy \mid f)$ models measurement values and their timestamps, including any dependence of measurement timing on f.

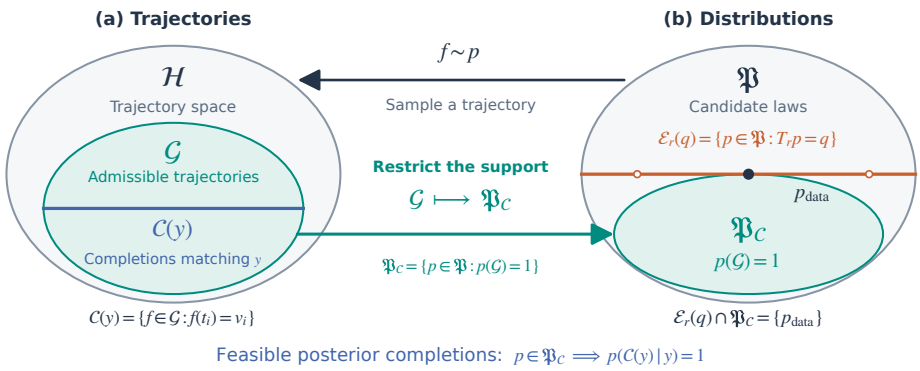


Figure 1: Trajectory constraints and distribution recovery. The unique intersection identifies p_{data}. Assuming the observed values satisfy the constraints, feasible posterior completions fill in unobserved values, preserve each observed value at its recorded time and obey the medical constraints.

Medical constraints define an admissible trajectory set $\mathcal{G} \subseteq \mathcal{H}$, incorporating physiological bounds and cross-channel relationships, such as systolic pressure being no lower than diastolic pressure at the same time. For a record y, feasible completions must also preserve its measurements:

$$\mathcal{C}(y) = \{f \in \mathcal{G} : f_c(t_{c,i}) = v_{c,i},\ c = 1, \ldots, D,\ i = 1, \ldots, n_c\}. \tag{9}$$

We assume $\mathcal{G}$ is closed and convex and $\mathcal{C}(y)$ is nonempty for the records being completed. That is, the observed values are consistent with the constraints under consideration and admit at least one feasible completion. For instance, a completed trajectory may be required to stay within prescribed physiological ranges and keep systolic blood pressure no lower than diastolic pressure, while preserving the recorded measurements. This formulation uses exact observations; measurement noise requires corresponding tolerances in the consistency constraints.

Restricting candidate laws to admissible trajectories gives the recovery problem

$$\min_{p \in \mathfrak{P}_\mathcal{C}} D_{\text{KL}}\left(q \parallel T_r p\right), \tag{10}$$

where $\mathfrak{P}_\mathcal{C} = \{p \in \mathfrak{P} : p(\mathcal{G}) = 1\}$ is the family of laws supported on admissible trajectories.

The objective still fits the observed distribution; the prior restricts which clean laws can explain it. Writing $\mathcal{E}_r(q) = \{p : T_r p = q\}$, the condition needed to resolve ambiguity is

Definition 1 (Prior-Restricted Identifiability)**.** *The target is identifiable under the prior when*

$$\mathcal{E}_r(q) \cap \mathfrak{P}_\mathcal{C} = \{p_{\text{data}}\}. \tag{11}$$

Thus the prior must leave exactly one admissible explanation of q; it need not make T_r injective over all clean laws. Fig. 1 summarizes these sets and the support condition linking trajectories to distributions.

4.2 Posterior Updates with Feasibility Projection

The left panel of Alg. 1 follows the update rule in Eq. (4), using its relaxed version in Eq. (5) during sampling. To address the identifiability problem when multiple clean distributions explain the same observations, Lu et al. (2026b) propose adding a clean anchor, giving the regularized objective

$$\min_{p \in \mathfrak{P}} D_{\text{KL}}\left(q \parallel T_r p\right) + \lambda D_{\text{KL}}\left(h_{\text{anchor}} \parallel p\right). \tag{12}$$

In practice, h_{anchor} is represented by a limited set of clean samples, which may not faithfully represent the true patient distribution; $\lambda \geq 0$ controls its influence. In the boxed *Mix with clean anchors* step, each training sample comes from this anchor set with probability $\lambda/(1+\lambda)$ and otherwise from the stored set of current completed trajectories. An unrepresentative anchor can therefore bias the recovered distribution, an unacceptable outcome when it distorts clinically meaningful conclusions. This claim is corroborated by our experiments in Section 5.

Algorithm 1 Practical SFBD-OMNI and constrained SFBD. Boxes highlight clean-anchor mixing and feasibility projection. The SFBD-OMNI loop follows Lu et al. (2026b, Algorithm 2); Constrained SFBD sets the fraction γ_k of stored completions replaced at each update following Eq. (6), with nonnegative error bounds δ_k satisfying Eq. (7).

Shared inputs: N observed records $\{y_j\}_{j=1}^N$, corruption sampler r, conditional model u_θ, training steps K. The stored set of current completed trajectories is $B = (B_1, \ldots, B_N)$, with one completion B_j per observed record y_j.

SFBD-OMNI (online)

Require: Clean samples A; weight λ; replacement fraction γ

Pretrain u_θ on A after re-corruption
Initialize $B_j \sim u_\theta(\cdot \mid y_j)$
for $k = 1, \ldots, K$ **do**
  Mix with clean anchors
  Sample f from A with probability $\lambda/(1+\lambda)$; otherwise from B.
  Draw $\tilde{y} \sim r(\cdot \mid f)$; train u_θ on $(f, \tilde{y})$
  if replacement step **then**
    Select indices J uniformly, $|J| \approx \gamma N$
    Resample $B_j \sim u_\theta(\cdot \mid y_j)$, $j \in J$
  end if
end for
return u_θ, B

Constrained SFBD (ours)

Require: Medical constraints; adaptive settings for Eq. (6)

Initialize B by forward filling records
for $k = 1, \ldots, K$ **do**
  Sample f from B
  Draw $\tilde{y} \sim r(\cdot \mid f)$; train u_θ on $(f, \tilde{y})$
  if replacement step **then**
    Choose δ_k satisfying Eq. (7)
    Update γ_k using Eq. (6)
    Select indices J uniformly, $|J| \approx \gamma_k N$
    Draw $\hat{f}_j \sim u_\theta(\cdot \mid y_j)$, $j \in J$
    Feasibility projection
    $B_j \leftarrow \Pi_{y_j}(\hat{f}_j)$, $j \in J$
  end if
end for
return u_θ, B

In contrast, our method replaces anchor mixing with the boxed feasibility projection step, projecting each generated trajectory $\hat{f}$ onto the constraint set $\mathcal{C}(y)$ defined by medical domain knowledge and the observed record y:

$$\Pi_y(\hat{f}) = \arg\min_{g \in \mathcal{C}(y)} \tfrac{1}{2}\|g - \hat{f}\|_{\mathcal{H}}^2. \tag{13}$$

Closedness, convexity, and nonemptiness give a unique projection. It preserves recorded values, enforces the trajectory constraints, and leaves feasible completions unchanged:

$$\hat{f} \in \mathcal{C}(y) \implies R(\hat{f}; y) = \Pi_y(\hat{f}) = \hat{f}. \tag{14}$$

Projection therefore excludes trajectories that violate the prior while leaving true trajectories unchanged when the prior is valid. The posterior updates fit the observed distribution, and the prior restricts which clean distributions can explain it. If p_{data} is the only admissible explanation, these two requirements identify the target. The analysis below establishes convergence under conditions that ensure target preservation and control posterior approximation error. To analyze the update at the level of probability distributions, let $\hat{u}_y^k$ be the learned posterior and γ_k the fraction of stored completions replaced at update k. The corresponding idealized updates are

$$\hat{\rho}^k = \int \hat{u}_y^k \, q(dy), \qquad \bar{p}^{k+1} = (1-\gamma_k)p^k + \gamma_k \hat{\rho}^k, \qquad p^{k+1} = R_{\#}\bar{p}^{k+1}. \tag{15}$$

Here $R_{\#}$ denotes the distribution after projection. Assuming p_{data} is a fixed point, $R_{\#}p_{\text{data}} = p_{\text{data}}$, the data-processing inequality (Kullback & Leibler, 1951, Theorem 4.1) gives

$$D_{\text{KL}}\left(p_{\text{data}} \parallel R_{\#}p\right) = D_{\text{KL}}\left(R_{\#}p_{\text{data}} \parallel R_{\#}p\right) \le D_{\text{KL}}\left(p_{\text{data}} \parallel p\right).$$

Thus projection cannot move an iterate farther from p_{data} in KL divergence and preserves the descent bound underlying SFBD-OMNI's convergence analysis. The formal conditions are stated in Assumption 2. Feasible stored completions remain unchanged, so only replacements need projection. Exact constraint projection gives

$$p^k \in \mathfrak{P}_{\mathcal{C}}, \qquad k \ge 1. \tag{16}$$

For the N observed records $\{y_j\}_{j=1}^N$, $B = (B_1, \ldots, B_N)$ is the stored set of current completed trajectories, with one completion B_j per record y_j, so its size is also N. We choose the fraction γ_k of

stored completions replaced at update k according to Eq. (6), where $\delta_k \geq 0$ satisfies Eq. (7). Algorithm 1 then selects approximately $\gamma_k N$ records uniformly at random and replaces their completions after projection.

Practical replacement-rate control. We compare proposed completions with the stored set of current completed trajectories using an unbiased estimator of squared maximum mean discrepancy (MMD), excluding self-pairs to remove finite-sample estimator bias (Gretton et al., 2012). Even two random groups drawn from the same stored set can differ because they contain different samples. We therefore also compare two disjoint subsets of stored completions, matching the sample sizes of the comparison between proposals and stored completions, to estimate the discrepancy caused by sampling alone. We square the MMD^2 estimates and smooth them over updates using exponential moving averages: v_k tracks the comparison between proposals and stored completions, while n_k tracks the comparison between two subsets of stored completions. When v_k is large relative to this sampling baseline n_k, we accept fewer replacements. Among $|J|$ uniformly proposed replacements, each is accepted with probability a_k, giving

$$\gamma_k = \frac{|J|}{2N} \min\left\{1, \sqrt{n_k / v_k}\right\}. \tag{17}$$

Here $|J|/N$ is the fraction of the N stored completions proposed for replacement, and a_k is their acceptance probability, so γ_k is the expected fraction actually replaced; large discrepancy spikes skip the update. This empirical diagnostic does not certify the KL-error bound in Eq. (7). Estimator and smoothing settings are specified in Section C.2; the convergence result below uses the stated posterior-error and projection assumptions.

4.3 From Feasible Updates to Distribution Recovery

Theorem 1 establishes convergence of the projected updates under regularity and posterior-error conditions, with clean-distribution recovery when the prior resolves ambiguity. The detailed conditions and proof are given in Section B.5.

Theorem 1. *Suppose the posterior-error bounds δ_k in Eq. (7) satisfy $\sum_k \delta_k < \infty$, and use the projected updates in Eq. (15) with γ_k chosen by Eq. (6). Under the weak conditions in Section B.5,*

$$D_{\mathrm{KL}}\left(q \parallel T_r p^k\right) \longrightarrow 0, \qquad \min_{0 \leq k < K} D_{\mathrm{KL}}\left(q \parallel T_r p^k\right) = O(K^{-1}).$$

With perfect posterior estimation ($\delta_k = 0$ for all k), the divergence decreases monotonically from the first feasible iterate. In both cases, p^k converges to the solution set $\mathcal{E}_r(q) \cap \mathfrak{P}_{\mathcal{C}}$. When the prior leaves only one clean distribution consistent with the observations, p^k converges to p_{data}.

Remark 1. *Here, data sampling or acquisition is viewed as a corruption process: it transforms a clean sample into an observed record according to $r(dy \mid f)$. The guarantees above apply to general known transformations satisfying the stated assumptions, including masking and irregular sampling. The proof relies on the KL chain rule and data-processing inequality, so medical trajectory recovery is one application of the same argument with domain-specific prior constraints.*

5 Experiments

In this section, we evaluate whether Projected SFBD recovers trajectory distributions and preserves clinically important events from sparse, irregular medical observations. We assess distributional recovery, hypotension detection, severity and recurrence, and within-stay variation in vital signs, then compare constraint projection with unconstrained and clean-anchor alternatives.

Experiment setup and metrics. We assess recovery under real-world missingness by linking MIMIC-III clinical charts to separately archived numerical vital-sign measurements recorded by bedside monitors in the MIMIC-III Waveform Database Matched Subset using patient identifiers and aligned timestamps (Johnson et al., 2016; Moody et al., 2020). Because bedside monitors record vital signs more frequently than clinical charts, they provide reference measurements at times without a recorded chart measurement. Monitor measurements are excluded from both training targets and conditioning inputs. We complement this real-data evaluation with the Pulse physiology simulator (Bray et al., 2019), whose complete trajectories permit a controlled comparison of constraint projection with constructed and oracle clean anchors. Cohort and protocol details are provided in Section C.

We use standard metrics in medical time-series analysis for both point-estimation and distribution-level evaluation. Table 6 in the appendix explains what each metric measures and how to interpret it. Formal definitions, finite-ensemble estimators, and aggregation conventions are given in Section C.3.

Models. Following function-space diffusion formulations (Kerrigan et al., 2023; Lim et al., 2025), we use temporally correlated Gaussian noise with a trace-class covariance so that perturbations remain in the trajectory space. This is the default for our real-data experiments unless otherwise stated. We choose radial basis function (RBF) kernels for their smooth correlations with explicitly controlled time scales (Rasmussen & Williams, 2006). Mixing short and long scales allows the noise prior to represent both rapid fluctuations and sustained changes. We observe little change in recovery performance when replacing the default mixture weights with weights fitted to observed-chart autocorrelation; see Section D.2 for details and further scale and weight comparisons. For re-corruption, we independently retain values at rates estimated separately for each channel. Architecture, time encoding, selection of the time–channel locations where the model generates values, and projection details are given in Section C.2.

5.1 Main Results

We compare Projected SFBD with CSDI, a conditional diffusion model generating multiple imputation samples (Tashiro et al., 2021), and three point-estimation baselines: SAITS, a self-attention model for time-series imputation (Du et al., 2023); linear interpolation between observations, an established physiological-signal imputation baseline (Xu et al., 2022); and forward-fill (last observation carried forward), used in the MIMIC-III clinical prediction benchmark (Harutyunyan et al., 2019). Forward-fill imputes each missing value using the most recent measurement of the same channel, if a previous measurement exists. Baseline configurations and evaluation details are in Section C.4.

Direct distribution recovery. For the MIMIC-III comparison, all methods use the same patients, timestamps, channels, and fixed-length bedside-monitor reference windows (see Section C). As shown in Table 1, Projected SFBD achieves the lowest energy score (ES) and variogram score (VS) across all three window lengths. ES evaluates the joint predictive distribution over a trajectory window, while VS focuses on changes between its time points. Together, the scores indicate that Projected SFBD better recovers both the joint distribution within each window and its temporal dependence. This matters for medical data, where matching typical values alone does not ensure that generated trajectories reproduce clinically relevant fluctuations.

Downstream clinical estimation. We assess whether recovered trajectories preserve clinically relevant events and patient-level summaries (Table 2). Sustained low blood pressure can compromise blood flow to vital organs, making its detection and severity clinically important. We define a hypotensive episode as mean arterial pressure (MAP) below 65 mmHg for at least ten minutes. The pressure threshold is clinically motivated by the recommended initial MAP target of 65 mmHg for adults with septic shock receiving vasopressors (Evans et al., 2021). Our ten-minute criterion focuses the evaluation on sustained events rather than isolated low readings. We calculate clinical summaries, such as episode counts, separately for each generated trajectory and then average the results, avoiding event loss from averaging trajectories first.

Detection. In this task, we evaluate whether the recovered trajectories reveal if a patient experienced any hypotensive episode during their stay. Each generated trajectory represents a possible completion of the sparse observations; the fraction containing an episode gives an estimated probability that hypotension occurred. We compare these probabilities with episode labels obtained from the monitor reference. Projected SFBD achieves the highest area under the receiver operating characteristic curve (AUROC, Fawcett, 2006), indicating better separation of stays with and without hypotension. It also detects more affected stays when the false-positive rate is fixed at 10%. These findings show that its distributional predictions help recover the occurrence of clinically important events, complementing the following evaluations of their severity and frequency.

Severity. This task evaluates whether recovered trajectories preserve the depth of low-blood-pressure events. Extreme values matter because average measurements can conceal clinically important deterioration: deeper or prolonged hypotension is associated with increased risk of acute kidney injury (Lehman et al., 2010). We therefore measure the nadir, the lowest MAP reached during a patient's stay, to distinguish a mild threshold crossing from a deeper pressure drop. We estimate it by

Table 1: Paired scores from the MIMIC-III evaluation using fixed-length windows of bedside-monitor measurements. Projected SFBD and CSDI use 16 generated trajectory samples; SAITS, linear, and forward-fill are deterministic. All methods use identical reference windows. Values are means; $\pm$ denotes the standard deviation across random training seeds, available only for SFBD. Bold marks the best score in each row.

Window	Score	Projected SFBD	CSDI	SAITS	Linear	Forward-fill
1 h	ES ↓	$\mathbf{0.9028} \pm 0.0132$	1.325	1.389	1.261	1.537
	VS ↓	$\mathbf{0.0957} \pm 0.0054$	0.2297	0.2308	0.1252	0.2356
3 h	ES ↓	$\mathbf{1.6794} \pm 0.0215$	2.441	2.615	2.376	2.903
	VS ↓	$\mathbf{0.1001} \pm 0.0038$	0.1847	0.1885	0.1291	0.2162
6 h	ES ↓	$\mathbf{2.4303} \pm 0.0316$	3.500	3.821	3.477	4.269
	VS ↓	$\mathbf{0.1002} \pm 0.0036$	0.1628	0.1671	0.1267	0.1963

Table 2: MIMIC-III downstream tasks. Linear and forward-fill are computed from the observed chart values supplied to each SFBD run. Nadir errors are in mmHg; range error concerns vital signs. Nadir and episode biases are signed: $\rightarrow 0$ means closer to zero is better; $\uparrow$ and $\downarrow$ indicate higher and lower is better. Bold marks the best mean in each row.

Task / metric	Projected SFBD	CSDI	SAITS	Linear	Forward-fill
Detection: AUROC ↑	$\mathbf{0.9353} \pm 0.0025$	0.847 ± 0.015	0.848 ± 0.010	0.867 ± 0.009	0.858 ± 0.013
Sensitivity at 90% spec. ↑	$\mathbf{81.9}\% \pm 1.0$	$77.8\% \pm 1.7$	$76.9\% \pm 1.9$	$76.9\% \pm 1.9$	$78.7\% \pm 2.7$
Severity: nadir bias $\rightarrow 0$	$\mathbf{-0.61} \pm 1.55$	$+5.25 \pm 0.12$	$+5.83 \pm 0.10$	$+5.12 \pm 0.15$	$+4.66 \pm 0.17$
Nadir MAE ↓	$\mathbf{5.63} \pm 0.42$	6.77 ± 0.26	6.98 ± 0.15	6.07 ± 0.27	5.90 ± 0.23
Episodes: count bias $\rightarrow 0$	$\mathbf{-4.47}\% \pm 10.71$	$-76.5\% \pm 0.1$	$-77.7\% \pm 0.2$	$-73.8\% \pm 0.6$	$-73.0\% \pm 0.2$
Per-stay count MAE ↓	$\mathbf{1.58} \pm 0.05$	2.84 ± 0.01	2.85 ± 0.01	2.73 ± 0.02	2.73 ± 0.01
Amplitude: range error ↓	$\mathbf{0.407} \pm 0.004$	0.524 ± 0.032	0.554 ± 0.004	0.593 ± 0.004	0.614 ± 0.003

taking the minimum MAP in each generated trajectory and averaging these minima, then compare it with the minimum in the monitor reference. Signed bias shows whether the estimated nadir is systematically too high or too low, while mean absolute error (MAE) measures error magnitude across patients. Projected SFBD achieves the smallest bias magnitude and the lowest MAE. The baselines generally overestimate the nadir, making low-pressure events appear less severe than they were. This complements detection by assessing how low pressure fell, although the nadir alone does not describe how long the low pressure persisted.

Episode count. Beyond whether hypotension occurred and how low pressure fell, we examine whether the recovered trajectories preserve repeated episodes during a patient's stay. This distinguishes recurrent pressure drops from one prolonged event, even when their total duration is similar, and reveals patterns of instability that detection and nadir alone cannot describe. We count each uninterrupted period satisfying the hypotension criterion as one episode, with a recovery to MAP at or above 65 mmHg separating successive episodes. Projected SFBD achieves substantially smaller count bias and per-stay MAE than the baselines, which miss roughly three quarters of episodes. Accurate counting requires preserving the recoveries between pressure drops: if reconstruction smooths over these recoveries, separate episodes appear as one prolonged event. The stronger episode-count recovery therefore complements the VS results, showing that the recovered temporal variation supports more accurate summaries of recurrent hypotension.

Amplitude. We also examine whether recovered trajectories preserve the extent of variation in a patient's vital signs during a stay. Trajectories with similar average values can have very different ranges, so accurate averages alone may conceal substantial fluctuations. We measure this variation using the within-stay range, the difference between the highest and lowest values, and compare its estimate from generated trajectories with the monitor reference. Projected SFBD achieves the lowest range error, suggesting that it better preserves the magnitude of patient-level variation. This complements nadir recovery by considering both ends of the trajectory, while episode count describes how often pressure crosses a clinical threshold.

Table 3: MIMIC-III: projection versus constructed clean anchors.

Metric	Projected SFBD	Clean anchor
CRPS ↓	$\mathbf{0.2158} \pm 0.0039$	0.2260 ± 0.0021
Coverage → 0.90	$\mathbf{0.8323} \pm 0.0318$	0.6084 ± 0.0084
ES, 1 h ↓	$\mathbf{0.9028} \pm 0.0132$	0.9403 ± 0.0055
ES, 3 h ↓	$\mathbf{1.6794} \pm 0.0215$	1.7469 ± 0.0109
ES, 6 h ↓	$\mathbf{2.4303} \pm 0.0316$	2.5220 ± 0.0146
AUROC ↑	$\mathbf{0.9353} \pm 0.0025$	0.9320 ± 0.0053
Nadir bias → 0	$-\mathbf{0.61} \pm 1.55$	$+2.25 \pm 0.90$
Episode bias → 0	$-\mathbf{4.47}\% \pm 10.71$	$-40.13\% \pm 9.84$
Range bias → 0	$+\mathbf{0.09} \pm 0.18$	-0.58 ± 0.09

Table 4: Pulse simulator: signed bias in clinical summaries (→ 0).

Method	PP < 25 dose	MAP < 65 time
Neither	$+2.931 \pm 0.271$	$+0.012 \pm 0.014$
Projection	$+\mathbf{2.036} \pm 0.228$	$-\mathbf{0.005} \pm 0.026$
Constructed anchor	$+3.272 \pm 1.338$	$+0.086 \pm 0.050$
Oracle anchor	$+2.674 \pm 1.813$	$+0.009 \pm 0.008$
Method	**SBP < 90 dose**	**MAP < 65 dose**
Neither	$+0.707 \pm 0.969$	$+1.028 \pm 0.669$
Projection	$+\mathbf{0.106} \pm 0.800$	$+\mathbf{0.129} \pm 0.651$
Constructed anchor	$+2.811 \pm 2.190$	$+2.138 \pm 1.666$
Oracle anchor	$+1.244 \pm 0.444$	$+1.224 \pm 0.107$

5.2 Ablation Study

In this section, we examine how the information used to guide recovery affects the resulting trajectory distribution, comparing physiological constraints with clean-anchor regularization. On MIMIC-III, where complete reference trajectories are unavailable, we construct anchors by linearly interpolating each stay's own charts and applying deterministic clinical rules for physiological relationships, normal values, and valid ranges (Section D.3). The comparison replaces projection with anchor training while retaining the same temporal and cross-channel noise structure. As shown in Table 3, projection gives interval coverage closer to the nominal level, lower energy scores at every evaluated window length, and substantially less undercounting of hypotension episodes. Signed range bias measures whether generated trajectories overestimate or underestimate the difference between a stay's highest and lowest values. Projected SFBD has a mean bias close to zero, whereas clean anchors systematically underestimate the range. Together with their poorer coverage and episode recovery, this suggests that constructed anchors suppress clinically relevant variation rather than faithfully recovering the trajectory distribution.

The Pulse simulator ablation provides a complementary test against complete ground-truth trajectories (Table 4). Relative to using neither anchors nor projection, projection reduces the magnitude of mean bias in all four low-blood-pressure summaries, while constructed anchors increase it in all four. Oracle anchors containing true clean values give mixed results: they improve pulse-pressure burden and time below the MAP threshold, but worsen systolic and mean-arterial-pressure burden. Clean anchor values therefore do not by themselves ensure accurate estimates of the severity and duration of low-pressure events. These results suggest that a limited set of clean samples may not characterize the trajectory distribution sufficiently to improve recovery; moreover, the complete clean trajectories used as oracle anchors are generally unavailable in the clinical setting considered here. Across these comparisons, physiological constraints provide a useful alternative to anchor regularization for recovering clinically relevant properties of the trajectory distribution.

6 Conclusion

We introduced Projected SFBD, a general approach that incorporates domain knowledge into distribution recovery by projecting proposed completions onto a feasible set instead of relying on clean reference trajectories. Our convergence theory extends beyond medical time series to settings where the transformation from complete samples to observations is specified and the domain constraints preserve the target distribution. On MIMIC-III, Projected SFBD improves trajectory-distribution scores and clinically relevant estimates over strong baselines; MIMIC-III and Pulse simulator ablations further support projection as a more reliable way to preserve variation than the evaluated clean-anchor schemes. These results show how domain constraints can guide distribution recovery when complete trajectories are unavailable.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery and review; improve the clarity and presentation of the manuscript; standardize notation, simplify exposition, and reorganize the presentation of proofs; and prepare figures from the authors' experimental results. Coding agents were also used to refine research code, diagnose and correct implementation errors, and assist with computational-resource management. The research questions, central ideas, mathematical arguments, methodological contributions, and choice and formulation of the medical application were developed by the authors. The authors directly checked the cited literature and reviewed and validated all AI-assisted text, notation, proof presentation, figures, code changes, and experimental outputs. The authors take full responsibility for the manuscript, software, results, and conclusions.

ETHICS STATEMENT

This study uses retrospective, deidentified ICU records from MIMIC-III and its matched waveform subset, accessed under PhysioNet's credentialing and data-use requirements, together with synthetic data from the Pulse simulator. No new patient data were collected, and no clinical intervention was performed. The recovered trajectories represent model-based uncertainty about unobserved values; they are not measurements and should not be used for diagnosis, alarm generation, or treatment without prospective external validation. Although the source records are de-identified, any release of trained models or patient-level generated outputs should follow the applicable data-use terms and undergo privacy review. We report only aggregate results and take responsibility for the analyses and claims.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section B states the formal conditions and provides the complete proof of Theorem 1. Section C documents cohort construction, timestamp alignment, preprocessing, architecture, training and sampling settings, diffusion covariance, query generation, re-corruption, constraint projection, adaptive replacement control, evaluation metrics, and baseline configurations. Section D records the ablation protocols and aggregation across random seeds, and Section E gives the finite-dimensional projection used in implementation. The tables state the number of random seeds and, where applicable, compare methods at identical evaluation coordinates. The experiments use MIMIC-III and the MIMIC-III Waveform Database Matched Subset, which require independent credentialed access through PhysioNet, and the Pulse simulator. Patient-level MIMIC-III data cannot be redistributed. The source package includes the manuscript and plotting scripts. Upon acceptance, we will release an open-source implementation containing the training and evaluation code, resolved configurations, and commands needed to reproduce the reported results.

REFERENCES

Asad Aali, Marius Arvinte, Sidharth Kumar, and Jonathan I. Tamir. Solving inverse problems with score-based generative priors learned from noisy data. In *57th Asilomar Conference on Signals, Systems, and Computers*, pp. 837–843, 2023.

Asad Aali, Giannis Daras, Brett Levac, Sidharth Kumar, Alex Dimakis, and Jonathan Tamir. Ambient diffusion posterior sampling: Solving inverse problems with diffusion models trained on corrupted data. In *International Conference on Learning Representations*, pp. 18073–18100, 2025.

Denis Agniel, Isaac S. Kohane, and Griffin M. Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*, 361:k1479, 2018.

Weimin Bai, Yifei Wang, Wenzheng Chen, and He Sun. An expectation-maximization algorithm for training clean diffusion models from corrupted observations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 19447–19471, 2024.

Ashish Bora, Eric Price, and Alexandros G. Dimakis. AmbientGAN: Generative models from lossy measurements. In *International Conference on Learning Representations*, 2018.

Aaron Bray, Jeffrey B. Webb, Andinet Enquobahrie, Jared Vicory, Jerry Heneghan, Robert Hubal, Stephanie TerMaath, Philip Asare, and Rachel B. Clipp. Pulse Physiology Engine: an open-source software platform for computational modeling of human medical simulation. *SN Comprehensive Clinical Medicine*, 1(5):362–377, 2019.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. BRITS: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alexandros G. Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 288–313, 2023.

Giannis Daras, Alex Dimakis, and Constantinos Costis Daskalakis. Consistent diffusion meets tweedie: Training exact ambient diffusion models with noisy data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10091–10108. PMLR, 2024.

Giannis Daras, Yeshwanth Cherapanamjeri, and Constantinos C. Daskalakis. How much is a noisy image worth? data scaling laws for ambient diffusion. In *International Conference on Learning Representations*, pp. 28165–28193, 2025a.

Giannis Daras, Adrian Rodriguez-Munoz, Adam Klivans, Antonio Torralba, and Constantinos Daskalakis. Ambient diffusion omni: Training good models with bad data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 38064–38101, 2025b.

Wenjie Du, David Côté, and Yan Liu. SAITS: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.

Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M. Coopersmith, Craig French, Flávia R. Machado, Lauralyn McIntyre, Marlies Ostermann, Hallie C. Prescott, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive Care Medicine*, 47(11):1181–1247, 2021.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Christopher A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. GP-VAE: Deep probabilistic time series imputation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1651–1661. PMLR, 2020.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

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.

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.

Bahjat Kawar, Noam Elata, Tomer Michaeli, and Michael Elad. GSURE-based diffusion model training with corrupted data. *Transactions on Machine Learning Research*, 2024.

Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion generative models in infinite dimensions. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9538–9563. PMLR, 2023.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951.

Li-wei Lehman, Mohammed Saeed, George Moody, and Roger Mark. Hypotension as a risk factor for acute kidney injury in ICU patients. In *Computing in Cardiology*, volume 37, pp. 1095–1098, 2010.

Jae Hyun Lim, Nikola B. Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, Christopher Pal, Arash Vahdat, and Anima Anandkumar. Score-based diffusion models in function space. *Journal of Machine Learning Research*, 26(158):1–62, 2025.

Haoye Lu, Qifan Wu, and Yaoliang Yu. Stochastic forward–backward deconvolution: Training diffusion models with finite noisy datasets. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40741–40768. PMLR, 2025.

Haoye Lu, Darren Lo, and Yaoliang Yu. SFBD Flow: A continuous-optimization framework for training diffusion models with noisy samples. arXiv preprint arXiv:2506.02371v2, 2026a. Version 2, revised April 4, 2026.

Haoye Lu, Yaoliang Yu, and Darren Lo. SFBD-OMNI: Bridge models for lossy measurement restoration with limited clean samples. In *International Conference on Learning Representations*, 2026b.

Benjamin Moody, George Moody, Mauricio Villarroel, Gari D. Clifford, and Ikaro Silva. MIMIC-III Waveform Database Matched Subset. PhysioNet, 2020. Version 1.0.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. Chapter 4: Covariance Functions.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24804–24816, 2021.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

Griffin M. Weber, William G. Adams, Elmer V. Bernstam, Jonathan P. Bickel, Kathe P. Fox, Keith Marsolo, Vijay A. Raghavan, Alexander Turchin, Xiaobo Zhou, Shawn N. Murphy, and Kenneth D. Mandl. Biases introduced by filtering electronic health records for patients with "complete data". *Journal of the American Medical Informatics Association*, 24(6):1134–1141, 2017.

Marije Wijnberge, Björn van der Ster, Alexander P. J. Vlaar, Markus W. Hollmann, Bart F. Geerts, and Denise P. Veelo. The effect of intermittent versus continuous non-invasive blood pressure monitoring on the detection of intraoperative hypotension, a sub-study. *Journal of Clinical Medicine*, 11(14):4083, 2022.

Maxwell A. Xu, Alexander Moreno, Supriya Nagesh, V. Burak Aydemir, David W. Wetter, Santosh Kumar, and James M. Rehg. PulseImpute: A novel benchmark task for pulsative physiological signal imputation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26874–26888, 2022.

A Learning Distributions from Corrupted Observations

Our setting is related to ambient generative modeling, which learns a clean distribution without directly observing clean samples. AmbientGAN establishes that adversarial training can recover a data distribution from lossy measurements when the measurement process is sufficiently informative (Bora et al., 2018). Ambient Diffusion replaces adversarial training with a diffusion objective: it applies additional measurement corruption and learns to predict the original corrupted sample, recovering clean-distribution information under conditions that include random inpainting and compressed sensing (Daras et al., 2023). For additive Gaussian corruption, SURE-Score estimates a clean score from noisy training samples using Stein's unbiased risk estimate (Aali et al., 2023); GSURE-Diffusion extends unbiased-risk training to randomized linear measurements (Kawar et al., 2024); and Consistent Diffusion Meets Tweedie enforces cross-noise-level consistency to obtain asymptotically exact sampling below the observed noise level (Daras et al., 2024). These methods demonstrate that clean-distribution learning can be possible without paired clean targets, but their guarantees depend on the structure and identifiability of the measurement process.

Recent work also clarifies the finite-data cost of relying only on corrupted samples. Daras et al. (2025a) show that noisy samples can be substantially less informative than clean ones at practical dataset sizes and that a small clean subset can sharply reduce the required number of corrupted samples. Ambient Diffusion Omni instead uses degraded or out-of-distribution samples mainly at diffusion times where added Gaussian noise reduces their mismatch with high-quality data, trading the bias introduced by imperfect samples against the variance reduction from having more data (Daras et al., 2025b). This corruption-agnostic strategy does not require an explicit observation model, but its additional noising can discard information that remains useful in the measurements.

EMDiffusion is closer algorithmically to SFBD-OMNI: given a known imaging operator, it alternates between reconstructing clean images with the current diffusion prior and retraining that prior on the reconstructions (Bai et al., 2024). SFBD-OMNI derives a related alternation from a KL ambient-projection objective, treats the corruption mechanism as a black-box sampling kernel, and explicitly distinguishes identifiable corruptions from cases requiring additional information (Lu et al., 2026b). Projected SFBD supplies this additional information through domain constraints rather than constructed clean anchors. The constraints restrict the admissible trajectory laws while preserving a distribution over feasible completions.

Diffusion-based inverse-problem solvers address a different task. Methods for medical-image reconstruction and diffusion posterior sampling combine a pretrained clean-data diffusion prior with a measurement model to reconstruct or sample the latent object for an individual observation (Song et al., 2022; Chung et al., 2023). Ambient Diffusion Posterior Sampling connects the two regimes by learning a prior from linearly corrupted measurements and then using that prior for inverse reconstruction (Aali et al., 2025). These methods ultimately target reconstruction conditioned on a particular measurement, whereas Projected SFBD learns the population distribution itself from a dataset of incomplete observations and uses the observation process during training.

B Proof and Formal Conditions for Theorem 1

This appendix presents the theoretical argument in dependency order: projection assumptions, feasibility, KL nonexpansiveness, the joint-law formulation, and the proof of Theorem 1.

B.1 Formal Setting and Projection Assumptions

Assumption 1 (Target Feasibility)**.** *The target trajectory law satisfies $p_{\mathrm{data}}(\mathcal{G}) = 1$. Equivalently, under the exact sampling model, for q-almost every record y,*

$$p_{\mathrm{data}}(\mathcal{C}(y) \mid y) = 1. \tag{18}$$

Thus true posterior trajectories are already feasible and are left unchanged by Π_y.

Assumption 2 (Target-Preserving Projection Kernel)**.** *The implemented record association and constraint projection induce a common measurable Markov kernel on $\mathcal{H}$, denoted by R, for candidate and target laws. This kernel preserves the target:*

$$R_{\#} p_{\mathrm{data}} = p_{\mathrm{data}}. \tag{19}$$

This is an additional condition on the projection operator at the distribution level; pointwise feasibility of Π_y alone does not establish it. Under this condition, the data-processing inequality gives the KL nonexpansiveness bound proved in Lemma 2 below.

B.2 Feasibility of Projected Trajectories

Lemma 1. *For every record $y = ((t_i, v_i))_{i=1}^n$ with nonempty $\mathcal{C}(y)$, a projected trajectory $f^+ = \Pi_y(\hat{f})$ satisfies*

$$\Pr[f^+ \in \mathcal{C}(y) \mid y] = 1, \qquad f^+(t_i) = v_i \quad \textit{almost surely for every } i.$$

Consequently the law of projected trajectories is supported on $\mathcal{G}$.

Proof. The projection takes its values in $\mathcal{C}(y) \subseteq \mathcal{G}$, and the definition of $\mathcal{C}(y)$ fixes all recorded evaluations. □

B.3 KL Nonexpansiveness of Projection

The target-preserving property gives a simple but useful guarantee.

Lemma 2. *Suppose Assumption 2 holds. Then for any candidate clean distribution p,*

$$\boxed{D_{\mathrm{KL}}\left(p_{\mathrm{data}} \parallel R_{\#}p\right) \le D_{\mathrm{KL}}\left(p_{\mathrm{data}} \parallel p\right).} \tag{20}$$

Proof. Since the projection operator preserves the target distribution,

$$R_{\#}p_{\mathrm{data}} = p_{\mathrm{data}}.$$

Therefore,

$$D_{\mathrm{KL}}\left(p_{\mathrm{data}} \parallel R_{\#}p\right) = D_{\mathrm{KL}}\left(R_{\#}p_{\mathrm{data}} \parallel R_{\#}p\right) \le D_{\mathrm{KL}}\left(p_{\mathrm{data}} \parallel p\right), \tag{21}$$

where the inequality follows from the data-processing inequality for KL divergence (Kullback & Leibler, 1951, Theorem 4.1). □

B.4 Joint-Law Formulation

Eq. (10) has the equivalent support-constrained joint-law formulation

$$\min_{p \in \mathfrak{P}_{\mathcal{C}}} \inf_{\pi:\, \pi_Y = q} D_{\mathrm{KL}}\left(\pi \parallel p(df)\, r(dy \mid f)\right), \tag{22}$$

where π couples trajectories and records and π_Y is its record marginal. For fixed p with finite objective, the inner minimum equals $D_{\mathrm{KL}}\left(q \parallel T_r p\right)$ and is attained by coupling q with the posterior under p.

B.5 Proof of Theorem 1

The observation kernel $r(dy \mid f)$ is known and common to the target and candidate laws, with $q = T_r p_{\mathrm{data}}$. We assume Assumptions 1 and 2, $H_0 := D_{\mathrm{KL}}\left(p_{\mathrm{data}} \parallel p^0\right) < \infty$, and feasibility of every projected iterate, $p^k \in \mathfrak{P}_{\mathcal{C}}$ for $k \ge 1$. The first condition below supports convergence and the best-iterate rate, the second is additionally used for monotonicity under exact posterior estimation, and the third supports convergence in the space of clean laws.

1. The true, exact candidate, and learned posteriors are well-defined. Writing $\mathcal{E}(u) = \int D_{\mathrm{KL}}\left(u_y^{\mathrm{data}} \parallel u_y\right) q(dy)$, the relevant errors are finite and the bounds δ_k in Eq. (7) satisfy $\sum_{k=0}^{\infty} \delta_k < \infty$. The adaptive rule in Eq. (6) uses these bounds, with $\beta, \gamma_{\max} \in (0, 1)$ and $\gamma_{\mathrm{base}}, v_0 > 0$. Exact posteriors permit $\delta_k = 0$; the initialization and $\beta > 0$ ensure $v_{k+1} > 0$ at every finite iteration.

2. For monotonicity and the last-iterate rate, $\hat{u}_y^k = u_y^k$ for q-almost every y, the feasible candidate family is closed under posterior aggregation and relaxed mixing, and $R_{\#}p = p$ for every $p \in \mathfrak{P}_{\mathcal{C}}$. Under the exact sampling model, pointwise projection has this identity property when feasible trajectories are paired with compatible records. This property is stronger than preserving p_{data} alone and is not required for the approximate-update result.
3. For convergence to the solution set and clean-distribution recovery, the iterates are relatively compact for weak convergence, T_r is continuous for this convergence, and $\mathfrak{P}_{\mathcal{C}}$ is closed.

Uniform boundedness and a common Lipschitz continuity bound give relative compactness of trajectories in the uniform topology by Arzelà–Ascoli; compactness in the chosen Hilbert-space topology requires a separate justification.

Proof of Theorem 1. The proof has four steps: a one-step KL bound, convergence under adaptive updates, monotonicity with exact posteriors, and convergence to the admissible solution set.

Step 1: One-step KL bound. Let $H_k = D_{\mathrm{KL}}\left(p_{\mathrm{data}} \parallel p^k\right)$ and $D_k = D_{\mathrm{KL}}\left(q \parallel T_r p^k\right)$. Since the same observation kernel acts on candidate and target laws, the KL chain rule gives $H_k = D_k + \mathcal{E}(u^k)$. Averaging the learned posteriors over q and applying convexity of KL yields

$$D_{\mathrm{KL}}\left(p_{\mathrm{data}} \parallel \hat{\rho}^k\right) \le \mathcal{E}(\hat{u}^k) \le H_k - D_k + \delta_k.$$

Applying the KL-nonexpansiveness result in Lemma 2 and convexity for the relaxed update gives

$$H_{k+1} \le D_{\mathrm{KL}}\left(p_{\mathrm{data}} \parallel \bar{p}^{k+1}\right) \le (1-\gamma_k)H_k + \gamma_k D_{\mathrm{KL}}\left(p_{\mathrm{data}} \parallel \hat{\rho}^k\right) \le H_k - \gamma_k D_k + \gamma_k \delta_k.$$

Step 2: Adaptive convergence and rate. Summing over k telescopes the H_k terms:

$$\sum_{k=0}^{K-1} \gamma_k D_k \le H_0 - H_K + \sum_{k=0}^{K-1} \gamma_k \delta_k \le H_0 + \gamma_{\max} \sum_{k=0}^{\infty} \delta_k. \tag{23}$$

The right-hand side is finite. Summability of δ_k implies $\delta_k \to 0$, and the adaptive recursion satisfies

$$v_{k+1} = \beta^{k+1} v_0 + (1-\beta) \sum_{j=0}^{k} \beta^{k-j} \delta_j^2.$$

Since $\sum_j \delta_j < \infty$, we have $\sum_j \delta_j^2 < \infty$. For any fixed J, the finite sum over $j \le J$ vanishes as $k \to \infty$, while the remaining tail is bounded by $\sum_{j>J} \delta_j^2$, which can be made arbitrarily small. Hence $v_{k+1} \to 0$ and $\gamma_k = \gamma_{\max}$ for all sufficiently large k. It follows from Eq. (23) that $\sum_k D_k < \infty$, hence $D_k \to 0$. Moreover,

$$\min_{0 \le k < K} D_k \le \frac{H_0 + \sum_{k=0}^{K-1} \gamma_k \delta_k}{\sum_{k=0}^{K-1} \gamma_k} = O(K^{-1}),$$

because the numerator is bounded and the denominator grows linearly for sufficiently large K.

Step 3: Monotonicity with exact posteriors. For exact posterior updates, the excess error vanishes, and the descent bound sharpens to $H_{k+1} \le H_k - \gamma_k D_k$, regardless of which valid error bounds are used to choose γ_k. Let $k_0 = 0$ if $p^0 \in \mathfrak{P}_{\mathcal{C}}$ and $k_0 = 1$ otherwise. For $k \ge k_0$, feasibility of p^k and $q \ll T_r p^k$ imply $u_y^k(\mathcal{G}) = 1$ for q-almost every y. Thus ρ^k and $(1-\gamma_k)p^k + \gamma_k \rho^k$ are feasible, and the additional identity property of R gives $p^{k+1} = (1-\gamma_k)p^k + \gamma_k \rho^k$. In particular, projection leaves this exact update unchanged.

To prove monotonicity of the observed objective, use the joint law $\pi^k(df, dy) = q(dy) u_y^k(df)$, whose clean marginal is ρ^k. The KL chain rule in the two conditioning directions gives

$$D_{\mathrm{KL}}\left(\pi^k \parallel p^k(df) r(dy \mid f)\right) = D_k, \qquad D_{\mathrm{KL}}\left(\pi^k \parallel \rho^k(df) r(dy \mid f)\right) = D_k - D_{\mathrm{KL}}\left(\rho^k \parallel p^k\right).$$

Marginalizing to the observed record and applying the data-processing inequality therefore yields

$$D_{\mathrm{KL}}\left(q \parallel T_r \rho^k\right) \le D_k - D_{\mathrm{KL}}\left(\rho^k \parallel p^k\right).$$

Since T_r is linear and KL is convex in its second argument, every $\gamma_k \in (0, 1]$ satisfies

$$D_{k+1} \le (1-\gamma_k)D_k + \gamma_k D_{\mathrm{KL}}\left(q \,\|\, T_r\rho^k\right) \le D_k - \gamma_k D_{\mathrm{KL}}\left(\rho^k \,\|\, p^k\right) \le D_k, \qquad k \ge k_0. \quad (24)$$

This is the exact alternating-minimization descent underlying SFBD-OMNI (Lu et al., 2026b, Appendix F), applied here with adaptive relaxation and an inactive projection. Combining it with $D_k \to 0$ gives $D_k \downarrow 0$ from k_0 onward. Moreover, monotonicity and the error-free telescoping bound give, for $K > k_0$,

$$D_K \sum_{k=k_0}^{K-1} \gamma_k \le \sum_{k=k_0}^{K-1} \gamma_k D_k \le H_{k_0} - H_K \le H_{k_0}, \qquad D_K \le \frac{H_{k_0}}{\sum_{k=k_0}^{K-1} \gamma_k} = O(K^{-1}).$$

The adaptive rule again makes the denominator grow linearly. For approximate posteriors, neither this monotonicity nor the last-iterate rate follows from the summable excess-error condition alone.

Step 4: Convergence to the admissible solution set. For convergence to the solution set, write $\mathcal{S} = \mathcal{E}_r(q) \cap \mathfrak{P}_{\mathcal{C}}$, which is nonempty because $p_{\mathrm{data}} \in \mathcal{S}$. Consider any weakly convergent subsequence with limit $\bar{p}$. Continuity of T_r and lower semicontinuity of KL imply $D_{\mathrm{KL}}\left(q \,\|\, T_r\bar{p}\right) = 0$, so $T_r\bar{p} = q$. Closedness of $\mathfrak{P}_{\mathcal{C}}$ gives $\bar{p} \in \mathfrak{P}_{\mathcal{C}}$. Thus every weak limit point belongs to $\mathcal{S}$.

More precisely, let d_{w} be a metric inducing weak convergence of probability laws on the separable Hilbert space $\mathcal{H}$. Convergence to the solution set means

$$\mathrm{dist}_{\mathrm{w}}(p^k, \mathcal{S}) := \inf_{p \in \mathcal{S}} d_{\mathrm{w}}(p^k, p) \longrightarrow 0.$$

If this failed, some subsequence would stay at distance at least $\varepsilon > 0$ from $\mathcal{S}$. Relative compactness would give a further subsequence converging weakly to $\bar{p} \in \mathcal{S}$, contradicting $\mathrm{dist}_{\mathrm{w}}(p^k, \mathcal{S}) \le d_{\mathrm{w}}(p^k, \bar{p}) \to 0$ along that subsequence. This proves set convergence for both exact and approximate posterior updates. If prior-restricted identifiability holds, $\mathcal{S} = \{p_{\mathrm{data}}\}$, and the same conclusion becomes $p^k \to p_{\mathrm{data}}$ weakly. □

C Experimental Details

C.1 Cohorts and Evaluation Scope

The continuous MIMIC-III cohort contains 6,355 ICU stays, with one stay per subject and lengths between 12 and 120 hours. Its 16 channels are HR, SBP, DBP, MAP, SpO_2, respiratory rate, lactate, creatinine, hemoglobin, white blood cell count, glucose, potassium, sodium, BUN, bicarbonate, and platelets. The 5,686-stay chart file and 669-stay waveform-matched file have disjoint subjects, but both groups contribute chart observations to training, making the monitor evaluation transductive. Monitor measurements are excluded from training targets and conditioning inputs. The same high-frequency bedside-monitor reference set informed configuration selection, so it is not an untouched test set. The 1-minute monitor medians are aggregated within five-minute bins; empty bins contribute no reference value. A channel needs at least 12 finite reference minutes to be considered. Before restricting to the 12-hour evaluation span, the eligible counts are 669 for HR, 665 for RR, 660 for SpO_2, and 303 each for SBP, DBP, and MAP. The final span restriction can reduce these counts slightly. No dense laboratory reference is available.

C.2 Continuous-Time Training Configuration

Architecture and time encoding. Each observed or queried value is represented by its timestamp, channel, and value. A conditional transformer approximates the trajectory posterior, using rotary positional encoding at the actual timestamps. A persistent set stores the current completed trajectories; each replacement step proposes new completions, applies the feasibility projection, and trains after re-corruption. Recorded values remain fixed during sampling and projection. Table 5 gives the architecture and training settings.

Completion queries and re-corruption. Time–channel locations specify where to evaluate a completion and do not impose a common acquisition grid. Each channel's whole-stay query intensity is set from its observed events per hour. Local queries lie in the middle half of gaps between

consecutive measurements. The mixture allocates half the intensity to each generator. Re-corruption independently retains completed time–channel–value entries at the estimated channel-specific rate: the number of retained observations divided by the number of observations plus queries. Evaluation-only monitor coordinates are excluded from this estimate. The resulting operator is value-independent.

Table 5: Configuration for the 12,000-step MIMIC-III ablations.

Component	Setting
Transformer	3 layers, width 96, 4 heads; 531,361 parameters
Time encoding	RoPE at actual timestamps in hours, scale 0.5
Diffusion target	v-prediction; min-SNR weighting, cap 5
VP schedule	$\beta(s) = 0.1 + 19.9s$, $s \in [0, 1]$
Sampler	30 DDIM steps, colored noise, $\eta = 1$
Guidance	1; no classifier-free guidance in these runs
Optimizer	AdamW; learning rate 10^{-3}, weight decay 10^{-4}
Batch / gradient clipping	32 stays / global norm 1
Training	12,000 steps, fp32; seeds 0, 1, 2
Evaluation	16 independent generated trajectory samples per stay
Queue	One record per stay, forward-fill initialization
Replacement proposals	5% of stays every 5 gradient steps
Clean anchors	None; clean and dense-anchor fractions zero
Query mixture	Equal local and whole-stay components
Model-input / query limits	1,500 chart time–channel–value entries; 600 prediction locations before adding simultaneous BP prediction locations

Constraint projection. We apply channel-specific physiological range bounds to generated values and jointly enforce the blood-pressure relationships described below. After constructing and capping the initial prediction locations, we add any missing systolic blood pressure (SBP), diastolic blood pressure (DBP), and mean arterial pressure (MAP) prediction locations at each relevant blood-pressure timestamp. This supplies simultaneous SBP, DBP, and MAP predictions at the same timestamps for the coupled BP constraints. These locations request model predictions; actual measurement times and recorded values remain fixed. In the recorded real-data ablations, disabling joint constraint projection also disables the addition of simultaneous BP prediction locations, while channel-specific range bounds remain active, so the comparison evaluates the implemented projection procedure as a whole rather than isolating the projection map at fixed prediction locations.

The real-data implementation enforces DBP $\leq$ MAP $\leq$ SBP and a minimum pulse-pressure constraint of 10 mmHg, together with the source-specific MAP constraint derived in Section E.3. Its arterial offset and tolerance are $(0, 10)$ mmHg, and its cuff values are $(-4.64, 13)$ mmHg. These are approximate prior restrictions, not universally satisfied laws. For example, 0.68% of audited real reference triples violate the minimum pulse-pressure constraint.

The real MAP constants are hard-coded and their recorded provenance uses reference residual quantiles. A later co-observed-only calibration gives an arterial center of -0.112 and half-width 9.765 mmHg at the 99th percentile, close to the deployed values. That agreement does not turn the historical experiment into an independently calibrated test. The dense arterial-reference residual has mean -0.146 and SD 3.670 mmHg over 37,235 triples. This audit does not validate the cuff band or measure the MAP band's incremental coverage cost.

The continuous implementation uses 15 passes through the projection procedure and can leave residual MAP-band violations. Observed members remain fixed. In an audit of the historical MIMIC-IV grid, 2,064 co-observed pairs violate DBP $\leq$ MAP and 608 violate MAP $\leq$ SBP. One of 1,895,040 cells has observed DBP above observed SBP and a generated MAP, making the conditional ordering set empty. Measurements aggregated from different times or devices can produce inconsistent grid triples; the implementation skips fully fixed constraints and reports remaining constraint violations. These observations motivate the explicit feasibility and target-preservation assumptions in the theory.

Diffusion noise. For event times t_i and channel labels c_i, the noise covariance is

$$K_{ij} = R_{c_i c_j} \sum_{\ell=1}^{3} w_\ell \exp\left[-\frac{(t_i - t_j)^2}{2a_\ell^2}\right],$$

The RBF mixture combines smooth correlations at distinct time scales (Rasmussen & Williams, 2006), with weights $(w_1, w_2, w_3) = (0.3, 0.3, 0.4)$ and scales $(a_1, a_2, a_3) = (0.083, 2, 12)$ hours, corresponding approximately to five minutes, two hours, and twelve hours. Diagonal jitter is 10^{-3}. Independent channel noise uses $R = I$; the correlated configuration estimates the vitals block from co-located chart measurements. The scales are empirical choices, not identified physiological constants. All observations are reset to the recorded values after each sampling step. Section D.2 reports the scale and weight comparisons.

MMD estimator and replacement control. For the practical update in Eq. (17), we linearly interpolate the generated part of each channel onto 16 relative-time points, requiring at least two query points and a positive time span. We estimate squared MMD between proposed trajectory embeddings and embeddings from the stored completed trajectories using the unbiased U-statistic, which excludes within-sample self-pairs (Gretton et al., 2012). The RBF kernel uses a median-distance bandwidth, with up to 128 embeddings per side; estimates are computed per channel and averaged across channels. To estimate the discrepancy caused by sampling alone, we apply the same comparison to two disjoint subsamples of the stored completed trajectories, matching the sample sizes used for proposed and stored completions. We square the MMD^2 estimates from both comparisons and maintain exponential moving averages with decay 0.9: each average retains 90% of its previous value and adds 10% of the current squared estimate. The average comparing proposals with the stored set is v_k; the average comparing two stored-set subsamples is n_k, the reference level for sampling fluctuations in Eq. (17). A spike reaching twice the previous smoothed squared diagnostic skips the update; otherwise the acceptance probability follows the cap in Eq. (17). This empirical proxy does not certify Eq. (7); neither its error summability nor the projection operator's target preservation is verified by the experiments.

C.3 Evaluation Metrics

Table 6 summarizes the evaluation metrics. A generated trajectory sample represents one possible completed trajectory, and a window is a short segment of that trajectory.

Table 6: A guide to the evaluation metrics. ↓ indicates lower is better; coverage should approach 0.90. CRPS and ES follow Gneiting & Raftery (2007); VS follows Scheuerer & Hamill (2015).

Metric	Better	What it measures
Individual measurements		
CRPS	↓	Continuous ranked probability score. Evaluates the predicted distribution for each measurement, accounting for both accuracy and uncertainty.
90% coverage	→ 0.90	Fraction of reference values inside the central 90% predictive interval. Low coverage means the intervals miss reference values too often.
Trajectory windows		
ES	↓	Energy score: the multivariate counterpart of CRPS. Evaluates a predicted window as a vector against the same patient's reference window.
VS	↓	Variogram score. Checks whether differences between time points match the reference, revealing errors in temporal dependence.
Clinical summaries		
Range error	↓	Error in the estimated difference between the highest and lowest values during a stay; assesses recovery of the extent of variation.
Duration error	↓	Error in the estimated time spent above or below a specified threshold; assesses how long abnormal episodes last.
Dose error	↓	Error in the accumulated deviation beyond a threshold; accounts for both the duration and severity of abnormal episodes.

Individual measurements

Let $\mathcal{I}_c$ be the reference coordinates scored for channel c, with $n_c = |\mathcal{I}_c|$, reference values y_j, and generated trajectory samples $x_j^{(1)}, \ldots, x_j^{(S)}$. Each comparison uses the same reference coordinates

for all methods. In the continuous benchmark, values are standardized using channel statistics from observed data; metrics are computed within each channel and then averaged equally over the eligible channels in the reported group.

Marginal predictive accuracy. For a predictive cumulative distribution function F and reference y, the continuous ranked probability score is

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} \big(F(z) - \mathbf{1}\{y \le z\}\big)^2 \, dz = \mathbb{E}|X - y| - \tfrac{1}{2}\mathbb{E}|X - X'|,$$

with X, X' independent samples from F (Gneiting & Raftery, 2007). The per-cell benchmark evaluates the empirical predictive distribution:

$$\widehat{\mathrm{CRPS}}_j = \frac{1}{S}\sum_{s=1}^{S} |x_j^{(s)} - y_j| - \frac{1}{2S^2}\sum_{s,s'=1}^{S} |x_j^{(s)} - x_j^{(s')}|, \qquad \mathrm{CRPS}_c = \frac{1}{n_c}\sum_{j \in \mathcal{I}_c} \widehat{\mathrm{CRPS}}_j.$$

The group score is the mean of CRPS_c over eligible channels. The S^2 denominator includes diagonal pairs: this is the exact CRPS of the empirical distribution, rather than an unbiased estimate of the population score. The fair ensemble correction uses $S(S-1)$ and excludes diagonal pairs (Ferro, 2014); some historical burden experiments use that correction. A deterministic prediction is a point mass, whose CRPS equals absolute error.

Interval calibration. Let l_j and u_j be the empirical 5th and 95th percentiles of the generated trajectory samples, using linear interpolation between order statistics. For the coordinates with bedside-monitor references of channel c,

$$\mathrm{Cov}_{90,c} = \frac{1}{n_c}\sum_{j \in \mathcal{I}_c} \mathbf{1}\{l_j \le y_j \le u_j\}.$$

Trajectory windows

For a reference window $y \in \mathbb{R}^L$ and coherent posterior windows $x_1, \ldots, x_S$, we use the fair-ensemble energy score and the variogram score with exponent $1/2$ and equal weights (Gneiting & Raftery, 2007; Scheuerer & Hamill, 2015):

$$\widehat{\mathrm{ES}} = \frac{1}{S}\sum_{s=1}^{S} \|x_s - y\|_2 - \frac{1}{2S(S-1)}\sum_{s \ne s'} \|x_s - x_{s'}\|_2,$$

$$\widehat{\mathrm{VS}} = \binom{L}{2}^{-1} \sum_{i<j} \left(|y_i - y_j|^{1/2} - \frac{1}{S}\sum_{s=1}^{S} |x_{si} - x_{sj}|^{1/2} \right)^2.$$

ES excludes self-pairs of generated trajectory samples, giving the fair ensemble correction for independent posterior samples (Ferro, 2014), unlike the per-cell CRPS convention above. For a deterministic forecast ($S = 1$), the spread term is zero. VS uses the displayed sample average inside the square and is not asserted to be an unbiased Monte Carlo estimate of the population score. Scores are averaged over reference windows within each channel, then over eligible channels. Each predicted window remains paired with its own reference throughout scoring.

Windows are non-overlapping within each stay–channel series and identical across methods within a comparison. Real-data windows contain 12, 36, or 72 successive five-minute reference points; missing monitor bins can make their elapsed spans longer than the nominal one, three, or six hours. The hourly-mean evaluation first averages each 12-point block before constructing six- or twelve-hour windows. ES depends on vector dimension, so comparisons use the same window definition.

Pooled energy distance. As a supplementary diagnostic, pooled ED compares the distribution of all generated windows with that of the reference windows, without individual prediction–reference pairing (Székely & Rizzo, 2013). For generated windows $A = \{a_i\}_{i=1}^m$ and reference windows $B = \{b_j\}_{j=1}^n$, we use

$$d(A, B) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \|a_i - b_j\|_2, \qquad \widehat{\mathrm{ED}} = \sqrt{\big[2d(A,B) - d(A,A) - d(B,B)\big]_+}.$$

Generated windows pool all posterior samples, and self-distance terms include diagonal pairs. ED is computed per channel and averaged over eligible channels. It measures agreement between populations of windows and complements the paired ES and VS; it does not assess whether a generated window matches the particular patient for whom it was generated.

DOWNSTREAM QUANTITIES

Downstream trajectory functionals. For a scalar functional g of a stay's trajectory, the error of its posterior expectation is

$$E_c(g) = \frac{1}{N_c} \sum_{i=1}^{N_c} \left| \frac{1}{S} \sum_{s=1}^{S} g(x_{ic}^{(s)}) - g(y_{ic}) \right|, \qquad g_{\text{range}}(z) = \max_j z_j - \min_j z_j,$$

where N_c counts eligible stays and each $x_{ic}^{(s)}$ and y_{ic} is evaluated at the same coordinates. Range error uses g_{range}, so it compares the average of the generated trajectory samples' ranges with the reference range. It is neither the range of the posterior mean nor the average absolute range error of individual generated trajectory samples.

For a trajectory $z(t)$ on $[0, T]$ and threshold a, below-threshold dose and duration are

$$g^-_{\text{dose},a}(z) = \int_0^T (a - z(t))_+ \, dt, \qquad g^-_{\text{duration},a}(z) = \int_0^T \mathbf{1}\{z(t) < a\} \, dt,$$

where $(v)_+ = \max(v, 0)$. Above-threshold versions replace $(a - z(t))_+$ by $(z(t) - a)_+$ and $z(t) < a$ by $z(t) > a$. Grid evaluation replaces the integrals by sums weighted by the time step; duration reported in steps takes that weight as one. Both are scored by $E_c(g)$; cross-channel targets apply g to the jointly generated channels.

C.4 BASELINE CONFIGURATIONS

Real-data baselines. SAITS has two layers, width 128, four attention heads, and feedforward width 256; CSDI has four layers, eight heads, width 64, and 50 diffusion steps. The baseline architecture hyperparameters are chosen so that SAITS, CSDI, and Projected SFBD have similar numbers of trainable parameters. Both learned baselines train for 50 epochs with batch size 64.

Comparison protocol. The updated MIMIC-III window comparison uses the recorded Projected SFBD configuration trained for the full training schedule and CSDI/SAITS trained on the same cohort. The baseline evaluation protocol averages observed chart values within approximately hourly bins and reads predictions back at each target's bin; SFBD retains continuous timestamps. Linear and forward-fill are reconstructed from each stay's observed chart values within every scoring block. The recorded consistency check covers 382,145 high-frequency bedside-monitor measurements for the same patients, timestamps, and channels and finds zero difference in reference values between SFBD and the saved predictions from the learned baselines. Different total numbers of saved predictions reflect prediction-location construction outside this shared evaluation grid. Thus these scores compare complete methods, including their input representations; they do not isolate architecture or temporal resolution.

D ADDITIONAL EMPIRICAL ABLATIONS

D.1 MIMIC-III STRUCTURAL AND KERNEL COMPARISONS

Table 7 shows that combining correlated noise with projection gives the strongest CRPS and coverage, whereas replacing the multi-scale kernel with a single scale worsens both. Figure 2 complements these results with energy distance overall and separately for blood-pressure and other-vital channels. Projection runs also add simultaneous SBP, DBP, and MAP prediction locations; because no run adds these locations without enforcing the constraints, the effects of the locations and projection cannot be separated.

Table 7: Matched 12,000-step MIMIC-III ablations, mean ± standard deviation across three seeds. Full configurations, including projection query construction, are given in Section C.2; channel-specific range bounds remain active throughout. Bold marks the best values, with coverage ranked by proximity to 0.90.

Configuration	CRPS ↓	Coverage → 0.90
Independent noise; no projection	0.230 ± 0.001	0.717 ± 0.021
Correlated noise; no projection	0.229 ± 0.002	0.727 ± 0.033
Independent noise; projection	0.229 ± 0.001	0.727 ± 0.029
Correlated noise; projection	$\mathbf{0.227} \pm 0.001$	$\mathbf{0.753} \pm 0.017$
Correlated noise; projection; single 2 h kernel	0.238 ± 0.002	0.683 ± 0.053

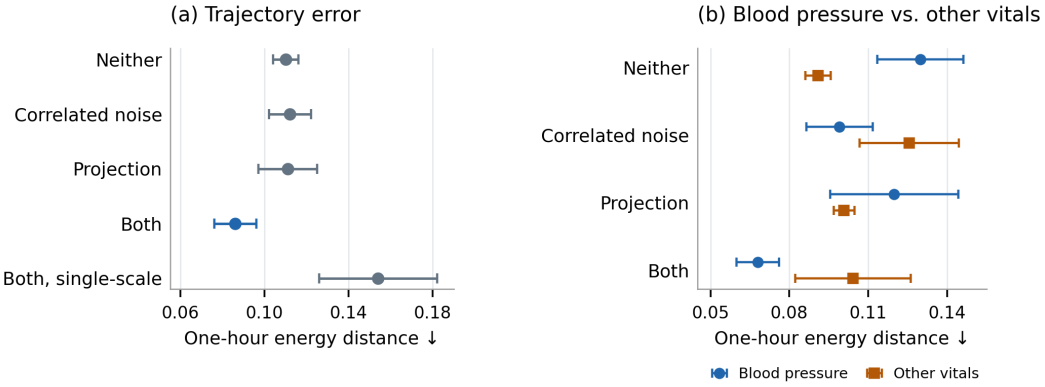


Figure 2: One-hour energy distance for the matched 12,000-step MIMIC-III ablations, reported overall and separately for blood-pressure and other-vital channels; points and bars show the mean and standard deviation across three seeds.

D.2 Noise-Mixture and Weight Sensitivity

The default multi-scale temporal kernel is defined in Section C.2. Holding the prediction locations and all other model settings fixed, we replace only its mixture weights with values fitted from observed-chart autocorrelation; Table 8 reports the matched comparison.

Table 8: Matched 12,000-step MIMIC-III comparison of the default and observed-chart-autocorrelation-fitted mixture weights, mean ± standard deviation across three seeds. Changes are differences of means; bold marks the best score, with coverage ranked by proximity to 0.90.

Metric	Default $(0.30, 0.30, 0.40)$	Fitted $(0.556, 0.348, 0.096)$	Change
CRPS ↓	$\mathbf{0.227} \pm 0.001$	$\mathbf{0.227} \pm 0.002$	0.000
Coverage → 0.90	$\mathbf{0.753} \pm 0.017$	0.743 ± 0.009	-0.010
ED, 1 h ↓	0.086 ± 0.010	$\mathbf{0.085} \pm 0.022$	-0.001
ED, 6 h ↓	0.168 ± 0.021	$\mathbf{0.167} \pm 0.009$	-0.001
Hourly-mean ED, 6 h ↓	$\mathbf{0.078} \pm 0.011$	0.079 ± 0.009	$+0.001$

The mean changes in Table 8 are small, but the one-hour ED changes for the three seeds are -0.012, $+0.016$, and -0.007, so the aggregate difference is not a per-seed equivalence bound. Because the fitted configuration changes all three weights, this experiment supports limited sensitivity only along that reallocation direction; it does not identify any one component or establish robustness over the entire weight simplex.

D.3 Constructed Clean Anchors

The deterministic `physiolinear` procedure uses only each stay's own observed chart values. It linearly interpolates between observations and supplements the interpolated trajectories with clinical domain knowledge: it uses the arterial relationship $\mathrm{MAP} = (\mathrm{SBP} + 2\,\mathrm{DBP})/3$, reverts to clinical normal values before a channel's first observation, and clamps values to physiological bounds. Anchor construction uses no bedside-monitor measurements, external data, observations from other patients, learned imputation model, or random sampling.

For the MIMIC-III comparison in Table 3, the anchor pool comprises 109 stays (2% of 5,486), with anchor trajectories constructed on a 15-minute grid over at most 12 hours. Constructed-anchor training replaces projection, using pure-anchor batch probability $p = 0.1$. Both configurations retain multi-scale temporal noise and fitted cross-channel correlation and are evaluated at the same dense-reference coordinates. This comparison is limited to the specified anchor construction, pool size, and batch probability.

E Discrete Constraint Projection

E.1 Discrete Representation of the Continuous Sampling Model

The population formulation models a law on trajectories and a sampling kernel $r(dy \mid f)$ on timestamp–value records. A numerical model instead uses an interpolation map $I_\Delta : \mathbb{R}^J \to \mathcal{H}$. For a coefficient law p_Δ, its represented trajectory law is $(I_\Delta)_{\#}p_\Delta$, and its record law is

$$T_{r,\Delta}p_\Delta = \int r(\cdot \mid I_\Delta x)\, p_\Delta(dx).$$

Thus the grid-based objective is

$$\min_{p_\Delta:\, p_\Delta(\{x:I_\Delta x\in\mathcal{G}\})=1} D_{\mathrm{KL}}\left(q \parallel T_{r,\Delta}p_\Delta\right).$$

It approximates the continuous problem by restricting the represented trajectory family. Its optimal value need not be zero when that family cannot reproduce the target record distribution.

For a linear basis, the norm matrix has entries $(W_\Delta)_{j\ell} = \langle I_\Delta e_j, I_\Delta e_\ell\rangle_{\mathcal{H}}$. For a record $y = ((t_i, v_i))_{i=1}^n$, define the evaluation matrix by $(B_y x)_i = (I_\Delta x)(t_i)$ and let $v = (v_1, \ldots, v_n)$. Observation consistency is $B_y x = v$, not generally equality of selected nodal coordinates. Computational nodes may be refined or augmented to represent observations exactly. If the chosen representation cannot fit all observed values, exact feasibility is not available; a measurement error model or a richer representation must be specified explicitly.

Each channel may use its own computational nodes and evaluation operator. Cross-channel constraints are evaluated through the represented functions at common times rather than by treating measurements recorded at different times as simultaneous, and each observed value remains fixed at its actual timestamp. The blood-pressure example in Section E.3 illustrates this construction.

For piecewise-linear interpolation, neighboring slope bounds imply a Lipschitz bound on the represented function, and convex pointwise bounds hold between feasible endpoints. These statements concern the interpolant, not the unknown trajectory. General continuous constraints require their own discretization or between-node checks. In particular, the continuous convergence theorem does not establish convergence as grid spacing tends to zero; that requires approximation and stability assumptions beyond the present results.

E.2 Projection onto General Affine Constraints

Using the finite-dimensional representation in Section E.1, eliminate the observation equalities $B_y x = v$ and express the remaining affine medical constraints as slabs

$$|a^\top z - b| \le c, \tag{25}$$

where z denotes free coefficients in an orthonormal nullspace basis for the observation equalities. For a general norm matrix W_Δ, the corresponding quadratic program uses its induced metric.

If

$$|a^\top \hat{z} - b| \le c,$$

the prediction is already feasible and

$$\boxed{z^+ = \hat{z}.}$$

Otherwise, define

$$v = a^\top \hat{z} - b.$$

The Euclidean projection is

$$\boxed{z^+ = \hat{z} - \frac{v - \operatorname{sgn}(v)c}{\|a\|_2^2} a.} \tag{26}$$

For several affine constraints,

$$Az \le b(y),$$

the projection becomes the small convex quadratic program

$$\boxed{z^+ = \arg\min_z \frac{1}{2}\|z - \hat{z}\|_2^2 \quad \text{s.t.} \quad Az \le b(y).} \tag{27}$$

Thus physiological bounds, temporal constraints, ordering relationships, and cross-channel relations can be enforced jointly without introducing an additional learned network.

E.3 Worked Example: MAP, SBP, and DBP

Consider multiple channels at a time t where predictions for those channels are requested. A value is observed here only if that channel was measured at t; otherwise it is generated for that channel at time t. Measurements taken at other times remain fixed at their own timestamps. We omit the time index below.

The implementation represents the approximate physiological relation among mean arterial pressure (MAP), systolic blood pressure (SBP), and diastolic blood pressure (DBP) with the source-specific convex constraint

$$\boxed{\left| MAP - \frac{SBP + 2DBP}{3} - c_{\text{src}} \right| \le \epsilon_{\text{src}}.} \tag{28}$$

The source-dependent offset c_{src} and tolerance ϵ_{src} are specified with the other deployed projection settings in Section C.2. The constraint is activated only if at least one of the channels involved is generated.

All three channels observed. If all three channels were measured at time t, their values are retained. Under the assumed relation, feasibility requires those recorded values to satisfy it; projection does not change measurements to repair an incompatible constraint.

Only MAP generated. If SBP and DBP are observed but MAP is missing, then

$$MAP \in \left[\frac{SBP_{\text{obs}} + 2DBP_{\text{obs}}}{3} + c_{\text{src}} - \epsilon_{\text{src}},\ \frac{SBP_{\text{obs}} + 2DBP_{\text{obs}}}{3} + c_{\text{src}} + \epsilon_{\text{src}} \right].$$

Hence

$$\boxed{MAP^+ = \operatorname{clip}\left(\widehat{MAP}, \mu_{\text{MAP,src}} - \epsilon_{\text{src}}, \mu_{\text{MAP,src}} + \epsilon_{\text{src}} \right),} \tag{29}$$

where

$$\mu_{\text{MAP,src}} = \frac{SBP_{\text{obs}} + 2DBP_{\text{obs}}}{3} + c_{\text{src}}.$$

The observed SBP and DBP values are unchanged.

Multiple channels generated. Suppose SBP is observed while DBP and MAP are generated. Substituting SBP_{obs} into Eq. (28) gives two affine inequalities in the generated pair (DBP, MAP). We apply the general quadratic program in Eq. (27), together with the ordering and range constraints, so the generated values are corrected jointly.