ProCTI: Prototype-Refined Global Conditioning for Diffusion-Based Time Series Imputation

Fariza Rashid[1]* **Duc Van Le**[2] **Rahat Masood**[2]
Gustavo Batista[2] **Aruna Seneviratne**[2] **Suranga Seneviratne**[1]
[1]University of Sydney [2]University of New South Wales

Abstract

Time series imputation has progressed from statistical and deep learning approaches to diffusion-based models, which have shown strong recent performance. Existing diffusion-based methods typically condition the reverse process using local contextual information from the current or neighbouring windows. Meanwhile, global dataset-level structure often remains implicit, limiting performance when local observations are sparse, noisy, or unrepresentative. To address this issue, we propose ProCTI, a diffusion-imputation framework that augments local conditioning with retrieved global dataset-level priors through learned prototypes. A hybrid conditioning mechanism integrates this global context with local signals during reverse diffusion, enabling more accurate reconstruction under varying missingness scenarios. Experiments across multiple benchmark datasets show that ProCTI outperforms strong baselines overall under random missingness, while remaining competitive under attribute-wise missingness. Furthermore, we use a latent-regime data model to characterise the precise conditions under which prototype-derived global conditioning provably improves imputation. We support this with a general theoretical analysis of local-global conditioning.

ProCTI Code Repository

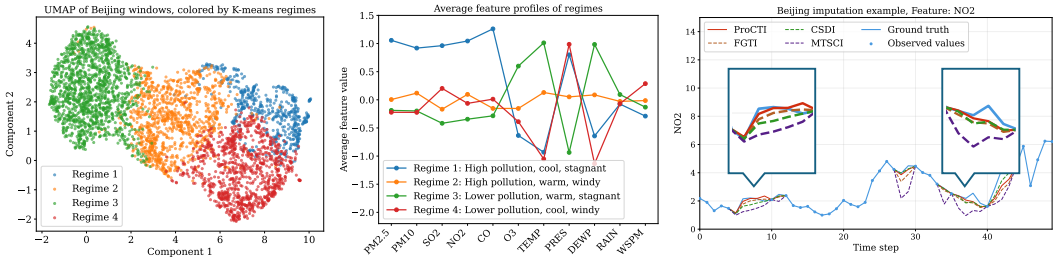
1 Introduction

Time series data appears in a wide range of real-world applications, including weather forecasting, finance, healthcare, cybersecurity, human-activity recognition, and audio signal processing [1–5]. Incomplete or missing data due to sensor malfunctions, human error, or communication failures is therefore a significant concern as it can degrade predictive performance and lead to harmful downstream decisions [6, 7]. Besides statistical, machine learning and deep learning methods [8–13], diffusion probabilistic models have emerged as a powerful paradigm for time series imputation [14, 15]. While existing diffusion-based approaches typically condition the reverse denoising process using signals derived from local context [15–17], broader dataset-level patterns are only implicitly encoded in learned parameters. Consequently, when local observations are sparse or noisy, the available context may be insufficient to accurately reflect the true underlying distribution, increasing uncertainty in the denoising process. Diffusion models conditioned only on local context may therefore struggle to generate accurate reconstructions of missing data.

In light of this limitation, we propose ProCTI, a diffusion-based time series imputation framework that explicitly incorporates dataset-level signals into the denoising process. Specifically, we aim to capture recurring global patterns of multivariate temporal behaviour that may exist across the

*Corresponding authors: fariza.rashid@sydney.edu.au, suranga.seneviratne@sydney.edu.au, g.batista@unsw.edu.au

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) (b) (c)

Figure 1: Regime structure and representative imputation behaviour in the Beijing dataset. (a) UMAP projection of normalised windows after PCA, coloured by K-means derived regimes. (b) Mean feature profiles of the discovered regimes, showing distinct combinations of pollution and meteorological conditions. (c) Example masked-window reconstruction for NO2, where ProCTI better matches the ground truth than competing baselines.

dataset. We refer to these patterns as regimes. To approximate such regimes, ProCTI introduces a learnable prototype bank within a prototype-conditioned diffusion model. Each partially observed input window queries the prototype bank to retrieve a window-specific global context vector, which refines the conditioning signal throughout the reverse diffusion process. This mechanism enables the model to repeatedly incorporate global structural priors during denoising, producing more accurate and coherent imputations.

As a pre-hoc motivating analysis, in Figure 1 we present a representative illustration of global regimes using the Beijing Air Quality dataset [18]. The left panel presents a UMAP visualisation of z-score normalised fixed-length windows, where each window is flattened and projected into a lower-dimensional space using PCA. Applying K-means clustering in this PCA-reduced space reveals four coarse groups of windows. The middle panel shows the corresponding average feature profiles, obtained by averaging each feature across all windows assigned to a given cluster. Because these profiles are derived from samples spanning the full dataset, they reveal recurring regimes characterised by different combinations of pollution indicators (PM2.5, PM10, SO2, NO2), temperature (TEMP), and wind speed (WSPM). Such dataset-level structure may be difficult to infer from a partially observed window alone. The right panel presents a masked-window imputation example, where leveraging this additional global context enables ProCTI to achieve lower MAE than competing baselines.

Overall, we make the following contributions:

- We introduce ProCTI, a new conditioning framework for diffusion-based time series imputation that explicitly combines retrieved global dataset-level priors with local context. We implement this framework through a prototype-conditioned diffusion model that improves imputation accuracy.
- We present a theoretical perspective that distinguishes local and global conditioning, using a latent-regime data model to characterise the precise conditions under which prototype-derived global conditioning provably improves imputation. This builds on standard supporting theoretical results that connect global conditioning to the diffusion training objective.
- We evaluate ProCTI against nine baselines on five real-world datasets: ProCTI achieves the best mean MAE on all five datasets under random missingness, reducing MAE by approximately 3–30% over the strongest diffusion baseline (FGTI), and is best or second-best in most attribute-wise missingness scenarios.

2 Related Work

Time series modelling Time-series modelling has seen rapid progress across generation, forecasting, and imputation tasks, with early approaches relying on statistical techniques and autoregressive methods [19, 20, 8]. More recent machine learning and deep learning methods emerged such as TIDER [10], which uses matrix factorisation to disentangle temporal components, and BRITS [11] and SAITS [21], which employ bidirectional Recurrent Neural Networks (RNNs) and self-attention

respectively, to capture long-range dependencies. In forecasting, non-Transformer approaches such as SCINet [22] and FreTS [23] model temporal interactions through recursive convolution and frequency-domain multilayer perceptrons respectively. Transformer-based forecasting methods [24–29] have also demonstrated strong performance by modeling long-range temporal dependencies through attention-based architectures, while time series generation methods like Diffusion-TS [14] and PaD-TS [30] are diffusion-based models which generate realistic synthetic sequences by learning temporal dynamics and preserving dataset-level statistical properties.

Conditioning in imputation models Existing diffusion-based imputation methods incorporate contextual information in varied ways, mostly relying on local or task-specific priors rather than explicitly learning dataset-level global structure. CSDI [15] conditions the reverse process directly on observed values and attention-based temporal/feature dependencies, without an explicit global prior. PriSTI [31] introduces context for spatiotemporal data by extracting coarse global priors from observed values together with geographic relationships, but this prior is specific to spatial sensor networks rather than general multivariate time series. MTSCI [16] conditions on neighbouring-window information and consistency constraints, thereby exploiting short-range temporal context instead of dataset-wide structure, while FGTI [17] incorporates frequency-domain priors through dominant- and high-frequency signals. Non-diffusion models such as TG-MSFM [32] use multi-scale flow trajectories to reconcile global trends with local details during deterministic reconstruction.

In contrast, we formalise conditioning for diffusion-based imputation by distinguishing *local context*, derived from the partially observed input window, and *global context*, representing recurring dataset-level structure shared across samples. We define their combination within the reverse diffusion process as complementary sources of guidance for reconstruction. Following this principle, ProCTI leverages a learned prototype bank to retrieve and inject global context alongside local observations into the denoising process for conditional imputation. This differs from PaD-TS [30], which preserves population-level statistics for unconditional time-series generation. We provide an overview of existing prototype-based methods in time-series modelling in Appendix A.1.

3 Prototype-Refined Global Conditioning

In this section, we present ProCTI, a diffusion-imputation framework that combines local window evidence with retrieved global dataset-level priors during reverse diffusion. As illustrated in Figure 2, ProCTI learns dataset-level prototypes that approximate recurring global regimes and uses them to produce instance-specific (i.e., window-level) conditioning signals. We first describe the prototype conditioning module, which combines global prototype guidance with local contextual information. We then formally distinguish local and global conditioning and use a latent-regime data model to characterise when combining them provably improves imputation. This builds on standard supporting results connecting global conditioning to the diffusion training objective.

3.1 Prototype-conditioning module

The prototype conditioning module contains a learnable prototype bank trained over the training dataset to approximate recurring global regimes. For each input window, cross-attention retrieves an attribute-wise global context vector from the prototype bank. This signal is then combined with local contextual information to form the conditioning input for the reverse diffusion process.

Problem setup. We consider a multivariate time series dataset X consisting of windows $x \in \mathbb{R}^{L\times K}$, where L represents the time length of the window and K represents the number of attributes. While the window may have naturally missing values, we apply an artificial keep-mask $m \in \{0,1\}^{L\times K}$, where $m_{t,k} = 1$ indicates that the value at time step t and attribute k is observed, and $m_{t,k} = 0$ otherwise. During training, the model learns to reconstruct values at masked positions through a diffusion denoising process.

Token construction. Given a mini-batch of B input windows $\mathbf{X} \in \mathbb{R}^{B\times L\times K}$ and the corresponding keep-masks $\mathbf{M} \in \{0,1\}^{B\times L\times K}$, where B denotes the batch size, we construct a 2-dimensional token at each time–attribute position:

$$\mathbf{z}_{b,t,k} = [\, x_{b,t,k},\ m_{b,t,k} \,] \in \mathbb{R}^2. \tag{1}$$

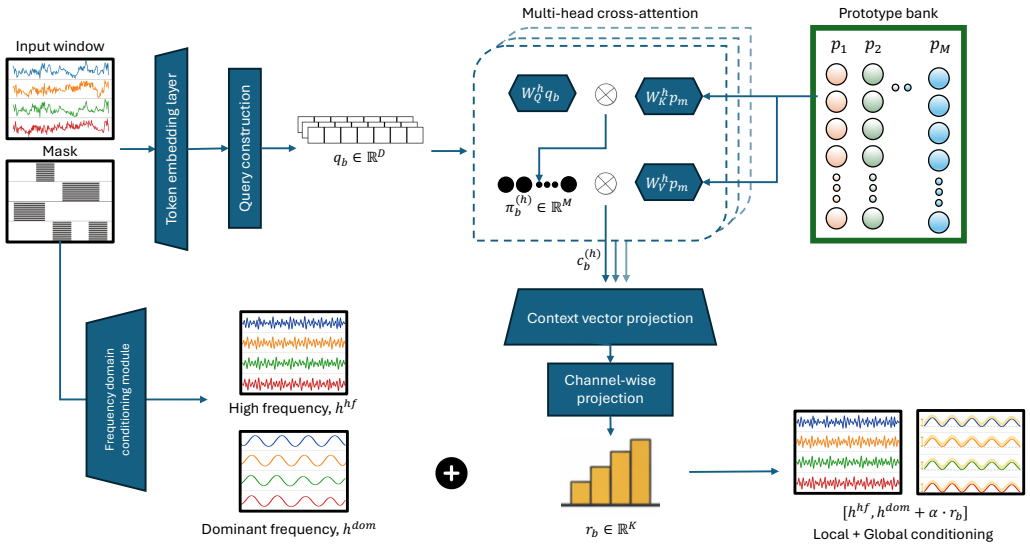


Figure 2: Prototype-conditioning module

Each token is projected into a D-dimensional embedding space via a learnable linear map:

$$\mathbf{e}_{b,t,k} = \mathbf{W}_{\text{proj}} \mathbf{z}_{b,t,k} \in \mathbb{R}^D. \tag{2}$$

The resulting tensor has shape $\mathbb{R}^{B \times L \times K \times D}$.

Query construction. To obtain a single representation per window, we reshape and average over all time–attribute embeddings to obtain a window-level query:

$$\mathbf{q}_b = \frac{1}{LK} \sum_{t=1}^{L} \sum_{k=1}^{K} \mathbf{e}_{b,t,k} \quad \in \mathbb{R}^D. \tag{3}$$

This query captures the instance-specific characteristics of the partially observed window, including both observed values and missingness patterns. This is then used to retrieve relevant global patterns from the prototype bank, which encodes dataset-level temporal regimes.

Prototype bank and cross-attention. We maintain a learnable prototype bank $P = \{p_m\}_{m=1}^M$ with $p_m \in \mathbb{R}^D$, intended to capture characteristic temporal regimes (periodic behaviour, irregular fluctuations, or cross-channel correlations) that may not be fully observable within any single partially observed window. Multi-head cross-attention from the window query q_b to the prototype bank yields a context vector $c_b \in \mathbb{R}^D$, which is mapped to the attribute domain by a learnable projection $W_r \in \mathbb{R}^{K \times D}$:

$$c_b = \text{MHCrossAttn}(q_b; P), \qquad r_b = W_r c_b \in \mathbb{R}^K. \tag{4}$$

The instance-dependent regime vector r_b thus translates globally learned temporal regimes into window-level, attribute-specific signals used to condition the diffusion process. Full multi-head cross-attention details are provided in Appendix A.2.

3.2 Global-conditioned diffusion model

We integrate the regime vector r_b from Section 3.1 into the diffusion-based imputation framework, refining the conditioning signal with instance-specific global information rather than relying on local observations or spectral features alone.

3.2.1 Conditional diffusion formulation

We follow the standard DDPM formulation [15]: the forward process gradually corrupts a clean input x_0 into x_T via a Markov chain with a predefined noise schedule, admitting the closed-form $x_t =$

$\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$ with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The reverse process $p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I})$ is parameterised by a denoising network $\epsilon_\theta(x_t, t)$ trained to predict the injected noise ϵ. For imputation, the reverse process is conditioned on a signal h derived from the partially observed window: $p_\theta(x_{0:T}^{ta} \mid h) = p(x_T^{ta}) \prod_t p_\theta(x_{t-1}^{ta} \mid x_t^{ta}, h)$, where x_0^{ta} denotes the missing (target) entries. Existing approaches typically construct h from local observations or window-level frequency representations; ProCTI refines this signal with prototype-derived global information.

3.2.2 Theoretical perspective on local and global conditioning

Let $X \in \mathbb{R}^{L \times K}$ denote a multivariate time series window, with observed entries O and missing entries Y; the imputation task is to model $p(Y \mid O)$. We formalise the distinction between local and global conditioning, then state four results characterising the value of hybrid conditioning during reverse diffusion.

Definition 1 (Local conditioning). Local conditioning refers to information derived directly from the observed values in the input window. Formally, we define a local conditioning representation as

$$h_{\text{loc}} = \phi_{\text{loc}}(O),$$

where ϕ_{loc} is a function of observed values and missingness patterns within the current window.

Definition 2 (Global conditioning). Global conditioning refers to information derived from dataset-level structure beyond the current input window. We define a latent regime variable Z as an unobserved random variable representing global temporal patterns governing the data distribution, such as periodic structure, long-range trends, or cross-channel dependencies. In ProCTI, this is approximated by a learned prototype bank, yielding a representation

$$h_{\text{glob}} = \phi_{\text{glob}}(O; P),$$

where P denotes the prototype bank learned over the training dataset.

Definition 3 (Local-global conditioning). We define hybrid conditioning as the combination of local and global components:

$$h = \psi(h_{\text{loc}}, h_{\text{glob}}).$$

To implement hybrid conditioning, we build upon a frequency-aware diffusion framework [17] that decomposes window-level frequency content into a high-frequency component $h^{hf} \in \mathbb{R}^{B \times L \times K}$ and a dominant-frequency component $h^{dom} \in \mathbb{R}^{B \times L \times K}$. The dominant-frequency component captures coarse temporal structure such as periodicity and long-range trends — properties associated with dataset-level regimes — while the high-frequency component encodes fine-grained instance-specific variation. We therefore inject the regime vector $r_b \in \mathbb{R}^K$ from the prototype module into the dominant-frequency branch only:

$$h_{b,t,k}^{dom+r} = h_{b,t,k}^{dom} + \alpha r_{b,k}, \qquad h = \left[h_{b,t}^{hf}; h_{b,t}^{dom+r}\right] \in \mathbb{R}^{2K}, \tag{5}$$

where α is a learnable scalar controlling the strength of global correction. The resulting prototype-refined conditioning h is fed to the denoising network $\epsilon_\theta(x_t, t \mid h)$ at each diffusion step.

Theoretical results. We next state a sequence of results: standard uncertainty and score-based arguments, followed by a latent-regime characterisation of precisely when hybrid conditioning improves imputation. Lemma 1 records that conditioning on a global variable Z, when informative, reduces residual uncertainty about Y. Proposition 1 decomposes the conditional score that the denoising network approximates into a local term and a regime guidance term, providing a constructive interpretation of how global conditioning intervenes in reverse diffusion. Proposition 2 connects this decomposition to the denoising mean-squared error that the diffusion training objective directly minimises. Proposition 3 specialises to a latent-regime data model and gives equivalent conditions on (Y, O, Z) under which the antecedents of the previous results hold, delineating both when ProCTI's prototype mechanism should be expected to help and when it should not.

Lemma 1 (Conditional uncertainty reduction). If Z carries information about Y beyond that already contained in O, namely $I(Y; Z \mid O) > 0$, then $H(Y \mid O, Z) < H(Y \mid O)$.

Proof. By the definition of conditional mutual information, $I(Y; Z \mid O) = H(Y \mid O) - H(Y \mid O, Z)$, so positivity of the left-hand side is equivalent to a strict decrease of conditional entropy. □

The lemma states the information-theoretic premise that global information, when relevant, reduces residual uncertainty in Y, but does not by itself describe how this reduction is realised inside the reverse diffusion process. The next two propositions make that connection explicit, first at the level of the score function approximated by the denoising network, then at the level of the mean-squared error directly minimised by training.

Proposition 1 (Score decomposition under hybrid conditioning). Under hybrid conditioning $h = (O, Z)$, the conditional score of the noisy state X_t admits the decomposition

$$\nabla_{x_t} \log p_t(x_t \mid O, Z) = \underbrace{\nabla_{x_t} \log p_t(x_t \mid O)}_{\text{local score}} + \underbrace{\nabla_{x_t} \log p(Z \mid x_t, O)}_{\text{regime guidance}}. \tag{6}$$

Proof. By Bayes' rule, $p_t(x_t \mid O, Z) = p(Z \mid x_t, O)\, p_t(x_t \mid O) / p(Z \mid O)$. Taking logarithms and differentiating in x_t, the term $\log p(Z \mid O)$ vanishes since it does not depend on x_t. □

The decomposition has the same structure as classifier guidance for conditional diffusion [33, 34] and assigns each conditioning source a distinct role: the local term is the score of the noised target given only O, while the regime guidance term pulls x_t toward states compatible with Z. ProCTI does not have direct access to Z at inference time; instead, its prototype-conditioning module learns an estimator $G = \phi_{\text{glob}}(O; P)$ of regime-relevant information and feeds G to the denoising network as a feasible surrogate for the regime guidance term in (6).

Proposition 2 (MMSE reduction under hybrid conditioning). Let $\mu^\star(x_t, h) = \mathbb{E}[X_0 \mid X_t = x_t, h]$ denote the Bayes-optimal denoiser under conditioning h, and let $h_1 = O$ and $h_2 = (O, Z)$. Then

$$\mathbb{E}\big[\|X_0 - \mu^\star(X_t, h_1)\|^2\big] - \mathbb{E}\big[\|X_0 - \mu^\star(X_t, h_2)\|^2\big] = \mathbb{E}\Big[\big\|\mu^\star(X_t, h_2) - \mu^\star(X_t, h_1)\big\|^2\Big] \geq 0, \tag{7}$$

with strict inequality whenever Z carries information about X_0 beyond that already in (X_t, O).

Proof. By the tower property, $\mu^\star(X_t, h_1) = \mathbb{E}[\mu^\star(X_t, h_2) \mid X_t, O]$. Decomposing $X_0 - \mu^\star(X_t, h_1) = (X_0 - \mu^\star(X_t, h_2)) + (\mu^\star(X_t, h_2) - \mu^\star(X_t, h_1))$, the cross-term has zero expectation by the orthogonality of conditional means. Squaring and taking expectations yields the identity, with strict inequality precisely when the two optimal denoisers differ on a set of positive measure. □

Proposition 2 is the operational counterpart of Lemma 1: it records the corresponding reduction in mean-squared error of the optimal denoiser, which is the quantity directly minimised by the diffusion training objective. Together, Propositions 1 and 2 give a unified picture: global conditioning enters the score additively as a regime guidance term, and this addition can only decrease the optimal denoising error, strictly so whenever the guidance term is informative.

Latent-regime data model. The previous three results all depend on the antecedent that Z carries information about Y (or X_0) beyond O. To pin down when this antecedent holds, we adopt a latent-regime model in which each window is associated with $Z \in \{1, \ldots, M^\star\}$ with $(O, Y) \mid Z \sim p(\cdot, \cdot \mid Z)$. The model is non-restrictive; any joint distribution over X admits such a representation and serves only to make the conditions for prototype-based improvement explicit.

Proposition 3 (Conditions for prototype informativeness). Under the latent-regime data model, the following are equivalent:

(i) $I(Y; Z \mid O) > 0$;
(ii) $Y \not\perp\!\!\!\perp Z \mid O$;
(iii) $\inf_f \mathbb{E}[\|Y - f(O)\|^2] > \inf_g \mathbb{E}[\|Y - g(O, Z)\|^2]$.

When any of these conditions hold, Lemma 1 yields a strict reduction in conditional entropy, and Proposition 2 yields a strict reduction in optimal denoising mean-squared error from hybrid over local conditioning. When they fail, that is when $Y \perp\!\!\!\perp Z \mid O$, no function of Z adds information about Y beyond O, and prototype-derived global conditioning cannot improve the optimal denoiser.

Proof. $(i) \Leftrightarrow (ii)$ is the standard equivalence between vanishing conditional mutual information and conditional independence. For $(ii) \Leftrightarrow (iii)$, the optimal predictor of Y given (O, Z) in mean-squared error is $\mathbb{E}[Y \mid O, Z]$, which coincides almost surely with $\mathbb{E}[Y \mid O]$ if and only if $Y \perp\!\!\!\perp Z \mid O$. □

Table 1: Comparison of imputation effectiveness across baseline methods, aggregated over missingness ratios $\{0.1, 0.3, 0.5, 0.7\}$. Each cell reports the mean with the standard deviation as a subscript. The best mean MAE and RMSE are shown in bold, and the second-best results are underlined. The full table of average results for each missingness ratio is in Appendix A.5.

Dataset	Metric	BRITS	CSDI	SCINet	TIDER	MTSCI	DiffusionTS	iTransformer	FGTI	PaD-TS	ProCTI
Gait	MAE	$0.352_{\pm.025}$	$0.449_{\pm.017}$	$0.475_{\pm.040}$	$0.746_{\pm.002}$	$0.501_{\pm.019}$	$0.367_{\pm.009}$	$0.537_{\pm.035}$	$\underline{0.332}_{\pm.026}$	$0.942_{\pm.032}$	$\mathbf{0.313}_{\pm.026}$
	RMSE	$\underline{0.533}_{\pm.037}$	$0.726_{\pm.014}$	$0.671_{\pm.054}$	$0.985_{\pm.003}$	$0.750_{\pm.019}$	$0.585_{\pm.014}$	$0.765_{\pm.039}$	$0.563_{\pm.035}$	$1.197_{\pm.032}$	$\mathbf{0.518}_{\pm.041}$
PhysioNet	MAE	$0.505_{\pm.032}$	$0.456_{\pm.034}$	$0.401_{\pm.091}$	$0.751_{\pm.030}$	$0.526_{\pm.075}$	$0.544_{\pm.026}$	$0.413_{\pm.098}$	$\underline{0.362}_{\pm.082}$	$0.865_{\pm.062}$	$\mathbf{0.259}_{\pm.058}$
	RMSE	$3.401_{\pm.158}$	$4.919_{\pm.203}$	$\underline{1.557}_{\pm.443}$	$3.465_{\pm.558}$	$1.609_{\pm.210}$	$3.054_{\pm.698}$	$1.925_{\pm.509}$	$2.970_{\pm.431}$	$3.153_{\pm.551}$	$\mathbf{1.383}_{\pm.299}$
Weather	MAE	$0.148_{\pm.046}$	$0.125_{\pm.015}$	$0.276_{\pm.085}$	$0.649_{\pm.006}$	$0.278_{\pm.115}$	$0.246_{\pm.037}$	$0.283_{\pm.116}$	$\underline{0.113}_{\pm.049}$	$0.176_{\pm.013}$	$\mathbf{0.095}_{\pm.037}$
	RMSE	$0.355_{\pm.040}$	$0.334_{\pm.020}$	$0.434_{\pm.094}$	$0.833_{\pm.008}$	$0.472_{\pm.119}$	$0.493_{\pm.056}$	$0.442_{\pm.111}$	$\underline{0.298}_{\pm.034}$	$0.362_{\pm.018}$	$\mathbf{0.274}_{\pm.040}$
Beijing	MAE	$0.217_{\pm.073}$	$0.186_{\pm.091}$	$0.341_{\pm.070}$	$0.722_{\pm.001}$	$0.282_{\pm.126}$	$0.256_{\pm.028}$	$0.405_{\pm.123}$	$\underline{0.159}_{\pm.053}$	$0.699_{\pm.073}$	$\mathbf{0.153}_{\pm.048}$
	RMSE	$0.482_{\pm.076}$	$0.464_{\pm.119}$	$0.559_{\pm.085}$	$1.003_{\pm.003}$	$0.577_{\pm.137}$	$0.582_{\pm.022}$	$0.660_{\pm.138}$	$\underline{0.437}_{\pm.070}$	$0.988_{\pm.079}$	$\mathbf{0.414}_{\pm.070}$
Stock	MAE	$3.880_{\pm.245}$	$2.842_{\pm.926}$	$2.671_{\pm.812}$	$5.278_{\pm.049}$	$3.762_{\pm.494}$	$4.302_{\pm.363}$	$\underline{2.622}_{\pm.971}$	$3.186_{\pm.422}$	$4.987_{\pm.150}$	$\mathbf{2.223}_{\pm.660}$
	RMSE	$4.299_{\pm.248}$	$3.223_{\pm.976}$	$3.482_{\pm.919}$	$5.712_{\pm.039}$	$4.120_{\pm.488}$	$4.801_{\pm.240}$	$\underline{3.226}_{\pm.963}$	$4.046_{\pm.230}$	$5.531_{\pm.089}$	$\mathbf{2.808}_{\pm.666}$

Proposition 3 delineates the regime in which ProCTI's prototype-conditioning mechanism is expected to provide gains. When Y is well-approximated by a deterministic function of O — as when an entire missing channel can be recovered by linear regression on the remaining channels — $Y \perp\!\!\!\perp Z \mid O$ holds approximately, and no global conditioning signal can improve on a sufficiently expressive local denoiser. We return to this prediction in Section 4, where it explains the dataset-dependence of ProCTI's gains observed under attribute-wise missingness. In the complementary regime, where Z carries residual information about Y given O, the prototype-conditioning module aims to recover this information from observations alone: under the latent-regime model with bank size $M \geq M^\star$ and a sufficiently expressive attention map ($\pi_m(O) \approx p(Z = m \mid O)$), the retrieved regime vector r_b functions as a soft sufficient statistic for Z given O, and Propositions 1 and 2 quantify the corresponding score and MMSE improvements.

4 Experiments

This section describes the experiments used to evaluate ProCTI against state-of-the-art baseline methods in terms of imputation effectiveness and robustness under attribute-wise missing scenarios. All experiments were conducted on a machine with an Intel Core i7-13700K CPU, two NVIDIA GeForce RTX 4090 GPUs (24GB each), and 128GB RAM.

4.1 Experimental Setup

Datasets We used five real-world time series datasets for our experiments, comparable to other similar work [14, 16, 17, 35]. *PhysioNet* [3] comprises multivariate clinical time-series data collected from ICU patients, with each record comprising measurements of approximately 40 physiological variables. The *Beijing* [18] dataset includes air-pollutant data collected from 12 air-quality monitoring sites in Beijing. The *Weather* [1] dataset is from the weather station of the Max Planck Institute for Biogeochemistry, comprising 14 meteorological attributes recorded between January 1st 2009 and December 31st 2016. The *Stock* [2] dataset reflects Google stock price data from 2004 to 2019, with each record representing a new day. Finally, the *Gait* [4] dataset comprises gait data collected using IMU sensors from 744 users during walks on level ground. We provide further details about dataset preparation for our experiments in Appendix A.3.

Baselines We evaluate ProCTI against a diverse set of baselines. We consider imputation methods including BRITS [11], CSDI [15], MTSCI [16], FGTI [17] and TIDER [10]. We include two forecasting models, SCINet [22] and iTransformer [29], that can be adapted to imputation via masked-value reconstruction [35]. We also evaluate two diffusion-based generation models, Diffusion-TS [14] and PaD-TS [30], that can likewise be adapted for conditional imputation.

Evaluation metrics For all of our experiments, we report the MAE and RMSE values to reflect imputation quality in terms of closeness to the ground truth. Therefore, the lower the value, the better the imputation. While MAE reflects the average error, RMSE reflects extreme deviations from the ground truth.

Table 2: Structured missingness evaluation results. Best results are in bold and second-best are underlined.

Dataset	Metric	Drop	BRITS	CSDI	SCINet	TIDER	MTSCI	DiffusionTS	iTransformer	FGTI	PaD-TS	ProCTI
Gait	MAE	1	0.735	0.801	0.745	0.745	0.741	1.053	0.745	**0.616**	1.037	0.619
		2	0.718	0.808	0.748	0.748	0.745	1.059	0.748	0.643	1.044	**0.640**
	RMSE	1	1.017	1.039	0.984	0.985	0.961	1.311	0.984	0.839	1.293	**0.821**
		2	0.971	1.047	0.988	0.988	0.967	1.316	0.988	0.877	1.300	**0.851**
PhysioNet	MAE	1	0.931	0.841	0.836	0.772	0.946	0.780	0.805	0.856	0.980	**0.747**
		2	0.962	0.694	0.708	0.768	0.740	0.693	0.692	0.645	0.944	**0.619**
	RMSE	1	1.246	1.010	1.031	0.938	1.147	1.002	0.998	1.042	1.215	**0.903**
		2	1.281	0.890	0.885	0.969	0.922	0.929	0.874	0.831	1.183	**0.781**
Weather	MAE	1	**0.205**	0.687	0.669	0.664	0.667	0.671	0.664	0.310	0.724	0.269
		2	**0.212**	0.693	0.679	0.674	0.671	0.696	0.673	0.318	0.755	0.288
	RMSE	1	**0.438**	0.859	0.840	0.838	0.830	0.874	0.838	0.593	0.932	0.479
		2	**0.480**	0.872	0.855	0.852	0.848	0.902	0.852	0.561	0.969	0.494
Beijing	MAE	1	1.107	0.708	0.737	0.745	0.647	0.875	0.736	0.429	0.863	**0.419**
		2	0.927	0.700	0.736	0.734	0.653	0.881	0.735	0.426	0.877	**0.436**
	RMSE	1	1.484	0.981	1.053	1.030	0.915	1.180	1.052	0.769	1.173	**0.716**
		2	1.309	0.990	1.053	1.094	0.937	1.191	1.053	0.776	1.189	**0.746**
Stock	MAE	1	**2.177**	4.975	5.350	5.073	4.882	4.655	5.092	5.422	4.776	4.580
		2	**2.355**	5.079	5.367	5.146	5.374	4.827	5.166	5.391	4.800	4.714
	RMSE	1	**2.528**	5.522	5.797	5.516	5.347	5.152	5.533	5.784	5.243	4.995
		2	**2.689**	5.581	5.778	5.550	5.793	5.282	5.566	5.787	5.308	5.082

4.2 Imputation effectiveness

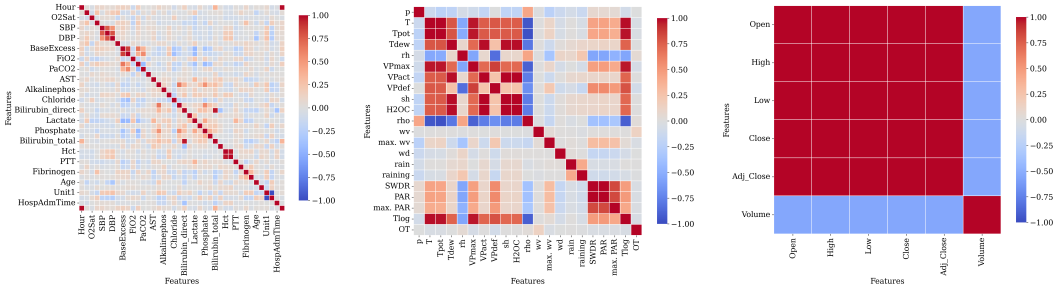
We compare imputation effectiveness under stochastic segment masking, artificially masking observed values to simulate missing data. We refer to the resulting proportion of unavailable values as the missing ratio r. Specifically, we implement the masking strategy of Zerveas et al. [36] as a two-state Markov process that generates alternating masked and unmasked segments whose lengths follow geometric distributions. The process is parameterised by a target missing ratio r and an average masked segment length l_m, with the corresponding average unmasked segment length given by $l_u = \frac{1-r}{r} l_m$. During training, we set $r = 0.15$, while evaluation is conducted at missing ratios of 0.10, 0.30, 0.50, and 0.70, where larger ratios correspond to increasingly challenging imputation settings. For both training and evaluation, we set $l_m = 6$, as moderately longer contiguous gaps provide a more challenging missingness pattern than short isolated gaps, while preserving sufficient observed context for reconstruction. Table 1 reports the mean and standard deviation of MAE and RMSE values over five random runs for each missingness ratio. We provide the full table of average results for each missingness ratio in Table 7 in Appendix A.5.

Overall, ProCTI outperforms all baselines in terms of average performance. This demonstrates the benefit of combining global prototype priors with local conditioning. Among the baselines, FGTI attains the second-best performance, yielding the second-lowest mean MAE for all scenarios and the second-lowest mean RMSE in 6 out of 10 scenarios. In comparison to FGTI, ProCTI reduces the average MAE by approximately 3–30%. The largest gains are observed on PhysioNet and Stock, followed by Weather, Beijing and Gait.

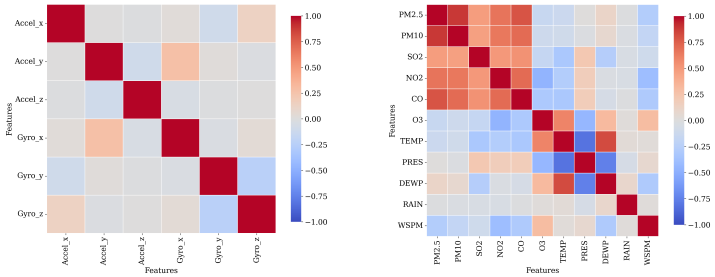
4.3 Robustness under attribute-wise missingness

We implemented attribute-wise missingness scenarios to evaluate robustness under structured missing patterns. This setting is motivated by real-world situations in which one or more attributes are missing for a period of time, for example, due to sensor malfunctions or system outages. Using the same training configuration as in the previous experiments, we conducted two attribute-wise missing protocols: Drop 1 and Drop 2, where one attribute and two attributes, respectively, were randomly selected and fully masked within the test windows. During evaluation, the reverse diffusion process reconstructed the missing attribute(s) solely from the remaining observed variables. For all datasets, we considered all available attributes except PhysioNet, which exhibited high sparsity. For this dataset, only attributes with more than 80% observed values were included in the experiment. Table 2 reports the MAE and RMSE values on the masked attributes for the different methods.

For the Gait, PhysioNet, and Beijing datasets, ProCTI outperforms all competing methods in terms of overall performance, achieving the lowest MAE and RMSE values in the majority of attribute-drop scenarios. FGTI consistently emerges as the second-best performing method. For the Weather



(a) PhysioNet (b) Weather (c) Stock



(d) Gait (e) Beijing

Figure 3: Feature correlation heatmaps for the five datasets: (a) PhysioNet, (b) Weather, (c) Stock, (d) Gait, and (e) Beijing.

and Stock datasets, however, BRITS displays the lowest MAE and RMSE values. Inspecting the feature-wise correlations of the five datasets, we find that BRITS's superior performance in these two cases can be attributed to its explicit feature-regression mechanism combined with bidirectional temporal recurrence. These are particularly effective when there is strong cross-feature redundancy. In the Stock dataset, for example, price variables such as Open, High, Low, Close, and Adj_Close are almost perfectly correlated, while many variables in the Weather dataset also exhibit linear dependencies and tightly coupled feature groups. Under attribute-drop evaluation, where an entire variable is missing, these relationships allow BRITS to reconstruct the missing channel directly from the remaining observed variables, consequently reducing the relative advantage of more complex diffusion-based or forecasting-adapted models. This pattern matches Proposition 3 in Section 3.2.2: the strong cross-feature redundancy in Stock and Weather approximately satisfies the conditional independence $Y \perp\!\!\!\perp Z \mid O$, the regime in which prototype-derived global conditioning provably cannot improve upon a sufficiently expressive local predictor. We provide heatmap visualisations of the datasets' feature-wise correlations in Figure 3.

In contrast, the Gait and PhysioNet datasets display substantially weaker feature correlations, indicating that missing-attribute recovery cannot rely primarily on simple cross-variable redundancy. For the Beijing dataset, although some strong pairwise correlations are present, its dependency structure is considerably more heterogeneous, containing mixtures of strongly positive, strongly negative, and weak relationships across pollutant, meteorological, and dispersion-related variables. Furthermore, as illustrated in Figure 1, Beijing windows form multiple recurring operating regimes with distinct feature profiles, suggesting that accurate missing-channel reconstruction depends on broader contextual interactions rather than uniform redundancy alone. Under such conditions, the antecedent of Proposition 3 holds: regimes carry residual information about Y beyond O, and ProCTI's prototype-derived global conditioning enables it to outperform BRITS and the remaining baselines, consistent with the entropy and mean-squared-error reductions established by Lemma 1 and Proposition 2.

Table 3: Ablation study of ProCTI across all datasets (mean values only). Variant A denotes the base FGTI model without prototype conditioning. Variants B–D use fixed $\alpha = 0.1$, while Variant E uses a learnable α.

Variant	Proto Bank	Proto Attn	Learnable α	Gait		PhysioNet		Weather		Beijing		Stock	
				MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
A	–	–	–	0.322	0.531	0.228	0.460	0.149	0.336	0.1566	0.4338	3.2606	4.0410
B	✓	–	–	0.318	0.521	0.219	0.424	0.113	0.296	0.1518	0.4062	2.6317	3.2982
C	✓	✓	–	0.317	0.520	0.217	0.417	0.119	0.304	0.1519	0.4044	2.4705	3.0971
D	frozen	✓	–	0.316	0.521	0.221	0.434	0.139	0.329	0.1591	0.4362	2.9831	3.7186
E (inj-dom)	✓	✓	✓	**0.316**	**0.518**	0.219	0.414	**0.096**	**0.280**	**0.1515**	**0.4043**	**2.2686**	**2.8323**
F (inj-hf)	✓	✓	✓	0.320	0.520	**0.214**	**0.412**	0.102	0.288	0.1910	0.4746	3.1198	3.7978
G (inj-both)	✓	✓	✓	0.316	0.526	0.223	0.422	0.104	0.289	0.1938	0.4752	2.8556	3.4772

4.4 Ablation analysis

We study the influence of the main components introduced in ProCTI, namely the prototype bank, the cross-attention interaction between input windows and prototypes, and the scaling parameter α used to inject regime information. We also study the effect of injecting regime information into both dominant and high-frequency branches together, as well as separately. To isolate the contribution of each component, we consider seven variants: (A) the original FGTI baseline without prototype conditioning, (B) adding only a learnable prototype bank, (C) enabling prototype cross-attention, (D) using a frozen prototype bank with attention, and (E - G) the full ProCTI model with a learnable α injected into the dominant frequency branch (inj-dom), high frequency branch (inj-hf), and both branches (inj-both) respectively. Full numerical results and discussion are reported in Table 3.

Across all five datasets, introducing a learnable prototype bank (Variant B) consistently improves upon the FGTI baseline (Variant A), highlighting the benefit of incorporating global prototype priors in addition to local frequency-based conditioning. Enabling prototype cross-attention (Variant C) yields further gains on most datasets, suggesting that dynamically matching each input window to relevant prototypes is more effective than relying on static prototype representations alone. In contrast, Variant D shows degraded performance relative to the preceding variants, indicating that a learnable prototype bank is more effective than a frozen one. Comparing the results across Variants E, F and G reflects that injecting a learnable α into the dominant frequency branch only yields lower error (with the exception of PhysioNet where Variant F returns a lower MAE and RMSE). Overall, Variant E yields the best results in the majority of cases, demonstrating the value of learning α (allowing the model to regulate how strongly regime information influences the local conditioning signal depending on the dataset) and injecting the regime information into the dominant frequency branch only.

5 Conclusion

In this paper, we introduced ProCTI, a diffusion-based imputation framework that combines local observations with reusable global priors through prototype retrieval. The proposed framework incorporates a learnable prototype-conditioning module that provides global context at the window-level, augmenting the conditioning signals of the reverse diffusion process. We further present a theoretical framework that distinguishes local and global conditioning in diffusion-based imputation models and demonstrate how ProCTI combines both, thereby improving imputation quality. Extensive experiments across diverse real-world datasets demonstrate that ProCTI consistently outperforms strong baselines under both random missingness and attribute-wise missingness settings. The gains of prototype-derived conditioning are conditional on regime structure being informative beyond observed entries (Proposition 3): when $Y \perp\!\!\!\perp Z \mid O$ holds approximately, as under strong cross-feature redundancy, simpler regression baselines remain competitive. Future work includes prototype-bank interpretability and designs for non-stationary regimes.

Acknowledgments and Disclosure of Funding

This research is funded by the Office of National Intelligence, Australia (Agreement No. GA396359), and the National Intelligence and Security Discovery Research Grant (NI240100159).

References

[1] Max Planck Institute for Biogeochemistry. Jena climate dataset. https://www.bgc-jena.mpg.de/wetter/, 2016. Accessed: 2026-04-16.

[2] Yahoo Finance. Google (goog) historical stock prices. https://finance.yahoo.com/quote/GOOG/history/?p=GOOG, 2026. Accessed: 2026-04-16.

[3] Matthew A Reyna, Christopher S Josef, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Shamim Nemati, Gari D Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. *Critical care medicine*, 48 (2):210–217, 2020.

[4] Thanh Trung Ngo, Yasushi Makihara, Hajime Nagahara, Yasuhiro Mukaigawa, and Yasushi Yagi. The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication. *Pattern Recognition*, 47(1):228–237, 2014.

[5] Ravin Gunawardena, Sandani Jayawardena, Suranga Seneviratne, Rahat Masood, and Salil S Kanhere. Single-sensor sparse adversarial perturbation attacks against behavioral biometrics. *IEEE Internet of Things Journal*, 11(16):27303–27321, 2024.

[6] Ahmed A Alwan, Allan J Brimicombe, Mihaela Anca Ciupala, Seyed Ali Ghorashi, Andres Baravalle, and Paolo Falcarin. Time-series clustering for sensor fault detection in large-scale cyber–physical systems. *Computer Networks*, 218:109384, 2022.

[7] Jinghan Du, Minghua Hu, and Weining Zhang. Missing data problem in the monitoring system: A review. *IEEE Sensors Journal*, 20(23):13984–13998, 2020.

[8] Mehran Amiri and Richard Jensen. Missing data imputation using fuzzy-rough methods. *Neurocomputing*, 205:152–164, 2016.

[9] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

[10] Shuai Liu, Xiucheng Li, Gao Cong, Yile Chen, and Yue Jiang. Multivariate time-series imputation with disentangled temporal representations. In *The Eleventh international conference on learning representations*, 2023.

[11] 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.

[12] 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*, pages 2260–2271, 2024.

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

[14] Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. *arXiv preprint arXiv:2403.01742*, 2024.

[15] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021.

[16] Jianping Zhou, Junhao Li, Guanjie Zheng, Xinbing Wang, and Chenghu Zhou. Mtsci: A conditional diffusion model for multivariate time series consistent imputation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3474–3483, 2024.

[17] Xinyu Yang, Yu Sun, Xiaojie Yuan, and Xinyang Chen. Frequency-aware generative models for multivariate time series imputation. *Advances in Neural Information Processing Systems*, 37: 52595–52623, 2024.

[18] Song Chen. Beijing Multi-Site Air Quality. UCI Machine Learning Repository, 2017. DOI: https://doi.org/10.24432/C5RK5G.

[19] David J Bartholomew. Time series analysis forecasting and control., 1971.

[20] Mark W Watson. Vector autoregressions and cointegration. *Handbook of econometrics*, 4: 2843–2915, 1994.

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

[22] Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. Scinet: Time series modeling and forecasting with sample convolution and interaction. *Advances in neural information processing systems*, 35:5816–5828, 2022.

[23] Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain mlps are more effective learners in time series forecasting. *Advances in Neural Information Processing Systems*, 36:76656–76679, 2023.

[24] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[25] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

[26] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*, 2021.

[27] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in neural information processing systems*, 35:9881–9893, 2022.

[28] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. arxiv 2022. *arXiv preprint arXiv:2211.14730*, 2022.

[29] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

[30] Yang Li, Han Meng, Zhenyu Bi, Ingolv T Urnes, and Haipeng Chen. Population aware diffusion for time series generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18520–18529, 2025.

[31] Mingzhe Liu, Han Huang, Hao Feng, Leilei Sun, Bowen Du, and Yanjie Fu. Pristi: A conditional diffusion framework for spatiotemporal imputation. In *2023 IEEE 39th international conference on data engineering (ICDE)*, pages 1927–1939. IEEE, 2023.

[32] Hangtian Wang and Mahito Sugiyama. Time-gated multi-scale flow matching for time-series imputation. In *The Fourteenth International Conference on Learning Representations*.

[33] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[34] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[35] Wenjie Du, Jun Wang, Linglong Qian, Yiyuan Yang, Zina Ibrahim, Fanxing Liu, Zepu Wang, Haoxin Liu, Zhiyuan Zhao, Yingjie Zhou, et al. Tsi-bench: Benchmarking time series imputation. *arXiv preprint arXiv:2406.12747*, 2024.

[36] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2114–2124, 2021.

[37] Bin Li, Carsten Jentsch, and Emmanuel Müller. Prototypes as explanation for time series anomaly detection. *arXiv preprint arXiv:2307.01601*, 2023.

[38] Zhi-Hao Yu, Lian-Tao Ma, Ya-Sha Wang, and Xu Chu. Imputation with inter-series information from prototypes for healthcare time series. *Journal of Computer Science and Technology*, 40(6): 1499–1511, 2025.

[39] Yu-Hao Huang, Chang Xu, Yueying Wu, Wu-Jun Li, and Jiang Bian. Timedp: Learning to generate multi-domain time series with domain prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17520–17527, 2025.

[40] Junwei Deng, Chang Xu, Jiaqi W Ma, Ming Jin, Chenghao Liu, and Jiang Bian. Oats: Online data augmentation for time series foundation models. *arXiv preprint arXiv:2601.19040*, 2026.

[41] Yuhang Duan, Lin Lin, and Xiaoshuai Wu. K-protodiff: Key prototypes-guided diffusion for time series generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 20959–20967, 2026.

[42] Kamile Stankeviciute, Ahmed M Alaa, and Mihaela Van der Schaar. Conformal time-series forecasting. *Advances in neural information processing systems*, 34:6216–6228, 2021.

A Appendix

A.1 Related work

Prototype-based representations in time series modelling Prototype-based representations have appeared across the broader time-series literature. In anomaly detection, ProtoAD [37] uses prototypes to characterise regular latent patterns and improve interpretability through example-based explanations. In healthcare imputation, PRIME [38] leverages prototype memory to capture inter-series information from similar patients within irregular clinical records. In generation, recent works such as TimeDP [39], OATS [40], and K-ProtoDiff [41] employ prototypes as domain prompts, guidance signals, or key-pattern constraints for synthesising realistic sequences. In contrast, ProCTI introduces prototypes for *diffusion-based conditional imputation* of general multivariate time series, where prototypes are learned as reusable dataset-level regimes and explicitly combined with local observations during reverse diffusion. This differs from prior prototype methods that focus on retrieval-style inter-series memory, interpretability, or unconditional generation rather than principled global conditioning for imputation.

A.2 ProCTI implementation details

A.2.1 Multi-head cross-attention details

We provide here the full multi-head cross-attention computation summarised in Section 3.1. Let the number of heads be H, with per-head dimension $d_h = D/H$. For each head $h \in \{1, \dots, H\}$ we have learnable projections $W_Q^{(h)}, W_K^{(h)}, W_V^{(h)} \in \mathbb{R}^{d_h \times D}$, and the projected query, key, and value vectors are

$$q_b^{(h)} = W_Q^{(h)} q_b, \quad k_m^{(h)} = W_K^{(h)} p_m, \quad v_m^{(h)} = W_V^{(h)} p_m,$$

with $q_b^{(h)}, k_m^{(h)}, v_m^{(h)} \in \mathbb{R}^{d_h}$. The attention weights are

$$\pi_{b,m}^{(h)} = \frac{\exp\left((q_b^{(h)})^\top k_m^{(h)} / \sqrt{d_h}\right)}{\sum_{j=1}^{M} \exp\left((q_b^{(h)})^\top k_j^{(h)} / \sqrt{d_h}\right)},$$

yielding the per-head context vector $c_b^{(h)} = \sum_{m=1}^{M} \pi_{b,m}^{(h)} v_m^{(h)} \in \mathbb{R}^{d_h}$. The outputs of all heads are concatenated and projected to obtain $c_b = W_O \operatorname{Concat}\left(c_b^{(1)}, \dots, c_b^{(H)}\right) \in \mathbb{R}^D$ with $W_O \in \mathbb{R}^{D \times D}$. The regime vector is then $r_b = W_r c_b \in \mathbb{R}^K$ as in Section 3.1.

A.2.2 Training and imputation algorithms

We provide the training and imputation algorithms for our proposed model ProCTI, in Algorithm 1 and Algorithm 2 respectively.

Algorithm 1 Training process of ProCTI

Input: Time series X, keep-mask M, prototype bank P, diffusion steps T
Output: Trained $\epsilon_\theta(\cdot)$ and prototype module

repeat
  $\hat{X}_0 \leftarrow X \odot (1 - M), \quad X^C \leftarrow X \odot M$
  $h^{hf}, h^{dom} \leftarrow \text{FreqFilter}(X^C)$
  $q_b \leftarrow \text{Query}(X^C, M)$
  $c_b \leftarrow \text{CrossAttn}(q_b, P), \quad r_b \leftarrow W_r c_b$
  $h_{b,t,k}^{dom+r} \leftarrow h_{b,t,k}^{dom} + \alpha r_{b,k}$
  $t \sim \text{Uniform}(\{1, \dots, T\}), \quad \epsilon \sim \mathcal{N}(0, I)$
  $\hat{X}_t \leftarrow \sqrt{\bar{\alpha}_t} \hat{X}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$
  $\hat{\epsilon} \leftarrow \epsilon_\theta(\hat{X}_t, t \mid X^C, h^{hf}, h^{dom+r})$
  Update parameters using $\mathcal{L}_{\text{diff}} = \|\epsilon - \hat{\epsilon}\|_2^2$
until converged

Algorithm 2 Imputation process of ProCTI

Input: Incomplete time series X, keep-mask M, trained prototype bank P, trained denoising network $\epsilon_\theta(\cdot)$, diffusion steps T

Output: Imputed time series $\tilde{X}$

$X^C \leftarrow X \odot M$
$h^{hf}, h^{dom} \leftarrow \text{FreqFilter}(X^C)$
$h^{dom+r} \leftarrow \text{ProtoCond}(X^C, M, P, h^{dom})$
$\hat{X}_T \sim \mathcal{N}(0, I)$
for $t = T, T-1, \ldots, 1$ **do**
    $\hat{\epsilon} \leftarrow \epsilon_\theta(\hat{X}_t, t \mid X^C, h^{hf}, h^{dom+r})$
    Obtain $\hat{X}_{t-1}$ from $p_\theta(\hat{X}_{t-1} \mid \hat{X}_t, X^C, h^{hf}, h^{dom+r})$
end for
$\tilde{X} \leftarrow X \odot M + \hat{X}_0 \odot (1 - M)$
return $\tilde{X}$

A.3 Dataset preparation

In this section, we provide a detailed description of the dataset preprocessing steps to create training, validation and test splits for our evaluation experiments. Since the Gait, PhysioNet, and Beijing datasets contain identifiable entities that can be used for partitioning (user IDs, patient IDs, and monitoring stations, respectively), we construct training, validation, and test splits such that each entity appears in only one split, preventing information leakage. This strategy differs from conventional approaches that randomly partition data at the window level, and is therefore designed to evaluate the ability of ProCTI and baseline methods to generalise across non-overlapping entities by capturing global patterns.

Specifically, for the Gait and PhysioNet datasets, we adopt user-disjoint and patient-wise splits, respectively, ensuring that all sequences associated with a given subject or patient are confined to a single partition. The training, validation, and test splits follow a 70:15:15 ratio at the entity level. For the Beijing dataset, we use a station-disjoint split, partitioning the 12 monitoring stations into training, validation, and test groups in an 8:2:2 ratio, and constructing windows independently within each station. For the Weather and Stock datasets, we adopt a chronological split along the temporal dimension, using a 70:15:15 ratio for training, validation, and testing. In all cases, non-overlapping fixed-length windowing is applied after splitting to generate sequences for model training and evaluation.

A.4 Supplementary experiments

A.4.1 Probabilistic evaluation

Table 4: Probabilistic evaluation on the test split across missingness ratios on PhysioNet data.

Metric	Missing	CSDI	Diffusion-TS	FGTI	PaD-TS	MTSCI	ProCTI
CRPS(↓)	0.10	0.1798	0.6579	0.1187	0.8727	0.4076	0.1286
	0.30	0.1867	0.8361	0.1362	0.8955	0.5980	0.1390
	0.50	0.2223	0.8485	0.1647	1.0370	0.6081	0.1593
	0.70	0.2284	0.7599	0.2266	0.9718	0.7307	0.2040
PI90 Coverage(90%)	0.10	0.8447	0.0368	0.8479	0.1383	0.8053	0.8467
	0.30	0.8460	0.0514	0.9097	0.2291	0.7806	0.8876
	0.50	0.8566	0.0734	0.9307	0.3375	0.7612	0.9017
	0.70	0.8666	0.1082	0.9265	0.4910	0.7368	0.8872
PI90 Width(↓)	0.10	0.5861	0.0647	0.5884	0.2509	1.2602	0.5704
	0.30	0.6368	0.0806	0.7693	0.4349	1.4165	0.7012
	0.50	0.7264	1.1117	1.0723	0.6777	1.5938	0.9137
	0.70	0.8134	0.1625	1.5036	1.0513	1.7813	1.1878

To complement the imputation evaluation reported in the main experiments (Section 4), we further assess the probabilistic quality of the predictive distributions produced by diffusion-based models. Using the PhysioNet and Beijing datasets under the Markov masking protocol, we adopt the Continu-

Table 5: Probabilistic evaluation on the test split across missingness ratios on Beijing data.

Metric	Missing	CSDI	Diffusion-TS	FGTI	MTSCI	PaD-TS	ProCTI
CRPS(↓)	0.10	0.1646	0.2368	0.1204	0.2286	0.6141	0.1185
	0.30	0.1700	0.2326	0.1282	0.2374	0.5920	0.1254
	0.50	0.1894	0.2393	0.1472	0.2639	0.5784	0.1424
	0.70	0.2119	0.2400	0.1679	0.2861	0.5657	0.1617
PI90 Coverage(90%)	0.10	0.8439	0.3016	0.8152	0.7506	0.1611	0.8536
	0.30	0.8479	0.3478	0.8279	0.7657	0.2946	0.8748
	0.50	0.8536	0.3950	0.8353	0.7825	0.4410	0.8925
	0.70	0.8679	0.4417	0.8301	0.8177	0.5870	0.9057
PI90 Width(↓)	0.10	0.8081	0.1771	0.5441	0.8911	0.2970	0.6532
	0.30	0.8539	0.2057	0.5857	1.0115	0.5531	0.7235
	0.50	0.9440	0.2409	0.6544	1.1976	0.8731	0.8427
	0.70	1.1174	0.2789	0.7461	1.4426	1.2630	1.0122

ous Ranked Probability Score (CRPS) [15, 17], and 90% prediction interval coverage (PI90 coverage) and width (PI90 width) [42] across different missingness ratios. CRPS evaluates the overall quality of the predictive distribution in comparison to the ground truth, where lower values are preferred. PI90 coverage measures how often the ground-truth values fall within the predicted 90% interval. Values closer to 0.90 are preferred, meaning that well-calibrated predictive intervals are neither too narrow nor excessively wide. Finally, PI90 width reflects interval sharpness, where narrower intervals are preferred, provided calibration remains close to the nominal 90% level. Table 4 and Table 5 show the results for PhysioNet and Beijing, respectively.

Overall ProCTI consistently achieves strong probabilistic performance across all missingness ratios, obtaining the best or second-best CRPS values while maintaining PI90 Coverage close to 90%. In contrast, some baselines such as Diffusion-TS and PaD-TS, produce narrow intervals but reveal severe under-coverage, indicating overconfident uncertainty estimates. Although FGTI achieves the lowest CRPS in some settings under PhysioNet, it generally requires wider prediction intervals to do so. Overall, ProCTI provides the most balanced trade-off between predictive accuracy, calibration, and sharpness, demonstrating that incorporating global prototype priors improves imputation performance as well as probabilistic reliability.

A.4.2 Resource consumption

We evaluate the resource consumption of ProCTI against all baselines in Figure 4. ProCTI achieves runtime comparable to FGTI and lower than other diffusion-based methods such as CSDI and MTSCI. Although its GPU memory usage is higher than most baselines, it remains similar to FGTI despite the added model complexity due to the prototype conditioning module. Given its overall superior imputation performance, we argue that this additional resource cost is justified.

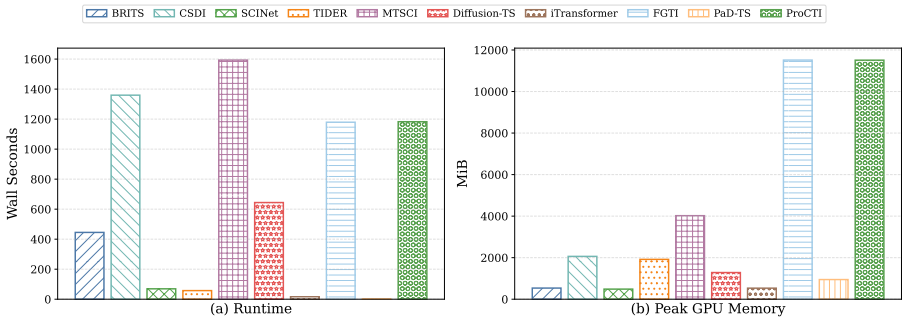


Figure 4: Runtime and peak GPU memory usage for different models on the Beijing dataset under the 30% missingness test setting.

A.4.3 Hyperparameter evaluation

In this section, we present a hyperparameter sensitivity analysis investigating the effect of prototype bank size and the regime-vector scaling parameter α. Table 6 and Figure 5 report ProCTI's imputation performance on held-out validation data, averaged across missingness ratios (10%, 30%, 50%, and 70%), for different values of these hyperparameters. Table 6 shows that a prototype bank size of 32 yields lower error in the majority of cases across MAE and RMSE (6/10 cells). Figure 5 indicates that $\alpha = 0.1$ wins on Beijing, Gait, and Stock, while $\alpha = 0.3$ gives lower MAE on PhysioNet and Weather. We therefore use $\alpha = 0.1$ as it gives the best overall performance, and a prototype bank size of 32 in ProCTI.

Table 6: Effect of prototype bank size (number of prototypes) on imputation performance. Best results for each dataset and metric are highlighted in bold.

Metric	Prototype bank size	Beijing	Gait	PhysioNet	Stock	Weather
MAE	16	0.200	0.310	0.222	0.894	0.102
	32	**0.158**	0.310	0.224	**0.835**	0.118
	64	0.202	**0.306**	**0.216**	1.030	**0.095**
RMSE	16	0.481	0.519	0.416	1.187	0.450
	32	**0.412**	**0.510**	0.438	**1.118**	**0.387**
	64	0.482	0.512	**0.413**	1.340	0.434

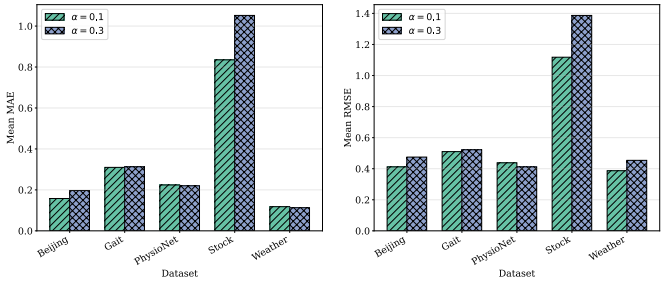


Figure 5: Effect of scaling parameter $\alpha \in \{0.1, 0.3\}$ on imputation performance. Lower values indicate better imputation performance.

A.5 Supplementary results

A.5.1 Imputation effectiveness

In Table 7, we report the detailed imputation performance of all methods under the Markov masking protocol, using MAE and RMSE at each missingness ratio, averaged over five repeated runs. Overall, ProCTI achieves the lowest errors in the majority of evaluated scenarios, while FGTI consistently delivers the second-best performance in most cases.

A.5.2 Statistical significance of imputation effectiveness

As the imputation performance of ProCTI and FGTI were very similar on some datasets such as Gait and Beijing, we statistically analysed the significance of the gains yielded by ProCTI. We conducted paired t-tests and Wilcoxon signed-rank tests across the 5 seeds for each missing ratio, with the one-sided hypothesis that ProCTI's per-seed error is lower than FGTI's (i.e., FGTI - ProCTI > 0). The paired t-test assumes the per-seed differences are approximately normally distributed, while the Wilcoxon signed-rank test is a non-parametric alternative that does not require this assumption and is more robust with small sample sizes. Results are shown in Table 8 (significance at the 0.05 shown in bold).

On Beijing, ProCTI significantly outperforms FGTI ($p < 0.05$) in 4/8 cells (t-test) and 1/8 (Wilcoxon). On Gait, ProCTI is significantly better in 5/8 cells under both tests, concentrated at $r \in 0.1, 0.3, 0.5, 0.7$; significance is lost at r = 0.7, consistent with the smaller gains in Table 1. These tests are based on 5 seeds which is a small sample for reliably estimating normality or the

Table 7: Comparison of imputation effectiveness across baseline methods. Best results are shown in bold and second-best results are underlined.

Dataset	Metric	Missing	BRITS	CSDI	SCINet	TIDER	MTSCI	DiffusionTS	iTransformer	FGTI	PaD-TS	ProCTI
Gait	MAE	0.1	0.324	0.436	0.436	0.749	0.484	0.358	0.501	0.307	0.906	**0.288**
		0.3	0.342	0.438	0.454	0.744	0.488	0.361	0.521	0.317	0.928	**0.299**
		0.5	0.360	0.449	0.482	0.745	0.505	0.369	0.544	0.336	0.954	**0.317**
		0.7	0.383	0.473	0.528	0.745	0.527	0.380	0.583	0.366	0.981	**0.347**
	RMSE	0.1	0.491	0.717	0.615	0.989	0.734	0.573	0.722	0.530	1.162	**0.477**
		0.3	0.518	0.715	0.645	0.982	0.737	0.575	0.749	0.543	1.184	**0.497**
		0.5	0.545	0.726	0.682	0.984	0.755	0.588	0.774	0.570	1.208	**0.528**
		0.7	0.577	0.746	0.741	0.984	0.775	0.604	0.814	0.610	1.235	**0.572**
PhysioNet	MAE	0.1	0.486	0.426	0.327	0.740	0.456	0.514	0.324	0.273	0.794	**0.203**
		0.3	0.481	0.427	0.339	0.796	0.473	0.574	0.353	0.322	0.837	**0.223**
		0.5	0.504	0.480	0.416	0.730	0.556	0.555	0.432	0.397	0.897	**0.278**
		0.7	0.551	0.489	0.524	0.737	0.617	0.534	0.543	0.459	0.934	**0.332**
	RMSE	0.1	3.245	4.758	1.355	2.996	1.868	2.023	1.484	2.446	2.404	**1.182**
		0.3	3.372	4.789	1.083	4.274	1.355	3.538	1.540	2.841	3.076	**1.074**
		0.5	3.366	5.203	1.679	3.264	**1.623**	3.413	2.120	3.459	3.609	1.657
		0.7	3.622	4.928	2.111	3.324	**1.590**	3.242	2.554	3.135	3.524	1.621
Weather	MAE	0.1	0.105	0.109	0.199	0.641	0.160	0.297	0.173	0.073	0.163	**0.062**
		0.3	0.124	0.120	0.226	0.651	0.221	0.249	0.215	0.085	0.169	**0.076**
		0.5	0.155	0.125	0.289	0.653	0.307	0.227	0.309	0.111	0.177	**0.097**
		0.7	0.209	**0.144**	0.389	0.651	0.425	0.213	0.435	0.182	0.193	0.146
	RMSE	0.1	0.315	0.309	0.343	0.821	0.338	0.570	0.346	0.295	0.338	**0.230**
		0.3	0.337	0.333	0.382	0.838	0.424	0.498	0.369	0.268	0.362	**0.266**
		0.5	0.357	0.335	0.452	0.837	0.508	0.465	0.462	0.281	0.368	**0.276**
		0.7	0.410	0.359	0.558	0.836	0.616	0.441	0.591	0.346	0.379	**0.326**
Beijing	MAE	0.1	0.141	**0.096**	0.332	0.721	0.152	0.225	0.285	0.105	0.619	0.103
		0.3	0.177	0.130	0.267	0.722	0.212	0.243	0.334	0.130	0.667	**0.128**
		0.5	0.243	0.217	0.331	0.722	0.331	0.268	0.439	0.176	0.720	**0.169**
		0.7	0.306	0.299	0.435	0.722	0.435	0.288	0.562	0.225	0.790	**0.214**
	RMSE	0.1	0.397	0.340	0.512	0.999	0.419	0.554	0.515	0.363	0.899	**0.337**
		0.3	0.449	0.400	0.485	1.007	0.519	0.578	0.590	0.406	0.959	**0.381**
		0.5	0.508	0.508	0.562	1.004	0.636	0.589	0.701	0.452	1.011	**0.438**
		0.7	0.574	0.609	0.677	1.003	0.733	0.606	0.832	0.526	1.084	**0.499**
Stock	MAE	0.1	3.582	1.875	1.729	5.278	3.144	3.810	1.520	2.678	4.773	**1.348**
		0.3	3.801	2.355	2.382	5.212	3.626	4.684	2.228	3.039	4.992	**2.121**
		0.5	3.983	3.151	2.931	5.293	3.988	4.340	2.960	3.370	5.073	**2.546**
		0.7	4.153	3.986	3.640	5.328	4.292	4.373	3.781	3.658	5.108	**2.876**
	RMSE	0.1	4.010	2.198	2.387	5.712	3.532	4.492	2.099	3.736	5.413	**1.913**
		0.3	4.203	**2.715**	3.174	5.660	3.946	5.076	2.861	4.018	5.517	2.720
		0.5	4.399	3.557	3.826	5.728	4.344	4.795	3.607	4.171	5.572	**3.162**
		0.7	4.584	4.423	4.540	5.750	4.658	4.843	4.338	4.260	5.620	**3.438**

sign/rank distribution. This likely explains why some cells miss significance despite ProCTI's lower mean error in Table 1. We expect significance to hold more consistently with additional seeds.

Table 8: Paired one-sided significance tests for ProCTI vs. FGTI across 5 seeds (H_1: ProCTI error $<$ FGTI error). p-values below 0.05 are in bold. With $n = 5$, the minimum attainable Wilcoxon p-value is $1/32 \approx 0.031$.

Dataset	r	Metric	t-stat	t-test p	Wilcoxon stat	Wilcoxon p
Beijing	0.1	MAE	-0.306	0.388	5	0.313
		RMSE	-2.590	**0.030**	1	0.063
	0.3	MAE	-3.111	**0.018**	0	**0.031**
		RMSE	-2.156	**0.049**	2	0.094
	0.5	MAE	-1.950	0.062	1	0.063
		RMSE	-2.677	**0.028**	1	0.063
	0.7	MAE	-1.049	0.177	6	0.406
		RMSE	-0.422	0.347	8	0.594
Gait	0.1	MAE	-3.906	**0.009**	0	**0.031**
		RMSE	-4.500	**0.005**	0	**0.031**
	0.3	MAE	-7.160	**0.001**	0	**0.031**
		RMSE	-1.814	0.072	3	0.156
	0.5	MAE	-3.780	**0.010**	0	**0.031**
		RMSE	-3.177	**0.017**	0	**0.031**
	0.7	MAE	-0.774	0.241	4	0.219
		RMSE	-1.396	0.118	2	0.094

A.5.3 Prototype utilisation

We analysed how many of the 32 prototypes were actually active at test time by computing the attention weight assigned to each of the 32 prototypes across the entire test sets' windows. We report two summary statistics: (1) the number of prototypes individually receiving more than half of a uniform share of attention (i.e., average weight above 1/(2×32)); and (2) the number of prototypes needed to jointly account for 90% of total attention mass. We present the results for missingness ratio 0.10 (seed 1) across all five datasets in Table 9 below.

We find that prototype utilisation is broad rather than concentrated on a small interpretable subset for each dataset. At least 31 of the 32 prototypes individually receive more than a half-share of attention, and between 27 and 29 of the 32 prototypes are needed to jointly account for 90% of attention mass. This suggests the prototype bank is not over-provisioned at bank size 32. Stock and Gait are marginally more selective under the 90%-cumulative measure (27 and 28 active prototypes respectively) than Weather, PhysioNet, and Beijing (29 each).

Table 9: Number of active prototypes per dataset under the usage threshold and the 90% cumulative usage criterion.

Dataset	n_{active} (threshold)	n_{active} (90% cumulative)
Stock	31	27
Weather	32	29
PhysioNet	32	29
Beijing	32	29
Gait	32	28

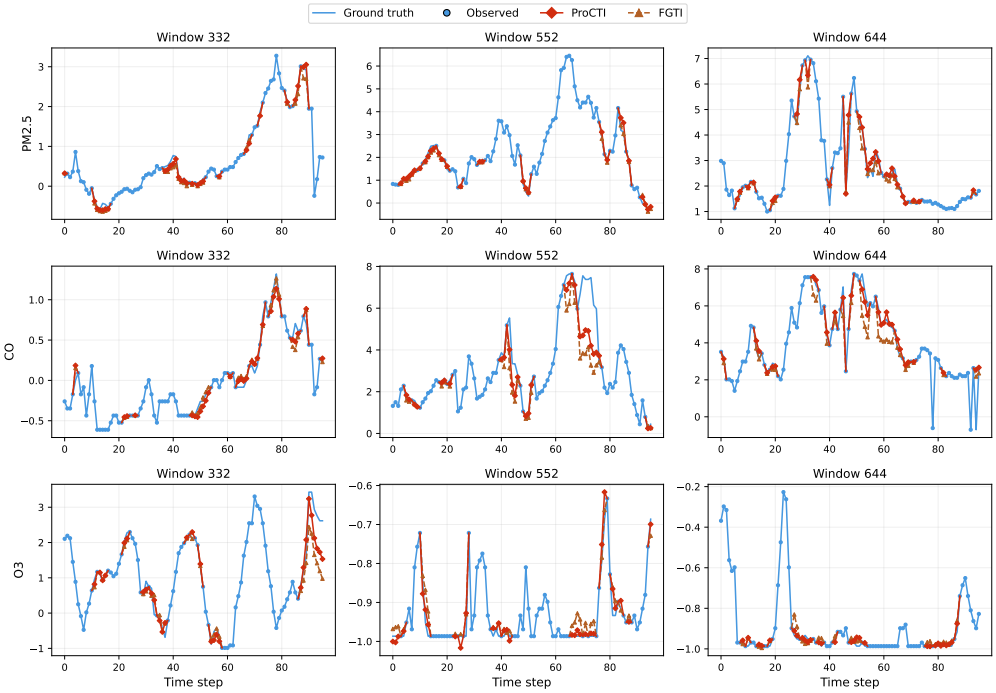
A.6 Visualisations

In Figure 6, Figure 7, and Figure 8, we present representative imputation examples for the Beijing, Gait, and PhysioNet datasets, comparing ProCTI, FGTI, and CSDI. Across all three datasets, CSDI imputations are more prone to deviating from the ground-truth trajectories, whereas ProCTI follows the true values within the masked regions more closely. ProCTI and FGTI often produce visually similar reconstructions, reflecting their shared frequency-aware diffusion conditioning framework. Nonetheless, closer inspection of several masked segments, together with the results in Table 1 and Table 7, shows that ProCTI consistently attains lower overall errors than FGTI, demonstrating the benefit of incorporating global conditioning signals.

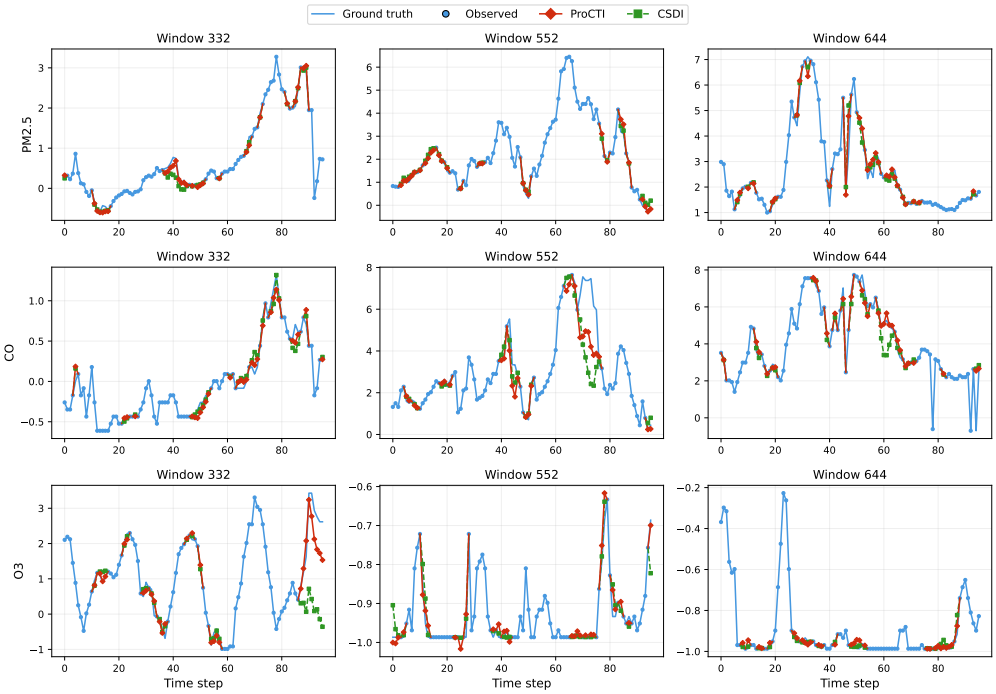
Finally, Figure 9 illustrates representative examples of prototype usage on window samples from the Stock dataset. Specifically, the figure shows the raw time series, the attention weights assigned by each window to the prototypes, the corresponding regime vector r_b values, and the resulting adjustments to the dominant-frequency component of the three attributes with the largest absolute r_b values.

A.7 Broader impact statement

The positive societal impacts of this work are directly related to time-series-data applications where missing or incomplete data are common, and improved imputation can positively affect downstream tasks. Example domains include healthcare, transportation, environmental monitoring, and industrial systems, where sensor readings or records may be missing or faulty. In healthcare, for instance, better recovery of missing clinical measurements may assist patient monitoring or diagnosis. On the other hand, potential unintended consequences in downstream applications due to incorrect imputations should also be considered. In such settings, technologies such as ProCTI should be used as decision-support tools rather than as replacements for human judgment, particularly in high-stakes scenarios.

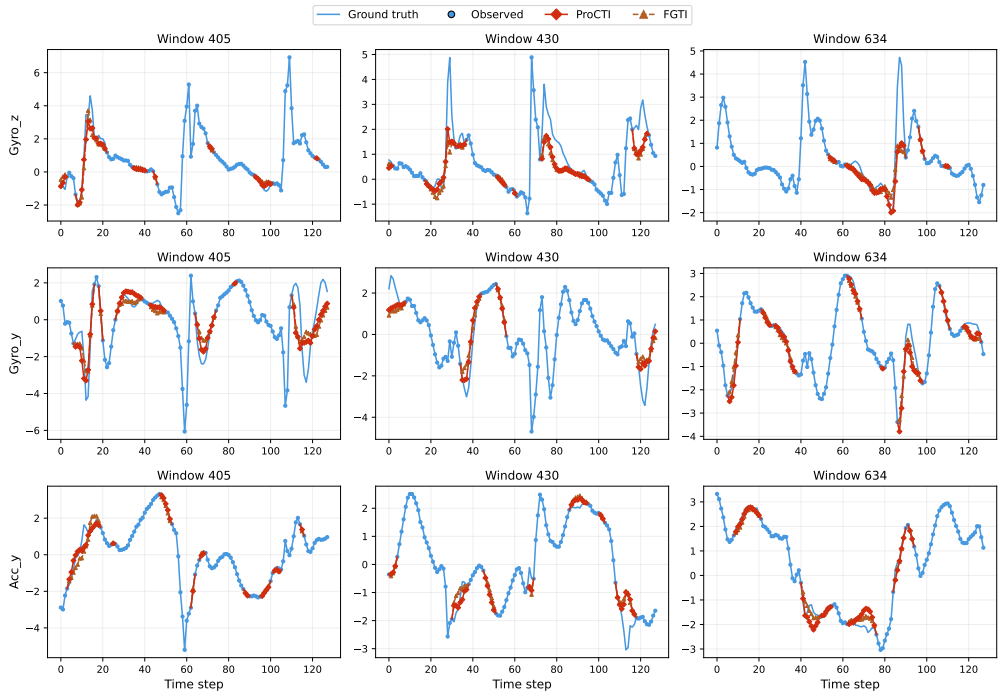


(a) ProCTI vs FGTI

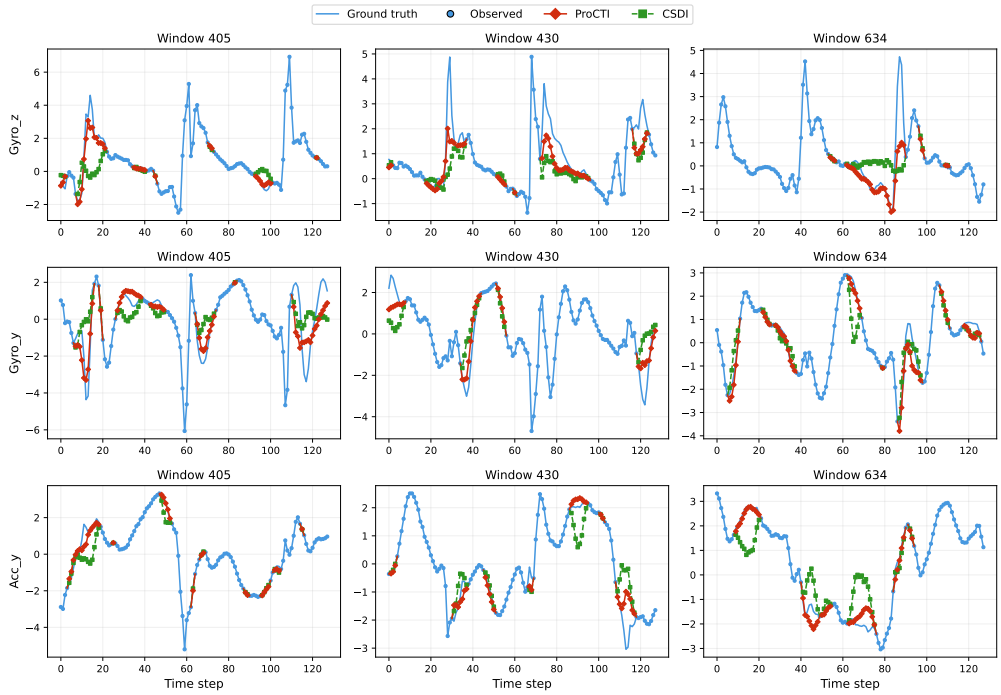


(b) ProCTI vs CSDI

Figure 6: Representative imputation comparison on the Beijing dataset.

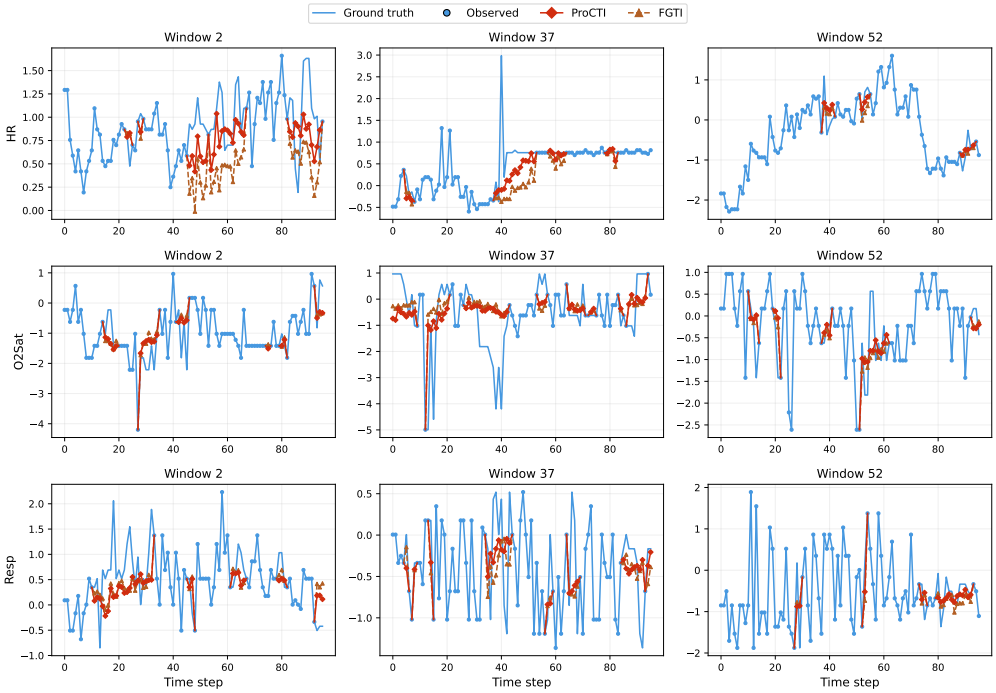


(a) ProCTI vs FGTI

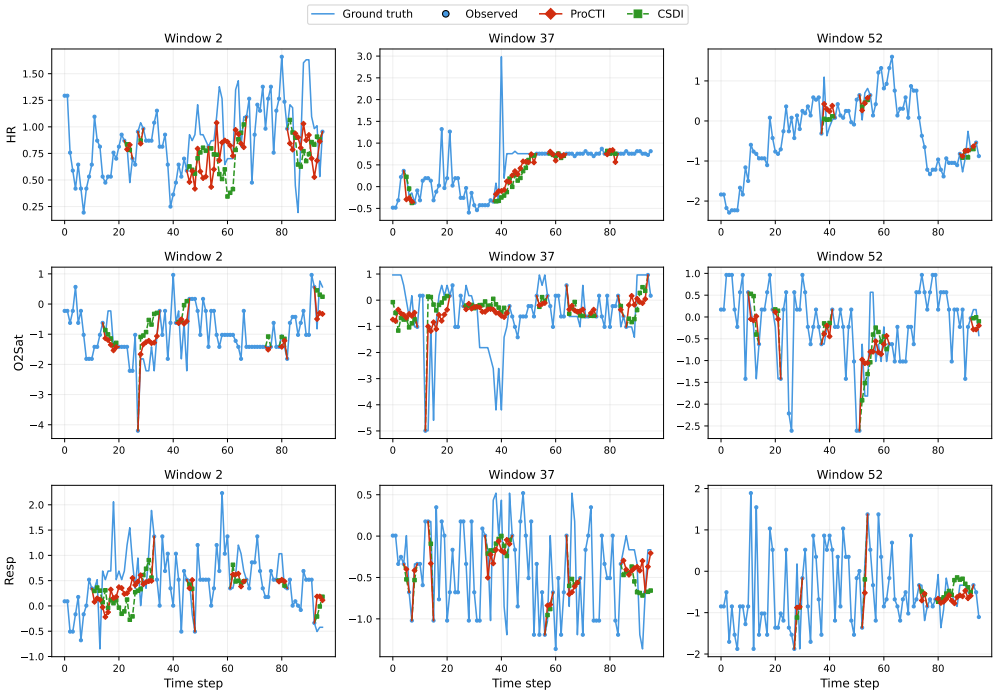


(b) ProCTI vs CSDI

Figure 7: Representative imputation comparison on the Gait dataset.

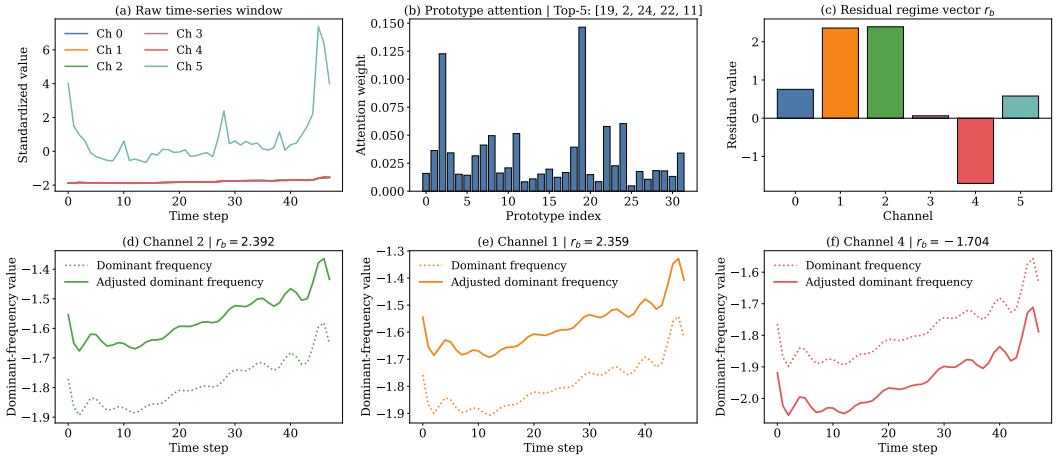


(a) ProCTI vs FGTI

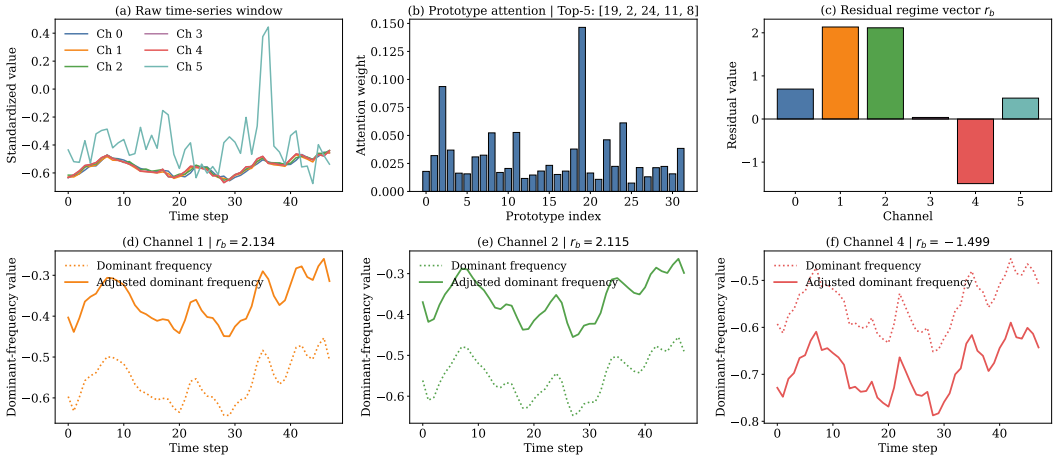


(b) ProCTI vs CSDI

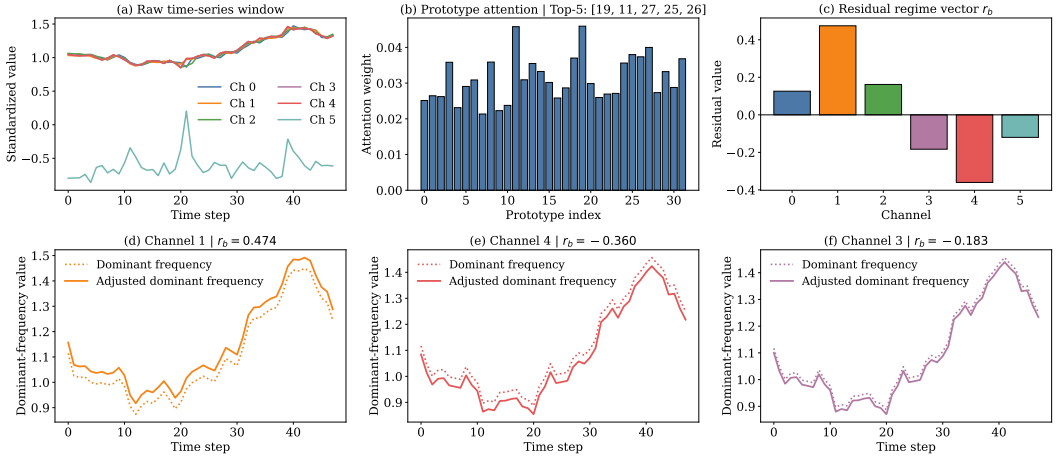
Figure 8: Representative imputation comparison on the PhysioNet dataset.



(a) Window 0, Top-3 $|r_b|$ channels: [2,1,4]



(b) Window 25, Top-3 $|r_b|$ channels: [1, 2, 4]



(c) Window 45, Top-3 $|r_b|$ channels: [1, 4, 3]

Figure 9: Prototype usage visualisation examples on the Stock dataset for four representative windows. Each panel shows (a) the raw time-series window, (b) prototype attention weights, (c) the residual regime vector r_b, and (d)–(f) the dominant-frequency signal before and after prototype-based adjustment for the three channels with the largest $|r_b|$. The prototype attention weights and r_b values are obtained after training for 50 epochs, at which point the learned scaling parameter is $\alpha = 0.0907$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction reflect the paper's contribution and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work in terms of resource consumption in Appendix A.4.2 and in terms of performance under attribute-wise missingness for datasets with high feature collinearity in Section 4.3.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All proofs are provided in Section 3.2.2.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include the detailed experimental setup in Section 4 and Appendix A.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We used publicly available datasets and we plan to publicly release our code once the status of the paper is confirmed.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We include the detailed experimental setup in Section 4 and Appendix A.2, A.3 and A.4.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All main results reported are averaged from five experiments of different random seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details about resource consumption in Appendix A.4.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper conforms to the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have provided a statement on broader impacts in Appendix A.7.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not pose high risks of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The data, code and models in the paper respect the license.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.