RECOVERABILITY-AWARE IMPUTATION WITH CONDITIONAL BAYES RISK ESTIMATION FOR MULTIVARIATE TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

The problem of missing values in time series is usually framed as a prediction problem where, given the contextual information, the model predicts every missing value. Though conceptually correct this approach overlooks a crucial aspect: the missing regions may differ in how recoverable they are given the available information. In this paper, we propose a recoverability-aware perspective of time series imputation that takes into account not only what to impute but also the uncertainty in this imputation. In our framework, the unrecoverability of the missing region is quantified via the conditional Bayes risk of the corresponding mask-aware missing region through a Contextual UnRecoverability Estimator (CURE) that learns this quantity from incomplete sequences. We then use the standard Bayes-risk decomposition to combine CURE with an imputer-specific residual-risk correction for selective imputation. Our experiments on synthetic dynamical systems demonstrate that regions with same size can have different levels of oracle unrecoverability for different generating configurations, and the imputers' errors correlate highly with the normalized oracle risk. Across five real-world datasets, seven imputers, MCAR and MAR missingness, and point and block corruption, CURE achieves a mean error-ranking correlation of 0.372, which increases to 0.507 after incorporating the imputer-specific correction. Total-risk ranking reduces AURC on average by 58.7% relative to random selection, compared with 53.8% for the strong imputer–CURE disagreement baseline, and yields lower AURC than disagreement in 56.3% of matched evaluations. For 50% coverage, selective RMSE is reduced by 38.4–62.7% for all five datasets relative to full coverage. These results support recoverability as a measurable basis for identifying which imputations can be trusted

1 INTRODUCTION

The central question in time-series imputation is not just what value should fill a gap, but whether that value is recoverable from the observations that remain. The distinction is pertinent as products of imputation models determine decision in medicine (Lu et al., 2023), environment (Lee et al., 2026), transportation (Zhang et al., 2024), and energy (Hwang & Suh, 2024) domains. Reconstrucrion of missing data in time series has rapidly advanced through modeling temporal and variable dependencies in recurrent neural networks, attention mechanisms, latent variable modeling and conditional generative methods (Cao et al., 2018; Fortuin et al., 2020; Tashiro et al., 2021; Du et al., 2023). Recent studies pose architectural and uncertainty modeling as central axes in the field(Wang et al., 2025). Yet reconstruction accuracy alone does not reveal whether an error arises because the imputer failed to exploit available information or because the observed context did not contain enough information for accurate recovery. Two regions of equal length or missing rates can appear in different temporal states with different boundary information, recurrence, and cross-variable information and therefore be of different difficulty. The unresolved challenge is to characterize the recoverability of a missing region from its available context separately from the additional error incurred by using an imputer of choice.

Existing methods approach this distinction indirectly. Deterministic imputers learn dependencies through reconstruction objectives (Cao et al., 2018; Du et al., 2023; Hu et al., 2024), while proba-

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.

bilistic models such as GP-VAE and CSDI represent distributions over plausible completions (Fortuin et al., 2020; Tashiro et al., 2021). Selective imputation methods further recognize that not every imputation should be retained (Blázquez-García et al., 2023; Weatherhead & Goldenberg, 2025), but their selection scores are obtained from uncertainty or errors associated with a fitted procedure. Such model-derived uncertainty need not coincide with reconstruction accuracy (Hossain & Milani, 2025). These developments leave a different question open, illustrated in Figure 1: *is a missing region difficult because the available observations leave its value ambiguous, or because a particular imputer introduces additional error?*

We study this question through the standard decomposition of conditional prediction risk into Bayes risk and model-specific excess risk (Lahlou et al., 2021; Kotelevskii et al., 2025). We operationalize the Bayes-risk term for missing-region imputation as *contextual unrecoverability*, defined relative to the observations available within the imputation context. We then introduce CURE (Contextual UnRecoverability Estimator) to estimate this quantity when oracle conditional moments are unavailable and learn an imputer-specific correction for estimating total conditional risk. The resulting scores support selective imputation by ranking missing regions according to their expected reconstruction risk. We evaluate the formulation first in controlled systems where conditional Bayes risk is available and then on five datasets with seven imputers under MCAR and MAR point and block masking. The real data evaluations concern artificially hidden entries that were originally observed. Applying the same interpretation to naturally missing entries additionally requires ignorability of the native missingness mechanism given the observed context.

Contributions.

- We formulate conditional Bayes risk as a missing region recoverability target that accounts for the observed mask, target region, and available time-series context, and distinguish it from imputer-specific excess risk.
- We design CURE to estimate contextual unrecoverability from incomplete sequences and combine it with an imputer-specific risk correction to rank point and block missing regions for selective imputation.
- We construct an oracle grounded evaluation across four controlled dynamical families and study five real datasets with seven imputers, testing whether recoverability captures a component of missing-region difficulty shared across models and whether model-specific risk improves selective decisions.

2 Related Work

Time series imputation in modern models involves recurrent, attention, latent variables, and generative models (Cao et al., 2018; Du et al., 2023; Fortuin et al., 2020; Tashiro et al., 2021; Hu et al., 2024). Despite their architectural differences, these approaches are designed to improve reconstruction capabilities of models. In contrast, our main concern lies in the estimation of the reconstruction risk that can be attributed to the context availability and the residual risk attributed to the imputer itself.

Risk-based uncertainty decomposition. The difference between irreducible uncertainty and model-specific bias is well known in the context of predictive uncertainty. The term aleatoric uncertainty usually refers to uncertainty that remains after conditioning on the available information, while epistemic uncertainty concerns the limitations of the predictor (Hüllermeier & Waegeman, 2021). Under squared loss, this distinction can be expressed through the standard decomposition of conditional prediction risk into Bayes risk and excess risk. DEUP (Lahlou et al., 2021) suggests an approach of this type by estimating generalization error and its aleatoric component, while newer insights offer formulations of predictive uncertainty within a decision-theoretic framework (Kotelevskii et al., 2025; Smith et al., 2025). Therefore, we do not claim that the Bayes-risk decomposition is novel in itself. Our contribution is to instantiate it for missing-region imputation, where the information set is determined jointly by the incomplete sequence, native observation mask, and target region. CURE estimates the resulting contextual Bayes-risk term, while a separate imputer-specific estimator models residual risk for selective imputation.

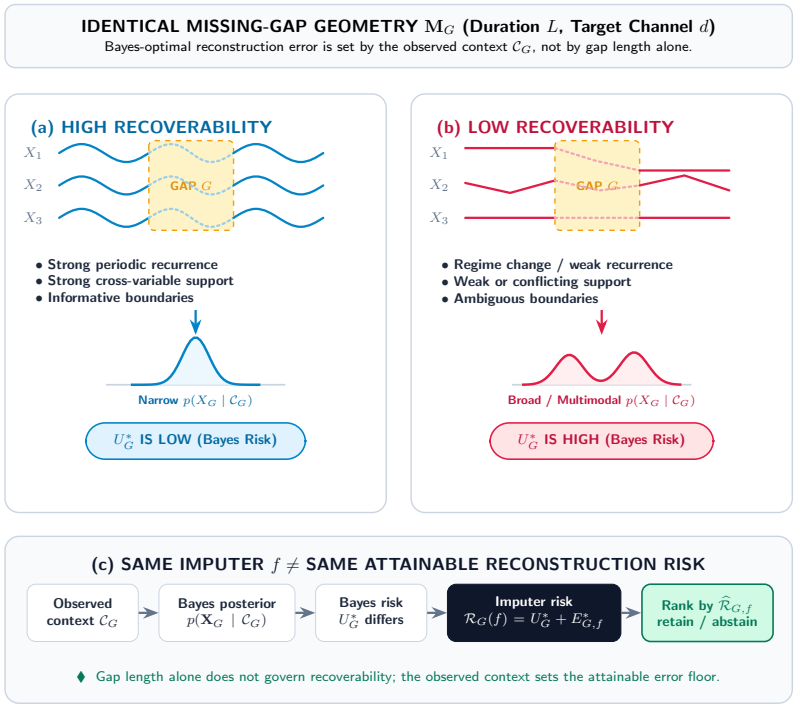


Figure 1: **Missingness geometry does not determine recoverability.** Equally sized missing regions can differ in the information supplied by recurrence, cross-variable redundancy, and boundary evidence. We define the resulting context-dependent minimum prediction risk as contextual unrecoverability U_G^*.

Uncertainty-aware and selective imputation. Uncertainty-aware and selective imputation Probability-based imputers like GP-VAE and CSDI generate probability distributions over reasonable imputations (Fortuin et al., 2020; Tashiro et al., 2021). Still, uncertainties derived from the model fitted do not necessarily correspond with the difficulty of reconstruction or the calibration of the model (Hossain & Milani, 2025). Selective imputation overcomes this problem by allowing some imputations to be dropped out. The work of Blázquez-García et al. (2023) incorporates the uncertainty of imputation along with fidelity of representation, whereas that of Weatherhead & Goldenberg (2025) keeps the predictions considering uncertainties of the model. Similarly, the earlier work on reject options and selective prediction makes use of predictive risk and coverage in the same way (El-Yaniv & Wiener, 2010; Shah et al., 2022) These approaches rank predictions using uncertainty associated with a fitted procedure. We instead first estimate the Bayes-risk component associated with the observed context and then incorporate imputer-specific excess risk. This distinction lets us test whether part of missing-region difficulty is shared across heterogeneous imputers and whether accounting for the remaining model-specific component improves selective decisions.

3 Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{T \times D}$ be a multivariate time-series window with native observation mask $\mathbf{M}^0 \in \{0,1\}^{T \times D}$. For a target region G with mask $\mathbf{R}_G$, hiding G gives $\mathbf{M}_G = \mathbf{M}^0 \odot (1 - \mathbf{R}_G)$ and context

$$\mathcal{C}_G = (\widetilde{\mathbf{X}}_G, \mathbf{M}_G, \mathbf{R}_G),$$

where $\widetilde{\mathbf{X}}_G = \mathbf{X} \odot \mathbf{M}_G$. Recoverability is therefore defined relative to the observations available within the input window. Evaluation is restricted to $G_{\text{eval}} = \{(t,d) \in G : M^0_{td} = 1\}$ using

$$\ell_G(\mathbf{a}, \mathbf{b}) = |G_{\text{eval}}|^{-1} \|\mathbf{a} - \mathbf{b}\|_2^2. \tag{1}$$

We assume the artificial masking design is ignorable,

$$P(A = G \mid \mathbf{X}_G, \mathbf{X}_{G^c}, \mathbf{M}^0) = P(A = G \mid \mathbf{X}_{G^c}, \mathbf{M}^0), \tag{2}$$

with positivity. Our MCAR and MAR constructions satisfy this requirement: MCAR masking is independent of the data, while MAR depends only on anchor variables that remain observed (Yoon et al., 2018; Jarrett et al., 2022). Real data claims therefore concern artificially hidden entries that were originally observed. Transfer to natively missing entries additionally requires ignorability of the native missingness mechanism given the observed context. Identification under MNAR requires additional assumptions (Ma & Zhang, 2021) and is outside the present scope.

We define *contextual unrecoverability* as the conditional Bayes risk

$$U_G^*(\mathcal{C}_G) = \inf_g \mathbb{E}[\ell_G(\mathbf{X}_{G_{\text{eval}}}, g(\mathcal{C}_G)) \mid \mathcal{C}_G]. \tag{3}$$

Under squared loss, the Bayes predictor is $\mathbf{X}^*_{G_{\text{eval}}} = \mathbb{E}[\mathbf{X}_{G_{\text{eval}}} \mid \mathcal{C}_G]$, giving

$$U_G^*(\mathcal{C}_G) = \frac{1}{|G_{\text{eval}}|} \operatorname{tr} \operatorname{Cov}(\mathbf{X}_{G_{\text{eval}}} \mid \mathcal{C}_G). \tag{4}$$

This is the standard Bayes-risk component under squared loss (Berger, 1985; Lahlou et al., 2021; Kotelevskii et al., 2025).

For a fixed imputer f, define

$$\mathcal{R}_G(f \mid \mathcal{C}_G) = \mathbb{E}[\ell_G(\mathbf{X}_{G_{\text{eval}}}, f(\mathcal{C}_G)) \mid \mathcal{C}_G].$$

The standard squared-loss decomposition gives

$$\mathcal{R}_G(f \mid \mathcal{C}_G) = U_G^*(\mathcal{C}_G) + E^*_{G,f}(\mathcal{C}_G), \qquad E^*_{G,f} = \frac{\|f(\mathcal{C}_G) - \mathbf{X}^*_{G_{\text{eval}}}\|_2^2}{|G_{\text{eval}}|}. \tag{5}$$

We use this established decomposition as the organizing lens for the method.

In practice, given an estimate $\widehat{U}_G$, the calibration target $Z_{G,f} = \ell_G(\mathbf{X}_{G_{\text{eval}}}, \widehat{\mathbf{X}}_{G,f}) - \widehat{U}_G$ satisfies

$$\mathbb{E}[Z_{G,f} \mid \mathcal{C}_G] = E^*_{G,f} + (U^*_G - \widehat{U}_G). \tag{6}$$

Thus, a learned correction $\widehat{E}_{G,f}$ may capture both imputer-specific excess risk and error in $\widehat{U}_G$. We therefore use $\widehat{\mathcal{R}}_{G,f} = \widehat{U}_G + \widehat{E}_{G,f}$ primarily as an estimate of total conditional risk. At fixed coverage, the oracle ranking for a given imputer is determined by $U^*_G + E^*_{G,f}$, while practical selective imputation ranks units by $\widehat{\mathcal{R}}_{G,f}$ from lower to higher estimated risk.

4 Learning Recoverability and Imputer-Specific Risk

Given target region G, the framework estimates the two quantities introduced in Section 3. CURE estimates contextual unrecoverability $\widehat{U}_G$, while a second model learns an imputer-specific residual-risk correction. Their sum forms the estimated conditional risk used for selective imputation (Figure 2).

4.1 CURE and Artificial-Holdout Learning

CURE uses a Transformer encoder (Vaswani et al., 2017) over the aligned inputs $(\widetilde{\mathbf{X}}_G, \mathbf{M}_G, \mathbf{R}_G)$. Missing entries are filled with zeros after mask construction, and the concatenated inputs are projected into a latent space with a learned temporal embedding:

$$\mathbf{H} = \mathrm{Enc}_\theta\Big(\mathrm{Proj}([\widetilde{\mathbf{X}}_G, \mathbf{M}_G, \mathbf{R}_G]) + \mathbf{P}\Big). \tag{7}$$

Two heads produce the conditional mean and a positive variance,

$$\widehat{\boldsymbol{\mu}} = h_\mu(\mathbf{H}), \qquad \widehat{\mathbf{v}} = \mathrm{softplus}(h_v(\mathbf{H})) + \epsilon. \tag{8}$$

For a point unit, $\widehat{U}_G = \widehat{v}_j$; for a block, $\widehat{U}_G$ is the mean predicted variance over its evaluable entries, following the Bayes-risk form in Eq. 4.

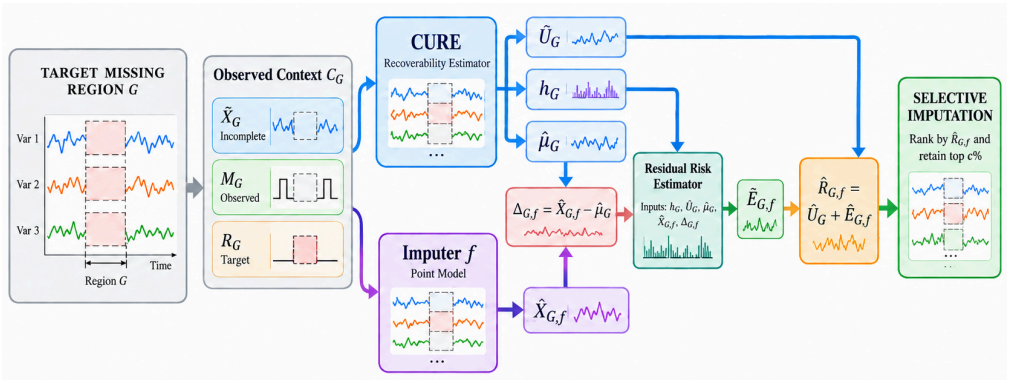


Figure 2: **Recoverability-aware selective imputation.** CURE estimates conditional mean $\widehat{\boldsymbol{\mu}}_G$, contextual unrecoverability $\widehat{U}_G$, and representation $\mathbf{h}_G$. For imputer f, a lightweight model estimates a residual-risk correction $\widehat{E}_{G,f}$, which is combined with $\widehat{U}_G$ to estimate total conditional risk.

CURE is trained on artificial holdouts from originally observed entries using

$$\mathcal{L}_{\text{CURE}} = \frac{1}{2|G_{\text{eval}}|} \sum_{j \in G_{\text{eval}}} \left[\frac{(X_j - \widehat{\mu}_j)^2}{\widehat{v}_j} + \log \widehat{v}_j \right]. \tag{9}$$

At population level, for positive conditional variance and finite second moments, this objective is minimized by the conditional mean and variance, without requiring the true conditional distribution to be Gaussian (Proposition 1 in Appendix A).

Training uses both point and block recovery tasks with equal probability and samples severity over the full experimental range, reducing mismatch between training and test masking patterns. Each training sequence contributes two recovery tasks. CURE is trained independently of the comparison imputers, so the same $\widehat{U}_G$ is used for every imputer evaluated on a given missing unit.

4.2 Imputer-Specific Residual-Risk Estimation

For reconstruction $\widehat{\mathbf{X}}_{G,f} = f(\mathcal{C}_G)$, define its deviation from CURE's estimated conditional mean as

$$\boldsymbol{\Delta}_{G,f} = \widehat{\mathbf{X}}_{G,f} - \widehat{\boldsymbol{\mu}}_G. \tag{10}$$

A lightweight MLP receives the CURE representation, recoverability estimate, conditional-mean estimate, imputer reconstruction, their signed, absolute, and squared deviations, and context descriptors:

$$\mathbf{z}_{G,f} = [\mathbf{h}_G, \widehat{U}_G, \widehat{\boldsymbol{\mu}}_G, \widehat{\mathbf{X}}_{G,f}, \boldsymbol{\Delta}_{G,f}, |\boldsymbol{\Delta}_{G,f}|, \boldsymbol{\Delta}^2_{G,f}, \mathbf{q}_G], \tag{11}$$

where $\mathbf{q}_G$ contains observation fraction, missingness geometry, severity, and target-variable information.

The estimator is fitted on a calibration split disjoint from imputer training, validation, and testing, using the residual target $Z_{G,f}$ defined in Eq. 6. As that equation shows, the fitted correction can contain both imputer-specific excess risk and systematic error in $\widehat{U}_G$. We therefore interpret $\widehat{E}_{G,f}$ as a residual-risk correction whose main purpose is to improve estimation of total conditional risk, not as an identified estimate of oracle $E^*_{G,f}$.

This stage is related to the excess-risk estimation in (Lahlou et al., 2021), but differs operationally in how the correction is conditioned and used. A separate estimator is calibrated for each imputer, dataset, missingness mechanism, and training seed. Its input in Eq. 11 combines CURE's mask-aware representation with the imputer prediction and its deviation from CURE's conditional-mean estimate. The resulting score estimates imputer-specific conditional risk for point and block missing regions and is used to rank those regions for selective imputation, rather than as a generic predictive-uncertainty score.

Negative calibration targets are retained. At inference, the predicted correction is constrained as $\widehat{E}_{G,f} = \max\{0, r_{\phi_f}(\mathbf{z}_{G,f})\}$, reflecting the non-negativity of oracle excess risk. Since Eq. 6 also permits negative residual corrections when $\widehat{U}_G$ overestimates U_G^*, this clipping is used as an implementation constraint.

At test time, $\widehat{U}_G$ is computed once for each missing unit and reused across imputers. For imputer f, we form $\widehat{\mathcal{R}}_{G,f} = \widehat{U}_G + \widehat{E}_{G,f}$ and rank units from lower to higher estimated conditional risk, as defined in Section 3. Architecture, calibration, and optimization details are given in Appendix E.

5 Experimental Setup

Data and imputers. Synthetic experiments use Gaussian AR/VAR, periodic and quasi-periodic, regime-switching, and nonlinear systems for which conditional Bayes risk and the Bayes predictor are available from the data-generating process. Real data experiments use ETTh1, ETTh2, ETTm1, ETTm2, and Beijing Air Quality, obtained and processed through BenchPOTS (Du, 2023). We evaluate linear interpolation (Lerp)(Du, 2023), BRITS(Cao et al., 2018), SAITS(Du et al., 2023), CSDI(Tashiro et al., 2021), GP-VAE(Fortuin et al., 2020), TimesNet(Wu et al., 2022), and ImputeFormer(Nie et al., 2024). Dataset statistics, preprocessing, and model configurations are reported in Appendices C–D.

Missingness and training protocol. We evaluate MCAR and MAR under point rates $\{0.10, 0.30, 0.50, 0.70\}$ and block lengths $\{T/12, T/6, T/4, T/2\}$. MAR point masking depends only on retained anchor variables, while MAR block onset depends on pre-gap anchor observations. Native missing entries are never used as ground truth. CURE training samples point and block holdouts with equal probability and draws severity across the full experimental range. Imputers and CURE are trained on the training split, model selection uses validation data, and each imputer-specific residual-risk estimator is fitted on a disjoint calibration split. Detailed mask construction and split discipline are given in Appendix E.

Synthetic evaluation. The synthetic study tests more than correlation with raw oracle risk. First, we compare missing regions at matched gap length and separately evaluate scale-normalized recoverability to distinguish contextual structure from changes in process scale. Second, we fix the gap length at $L = 24$ to remove gap size as an explanation for error ordering. Third, we relate configuration-level U_G^* to known properties of each generating process, including noise, cross-variable coupling, periodic recurrence, and regime observability. For Gaussian VAR and periodic systems, conditional variance is determined by the generating configuration and mask geometry, so their fixed-length variation is interpreted across configurations, not as value-dependent variation within a fixed configuration. Finally, the oracle study evaluates the exact risk decomposition in Eq. 5 and compares selective ranking by learned recoverability with gap length and random ordering. Full diagnostics are reported in Appendix G.

Real data evaluation. Recoverability is evaluated using Spearman correlation between $\widehat{U}_G$ and realized unit-level squared error, with the corresponding correlation reported for $\widehat{\mathcal{R}}_{G,f}$. Selective performance uses risk–coverage curves and AURC, comparing $\widehat{U}_G$, total risk, imputer–CURE disagreement $D_{G,f} = |G_{\text{eval}}|^{-1}\|\widehat{\mathbf{X}}_{G,f} - \widehat{\boldsymbol{\mu}}_G\|_2^2$, and random ordering. Point corruptions form scalar selection units, while each feature-specific contiguous block forms one unit whose error is averaged over evaluable entries. These are the four selection scores implemented in the real-data evaluation.

Replication and aggregation. We use three training seeds. The ETT datasets use five independent test-mask seeds and Beijing Air Quality uses three completed test-mask seeds. Test masks are shared across imputers. Recoverability correlations are aggregated into $5 \times 2 \times 8 \times 7 = 560$ dataset–mechanism–missingness–imputer conditions. Selective analyses retain replicate structure, giving $1{,}104$ test settings, $7{,}728$ imputer-level evaluations, and $23{,}184$ within-setting imputer-pair comparisons. We therefore report both aggregate summaries and the distribution of replicate-level selective performance rather than treating the evaluations as independent repetitions.

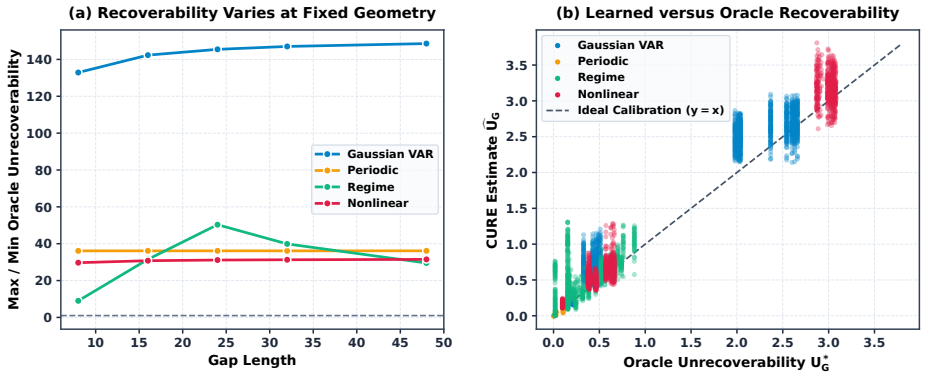


Figure 3: **Synthetic recoverability validation. (a)** Maximum-to-minimum oracle unrecoverability across generating configurations for gaps of the same length. Large ratios show that gap length alone does not determine conditional Bayes risk. **(b)** CURE estimates versus oracle conditional Bayes risk across the four synthetic process families.

6 Results and Discussion

6.1 Does Conditional Bayes Risk Characterize Missing-Region Difficulty, and Can CURE Estimate It?

For matched gap lengths, oracle unrecoverability varies markedly across generating configurations, with max/min ratios of $148.6\times$, $36.1\times$, $50.3\times$, and $31.5\times$ (Figure 3) for Gaussian VAR, periodic, regime-switching, and nonlinear dynamics. For Gaussian VAR and periodic data, these differences arise across generating configurations, not from value-dependent uncertainty within a fixed configuration. After scale normalization, oracle recoverability remains associated with practical-imputer error (Spearman 0.726–0.978). Known process factors also behave as expected: greater noise increases U_G^*, whereas stronger coupling and regime observability reduce it.

The exact conditional risk $U_G^* + E_{G,f}^*$ tracks realized practical-imputer error ($\rho = 0.927$–0.996), supporting the decomposition. CURE preserves oracle ordering ($\rho = 0.793$–0.928), and at fixed $L = 24$, learned unrecoverability remains correlated with error (0.672–0.900), indicating information beyond gap length. In synthetic selective evaluation, learned recoverability beats random in all 16 family–imputer combinations, reduces AURC by a median 59.8%, and outperforms gap length in $14/16$ comparisons. Absolute calibration is less uniform, so we emphasize ranking performance. Detailed parameter-level results are reported in Appendix G.

6.2 Does Learned Recoverability Predict Imputation Difficulty on Real Data?

Across the 560 aggregated conditions, $\widehat{U}_G$ is positively associated with realized error in 96.8% of cases, with mean Spearman correlation 0.372 (median 0.365). Associations are strongest and most uniform on Beijing Air Quality; ETTm2 is stronger on average but varies across imputers, while ETTh1 and several ETTh2 MAR conditions are weaker (Figure 4a). The 18 negative conditions show that recoverability is not a deterministic surrogate for realized error.

6.3 Is Imputation Difficulty Shared Across Different Imputers?

Across 23,184 within-setting imputer pairs, mean cross-imputer difficulty correlation is 0.450 and 98.0% are positive. Agreement is stronger for block than point missingness (0.533 vs. 0.367), with pairwise means ranging from 0.17 (Lerp–GP-VAE) to 0.67 (GP-VAE–TimesNet) in Figure 4b. The broadly positive agreement supports a difficulty component shared across models presented with the same observed context. The remaining variation is consistent with different imputers exploiting different contextual cues and motivates the imputer-specific excess-risk term.

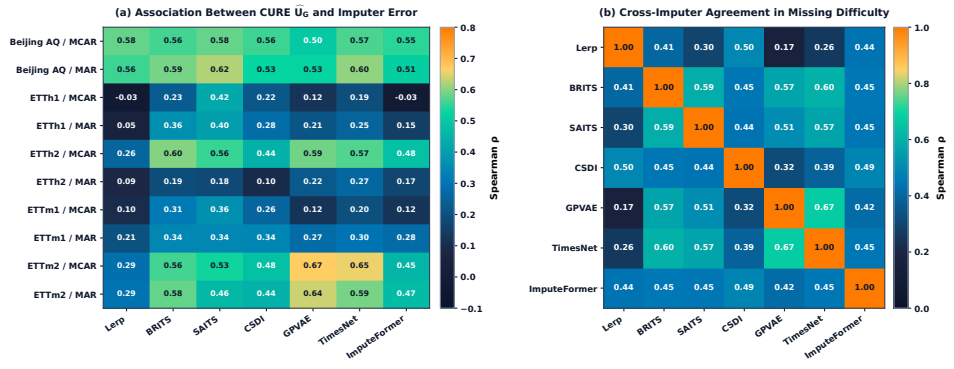


Figure 4: **Recoverability and cross-imputer difficulty.** (a) Mean Spearman correlation between $\widehat{U}_G$ and realized error by dataset, mechanism, and imputer. (b) Mean pairwise correlation between realized errors of imputers on shared missing units.

6.4 Does Imputer-Specific Excess Risk Improve Risk Estimation?

Adding $\widehat{E}_{G,f}$ increases mean error-ranking correlation from 0.372 for $\widehat{U}_G$ to 0.507 for $\widehat{\mathcal{R}}_{G,f}$; the median rises from 0.365 to 0.495, and total risk is better in 67.5% of the 560 conditions. The fraction with $\rho \geq 0.30$ rises from 60.7% to 90.9%. Figure 5 shows that the improvement is not universal, consistent with $\widehat{E}_{G,f}$ being a learned correction rather than oracle excess risk.

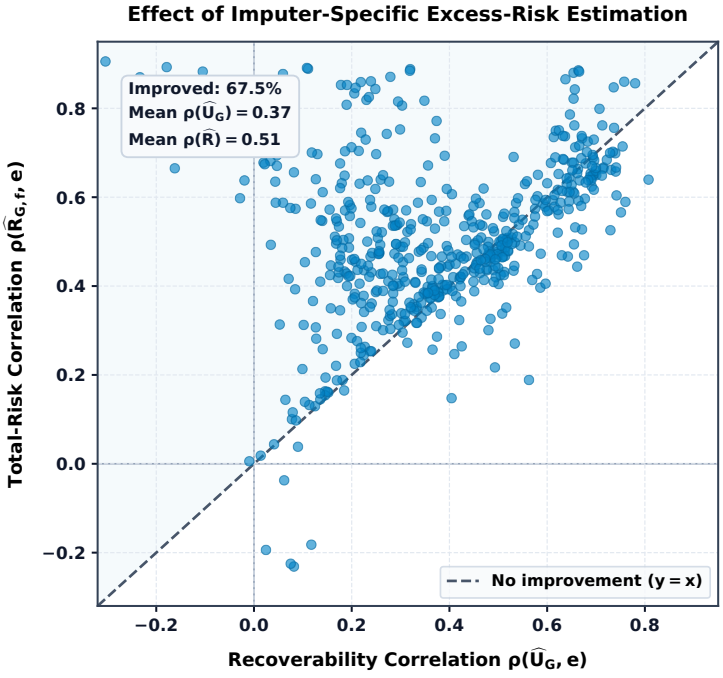


Figure 5: **Effect of imputer-specific excess-risk estimation.** Each point compares correlation with realized error using contextual unrecoverability alone (horizontal) and after adding estimated excess risk (vertical).

6.5 Does Recoverability-Aware Risk Improve Selective Imputation?

Across the $7{,}728$ replicate-level evaluations, ranking by $\widehat{U}_G$ reduces AURC relative to random selection by a median 43.8%. Incorporating the imputer-specific correction increases the median

Table 1: **Selective RMSE at 50% coverage.** Full denotes no abstention; lower is better. Reduction compares total-risk selection with full coverage.

Dataset	Full	Random	$\widehat{U}_G$	Disagree.	$\widehat{\mathcal{R}}_{G,f}$	Reduction
Beijing Air Quality	0.672	0.646	0.320	0.356	**0.250**	62.7%
ETTh1	0.636	0.631	0.568	0.416	**0.392**	38.4%
ETTh2	0.777	0.758	0.513	**0.414**	0.440	43.3%
ETTm1	0.472	0.471	0.422	0.255	**0.248**	47.6%
ETTm2	0.439	0.437	0.262	**0.220**	0.236	46.2%

reduction to 58.7%, compared with 53.8% for imputer–CURE disagreement. Figure 6 shows the distribution of these gains across training seeds, test masks, missingness settings, datasets, and imputers. Improvements over random occur in 88.3%, 97.8%, and 97.0% of matched evaluations, respectively. Total risk also gives lower AURC than disagreement in 56.3% of cases, with a median advantage of 5.2%, showing an average benefit without universal dominance.

These ranking gains translate into lower imputation error. At 50% coverage, total-risk selection reduces selective RMSE by 38.4%–62.7% relative to full coverage across all five datasets (Table 1), while random retention changes RMSE only slightly. Total risk gives the lowest RMSE on Beijing Air Quality, ETTh1, and ETTm1, whereas disagreement is lower on ETTh2 and ETTm2. At 50%, 70%, and 90% coverage, total-risk ranking reduces retained risk relative to random by median 64.7%, 52.8%, and 30.9%, respectively. Together, these results show that improved risk ranking translates into meaningful accuracy gains, while the best selection signal remains imputer dependent.

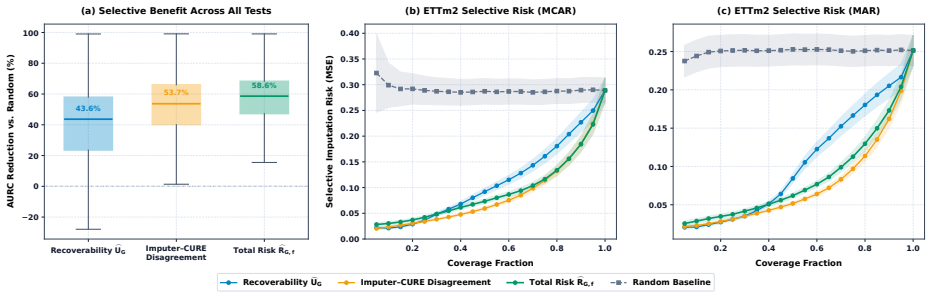


Figure 6: **Selective-imputation performance.** (a) AURC reduction relative to random selection across real-data test conditions. (b-c) Representative risk–coverage curves on ETTm2 dataset Across MCAR and MAR.

7 Conclusion

Multivariate time-series imputation is usually framed as estimating values for missing observations, but attainable accuracy also depends on how much information the remaining context provides about a missing region. We operationalized this idea using conditional Bayes risk over mask-aware missing regions and introduced CURE to estimate contextual unrecoverability from incomplete sequences. Combined with an imputer-specific residual-risk correction, the framework provides an estimated conditional risk for selective imputation.

Synthetic experiments showed substantial variation in oracle unrecoverability among equally sized regions across generating configurations, while scale-normalized and fixed-gap analyses confirmed that difficulty is not explained by geometry alone. On real data, mean error-ranking correlation increased from 0.372 for CURE to 0.507 after adding the imputer-specific correction. Total-risk ranking reduced AURC by a median 58.7% relative to random selection, compared with 53.8% for imputer–CURE disagreement, and reduced selective RMSE by 38.4–62.7% at 50% coverage. Recoverability remains relative to the available input context, and extending identification to naturally missing data under non-ignorable mechanisms and improving calibration under complex dynamics remain important directions.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs and providing feedback on research methodology or experiments. All AI-assisted readability edits were reviewed to ensure they preserved the intended meaning and technical content. All AI-assisted outputs were reviewed and validated by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition, 1985. ISBN 978-0-387-96098-2. doi: 10.1007/978-1-4757-4286-2.

Ane Blázquez-García, Kristoffer Wickstrøm, Shujian Yu, Karl Øyvind Mikalsen, Ahcene Boubekki, Angel Conde, Usue Mori, Robert Jenssen, and Jose A Lozano. Selective imputation for multivariate time series datasets with missing values. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9490–9501, 2023.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

Wenjie Du. PyPOTS: a Python toolbox for data mining on Partially-Observed Time Series. *arXiv preprint arXiv:2305.18811*, 2023.

Wenjie Du, David Côté, and Yan Liu. SAITS: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023. doi: 10.1016/j.eswa.2023.119619.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL `http://jmlr.org/papers/v11/el-yaniv10a.html`.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*, pp. 1651–1661. PMLR, 2020.

Zarin Tahia Hossain and Mostafa Milani. Beyond accuracy: An empirical study of uncertainty estimation in imputation. In *2025 IEEE International Conference on Big Data (BigData)*, pp. 7347–7356. IEEE, 2025.

Hanwen Hu, Shiyou Qian, Dingyu Yang, Jian Cao, and Guangtao Xue. Iterative time series imputation by maintaining dependency consistency. *ACM Transactions on Knowledge Discovery from Data*, 19(1):1–24, 2024.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Junhwa Hwang and Dongjun Suh. Cc-gain: Clustering and classification-based generative adversarial imputation network for missing electricity consumption data imputation. *Expert Systems with Applications*, 255:124507, 2024.

Daniel Jarrett, Bogdan C Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. Hyperimpute: Generalized iterative imputation with automatic model selection. In *International conference on machine learning*, pp. 9916–9937. PMLR, 2022.

Nikita Kotelevskii, Vladimir Kondratyev, Martin Takáč, Eric Moulines, and Maxim Panov. From risk to uncertainty: Generating predictive uncertainty measures via bayesian estimation. In *International Conference on Learning Representations*, volume 2025, pp. 92468–92504, 2025.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. Deup: Direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501*, 2021.

Jangho Lee, Max Berkelhammer, Joseph O'Brien, Gavin McNicol, Anna E. S. Vincent, Maxwell Grover, Aaron I. Packman, Bilal Kaludi, Ahram Cho, and Miquel Gonzalez-Meler. Imputation of urban environmental sensor data using gated attention bidirectional long short-term memory (GA-BiLSTM): methods, performance, and implications. *Environmental Monitoring and Assessment*, 198(3):262, feb 2026. ISSN 1573-2959. doi: 10.1007/s10661-026-15112-8. URL https://doi.org/10.1007/s10661-026-15112-8.

Chang Lu, Chandan K Reddy, Ping Wang, Dong Nie, and Yue Ning. Multi-label clinical time-series generation via conditional gan. *IEEE Transactions on Knowledge and Data Engineering*, 36(4): 1728–1740, 2023.

Chao Ma and Cheng Zhang. Identifiable generative models for missing not at random data imputation. *Advances in Neural Information Processing Systems*, 34:27645–27658, 2021.

Tong Nie, Guoyang Qin, Wei Ma, Yuewen Mei, and Jian Sun. Imputeformer: Low rankness-induced transformers for generalizable spatiotemporal imputation. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2260–2271, 2024.

Abhin Shah, Yuheng Bu, Joshua K Lee, Subhro Das, Rameswar Panda, Prasanna Sattigeri, and Gregory W Wornell. Selective regression under fairness criteria. In *International Conference on Machine Learning*, pp. 19598–19615. PMLR, 2022.

Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark Van Der Wilk, Adam Foster, and Tom Rainforth. Rethinking aleatoric and epistemic uncertainty. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24804–24816. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/cfe8504bda37b575c70ee1a8276f3486-Paper.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jun Wang, Wenjie Du, Yiyuan Yang, Linglong Qian, Wei Cao, Keli Zhang, Wenjia Wang, Yuxuan Liang, and Qingsong Wen. Deep learning for multivariate time series imputation: a survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/1187. URL https://doi.org/10.24963/ijcai.2025/1187.

Addison Weatherhead and Anna Goldenberg. Impute with confidence: A framework for uncertainty aware multivariate time series imputation. *arXiv preprint arXiv:2507.09353*, 2025. doi: 10.48550/arXiv.2507.09353.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pp. 5689–5698. PMLR, 2018.

Yimei Zhang, Xiangjie Kong, Wenfeng Zhou, Jin Liu, Yanjie Fu, and Guojiang Shen. A comprehensive survey on traffic missing data imputation. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):19252–19275, 2024.

A Moment Consistency of the CURE Objective

The heteroscedastic Gaussian objective used by CURE does not require the conditional distribution of a missing value to be Gaussian. The following result shows that, at population level, it targets the conditional mean and variance required by Eq. 4.

Proposition 1 (Conditional moment consistency)**.** *Let Y denote a target entry and let $\mathcal{C}$ denote its available context. Assume $\mathbb{E}[Y^2 \mid \mathcal{C}] < \infty$ and $\mathrm{Var}(Y \mid \mathcal{C}) > 0$. For measurable functions $\mu(\mathcal{C})$ and $v(\mathcal{C}) > 0$, consider the conditional population loss*

$$\mathcal{L}(\mu, v \mid \mathcal{C}) = \frac{1}{2}\mathbb{E}\left[\frac{(Y-\mu)^2}{v} + \log v \;\middle|\; \mathcal{C}\right].$$

Its unique minimizer is

$$\mu^*(\mathcal{C}) = \mathbb{E}[Y \mid \mathcal{C}], \qquad v^*(\mathcal{C}) = \mathrm{Var}(Y \mid \mathcal{C}).$$

Hence, Gaussian negative log-likelihood is consistent for the first two conditional moments even when the true conditional distribution of $Y \mid \mathcal{C}$ is non-Gaussian.

Proof. Let

$$m(\mathcal{C}) = \mathbb{E}[Y \mid \mathcal{C}], \qquad \sigma^2(\mathcal{C}) = \mathrm{Var}(Y \mid \mathcal{C}).$$

Conditioning on $\mathcal{C}$ and using the bias–variance identity gives

$$\mathbb{E}[(Y-\mu)^2 \mid \mathcal{C}] = \sigma^2(\mathcal{C}) + \big(\mu - m(\mathcal{C})\big)^2.$$

Therefore,

$$\mathcal{L}(\mu, v \mid \mathcal{C}) = \frac{1}{2}\left[\frac{\sigma^2(\mathcal{C}) + (\mu - m(\mathcal{C}))^2}{v} + \log v\right].$$

For any fixed $v > 0$, this expression is uniquely minimized at $\mu = m(\mathcal{C})$. Substituting this value yields

$$\mathcal{L}(m, v \mid \mathcal{C}) = \frac{1}{2}\left[\frac{\sigma^2(\mathcal{C})}{v} + \log v\right].$$

Differentiating with respect to v gives

$$\frac{\partial \mathcal{L}}{\partial v} = \frac{v - \sigma^2(\mathcal{C})}{2v^2},$$

which is negative for $v < \sigma^2(\mathcal{C})$ and positive for $v > \sigma^2(\mathcal{C})$. Thus the unique minimum occurs at $v = \sigma^2(\mathcal{C})$. □

Because Eq. 9 is additive over target entries, the result applies componentwise to CURE's mean and variance heads. Averaging the resulting conditional variances over a region therefore targets the region-level Bayes risk in Eq. 4. In implementation, CURE uses $\widehat{v} = \mathrm{softplus}(\cdot) + \epsilon$ for numerical stability. In implementation, CURE uses $\widehat{v} = \mathrm{softplus}(\cdot) + \epsilon$, so the predicted variance satisfies $\widehat{v} > \epsilon$. When $\mathrm{Var}(Y \mid \mathcal{C}) > \epsilon$, the population minimizer remains exactly representable. If the conditional variance is at or below the numerical floor, the infimum is approached as $\widehat{v} \downarrow \epsilon$.

B Selection Implications of Conditional Risk

For completeness, we state the selection implication used in Section 3. Let $\mathcal{G} = \{G_1, \ldots, G_N\}$ denote candidate missing regions.

Proposition 2 (Optimal selection for a fixed imputer)**.** *For a fixed imputer f and coverage k/N, expected selective risk is minimized by retaining the k regions with smallest*

$$\mathcal{R}_G(f \mid \mathcal{C}_G) = U_G^*(\mathcal{C}_G) + E_{G,f}^*(\mathcal{C}_G).$$

For the Bayes predictor, $E_{G,f}^* = 0$, so this reduces to ranking by U_G^*.

Proof. Order the conditional risks as $r_{(1)} \le \cdots \le r_{(N)}$. At coverage k/N, retaining the first k gives average conditional risk $k^{-1}\sum_{i=1}^{k} r_{(i)}$. Any alternative subset containing a larger-risk unit while excluding a smaller-risk unit can be improved by exchanging the two. Repeating this exchange yields the k smallest-risk regions. □

C Dataset Details

C.1 Synthetic Systems

The synthetic study contains four process families chosen to vary the amount and form of information available around a missing region. Gaussian AR/VAR systems provide linear temporal and cross-variable dependence. Periodic and quasi-periodic systems test recurrence across time. Regime-switching systems introduce changes in local dynamics. Nonlinear systems test recoverability when the conditional relationship is not linear.

The synthetic experiments use the known data-generating process to obtain an oracle estimate of conditional Bayes risk. The same missing-region definitions used for the real-data study are applied when constructing point and block recovery tasks. Generation parameters for each process family are reported with the synthetic experiment specification.

C.2 Real-World Datasets

Table 2 summarizes the real datasets after preprocessing. All sequences are loaded through BenchPOTS. Experimental corruption is introduced only after preprocessing.

Table 2: Real datasets used in the recoverability experiments. Split sizes and native missing rates should be reported from the processed arrays used in the final runs.

Dataset	Sampling	Features	Window	Native missing
ETTh1	1 hour	7	96	0%
ETTh2	1 hour	7	96	0%
ETTm1	15 min	7	96	0%
ETTm2	15 min	7	96	0%
Beijing Air Quality	1 hour	132	96	1.6%

For datasets with native missing values, the native mask is retained after loading. Such entries can contribute contextual information only when observed and are excluded from artificial evaluation targets when unavailable.

D Comparison Imputers and Model Settings

All comparison imputers are implemented through PyPOTS. Each model receives the same artificially corrupted input for a given test condition. Probabilistic methods are reduced to point reconstructions by averaging their generated samples.

Table 3: Configuration of the comparison imputers.

Model	Main configuration
Lerp	Linear interpolation with no learned parameters.
BRITS	Bidirectional recurrent imputation with hidden size 64.
SAITS	Two attention layers, model dimension 128, four heads, feed-forward dimension 128, dropout 0.1, and equal ORT and MIT loss weights.
CSDI	Four layers, four heads, 64 channels, 50 diffusion steps, conditional generation, and mixed target strategy. Ten samples are generated per imputation.
GP-VAE	Two-layer encoder and decoder with 64 hidden units, Cauchy kernel, $\beta = 0.2$, and latent dimension determined from the number of features with an upper bound of 32. Ten samples are generated per imputation.
TimesNet	Two layers, model dimension 64, feed-forward dimension 128, top-$k = 5$, six kernels, and dropout 0.1.
ImputeFormer	Two layers, embedding dimensions of 32, projection dimension 32, four temporal heads, feed-forward dimension 128, and dropout 0.1.

All trainable comparison models use at most 100 epochs with early stopping patience of 10. Batch size is adjusted for datasets with large feature dimensions to control memory use.

E Training, Missingness, and Calibration Details

E.1 Missingness Construction

Artificial evaluation targets are restricted to entries that were originally observed. Under MAR, anchor variables are excluded from artificial target masks and remain available to determine masking probabilities. Consequently, the artificial MCAR and MAR mechanisms satisfy the masking assumption in Eq. 2.

Point MCAR test masks are generated with PyGrinder RDO and are restricted to originally observed values. Training MCAR masks use exact random holdout over observed entries. Point MAR masks are generated with an interpretable logistic mechanism in which high-observation variables act as anchor variables. The anchor fraction is set to 0.20, and state dependence is controlled by $\beta = 1.5$.

For MAR blocks, the target variable is selected from non-anchor variables. Block onset is sampled according to the observed state of its anchor variable during the preceding four time steps. Values inside the future block are not used to determine the onset probability.

Point rates are $0.10, \quad 0.30, \quad 0.50, \quad 0.70$, and block fractions are $\frac{1}{12}, \quad \frac{1}{6}, \quad \frac{1}{4}, \quad \frac{1}{2}$.

E.2 CURE Training

CURE uses a two-layer Transformer encoder with model dimension 64, four attention heads, and feed-forward dimension 128. Dropout is set to 0.1. The mean and variance heads are linear projections from the encoder representation, with softplus applied to the variance output.

CURE is trained with AdamW using learning rate 10^{-3} and weight decay 10^{-4}. Training runs for at most 80 epochs with patience 10. Each training sequence contributes two recovery tasks. Point and block tasks are sampled with equal probability.

E.3 Residual-Risk Calibration

The original validation split is divided into a model-validation portion and a calibration portion. For sliding-window datasets, a gap of up to one sequence length is inserted between them to reduce raw-timestamp overlap.

Calibration point masks use rates $0.30, \quad 0.50$, while calibration block fractions use $\frac{1}{6}, \quad \frac{1}{4}$.

Two independent calibration-mask seeds are used. The excess-risk estimator is a multilayer perceptron with hidden sizes 64 and 32. It uses GELU activations and dropout 0.1. Training uses AdamW with learning rate 10^{-3} for at most 100 epochs with patience 10.

Negative values of $\ell_G(\mathbf{X}_{G_{\text{eval}}}, \widehat{\mathbf{X}}_{G,f}) - \widehat{U}_G$ are retained during calibration. Predicted excess risk is clipped to zero only at inference. Because $\widehat{U}_G$ replaces the unavailable U_G^*, this target can contain both imputer-specific excess risk and CURE estimation error. Accordingly, the fitted output is interpreted as a residual-risk correction for estimating total conditional risk.

F Evaluation Details

Overall imputation accuracy is computed on all artificially hidden entries using MSE, RMSE, and MAE. Selective analyses use at most 250,000 units per test condition when subsampling is required. For point corruption, each hidden scalar forms one selection unit. Its realized error is

$$e_{G,f} = (\widehat{X}_{G,f} - X_G)^2.$$

For block corruption, one contiguous feature-specific block forms one unit. Its error is the mean squared error over all evaluable entries in that block.

Recoverability validity is measured with Spearman correlation between $\widehat{U}_G$ and realized error. Total-risk ranking is evaluated with the corresponding correlation between $\widehat{\mathcal{R}}_{G,f}$ and error. Cross-imputer difficulty agreement is measured through pairwise Spearman correlations between realized errors on shared units.

Risk–coverage curves contain 20 coverage levels from 0.05 to 1.00. At each level, units are sorted in ascending selection score and the retained error is averaged. AURC is calculated with numerical integration over the resulting curve. Risk is also recorded near coverage levels of 0.50, 0.70, and 0.90.

Four rankings are compared:

$$\widehat{U}_G, \qquad \widehat{\mathcal{R}}_{G,f}, \qquad \text{model disagreement}, \qquad \text{random ranking}.$$

Model disagreement is measured by the squared deviation between the imputer prediction and CURE's estimated conditional mean.

G Additional Synthetic Diagnostics

This section provides the detailed diagnostics supporting the synthetic results in Section 6. We examine scale normalization, fixed-gap behavior, the oracle risk decomposition, known generating-process parameters, and selective ranking against a simple gap-length baseline.

Table 4: **Synthetic robustness checks.** ρ_{norm} reports the range of Spearman correlations between scale-normalized oracle recoverability and normalized imputation error across BRITS, SAITS, and CSDI. $\rho_{L=24}$ evaluates learned recoverability after fixing gap length.

Family	CURE vs. U_G^*	ρ_{norm} range	Mean $\rho_{L=24}$
Gaussian VAR	0.926	0.726–0.886	0.900
Periodic	0.928	0.925–0.978	0.772
Regime switching	0.793	0.842–0.939	0.672
Nonlinear	0.801	0.746–0.934	0.766

The normalized analysis shows that the association between oracle recoverability and practical-imputer error is not explained solely by differences in process scale. Fixing $L = 24$ likewise removes gap length as an explanation for the learned ranking, although for Gaussian VAR and periodic systems the remaining variation is primarily across generating configurations rather than value-dependent variation within a fixed configuration.

Table 5: **Oracle risk-decomposition check.** Spearman correlation between the exact conditional risk $U_G^* + E_{G,f}^*$ and realized imputation MSE at the scenario level. The main-text range of 0.927–0.996 refers to BRITS, SAITS, and CSDI; linear interpolation is included here for completeness.

Family	Linear	BRITS	SAITS	CSDI
Gaussian VAR	0.989	0.961	0.962	0.967
Periodic	0.999	0.971	0.966	0.991
Regime switching	0.992	0.996	0.996	0.982
Nonlinear	0.983	0.960	0.927	0.960

The close agreement between exact conditional risk and realized MSE is a sanity check of the risk decomposition under oracle quantities. It should not be interpreted as validation of the learned residual-risk estimator used on real data.

Table 6: **Generating-process associations with oracle unrecoverability.** Spearman correlations are computed from one mean U_G^* per generating configuration. Each family contains six configurations. The small number of configurations makes these results directional diagnostics rather than high-powered significance tests.

Family	Generating property	ρ	p
Gaussian VAR	Process noise	0.926	0.008
Gaussian VAR	Cross-variable coupling	-0.802	0.055
Periodic	Process noise	0.956	0.003
Regime switching	Regime reveal strength	-0.837	0.038
Nonlinear	Process noise	0.845	0.034
Nonlinear	Cross-variable coupling	-0.926	0.008

Process noise is positively associated with unrecoverability in the Gaussian VAR, periodic, and nonlinear families. Stronger regime observability and nonlinear cross-variable coupling show the opposite association. Gaussian coupling follows the same direction, although its association is borderline at $p = 0.055$. These results provide a process-level sanity check for the interpretation of U_G^*.

Table 7: **Synthetic selective-imputation comparison.** Counts indicate the number of imputers, out of four, for which learned recoverability gives lower AURC than the stated baseline. Percentage columns report the median paired AURC reduction within each family.

Family	vs. Random	vs. Gap length	Median Δ Random	Median Δ Gap
Gaussian VAR	4/4	4/4	67.8%	69.2%
Periodic	4/4	3/4	64.1%	62.0%
Regime switching	4/4	3/4	53.4%	37.1%
Nonlinear	4/4	4/4	53.3%	52.6%
Overall	16/16	14/16	59.8%	54.3%

Across all 16 family–imputer combinations, learned recoverability gives lower AURC than random ranking. It also outperforms gap-length ranking in $14/16$ comparisons. By comparison, gap length itself beats random in $12/16$ cases but achieves only a 4.7% median AURC reduction relative to random, compared with 59.8% for learned recoverability. Thus, missingness geometry carries some difficulty information, but it accounts for only a small fraction of the selective gain obtained from learned recoverability in this synthetic study.