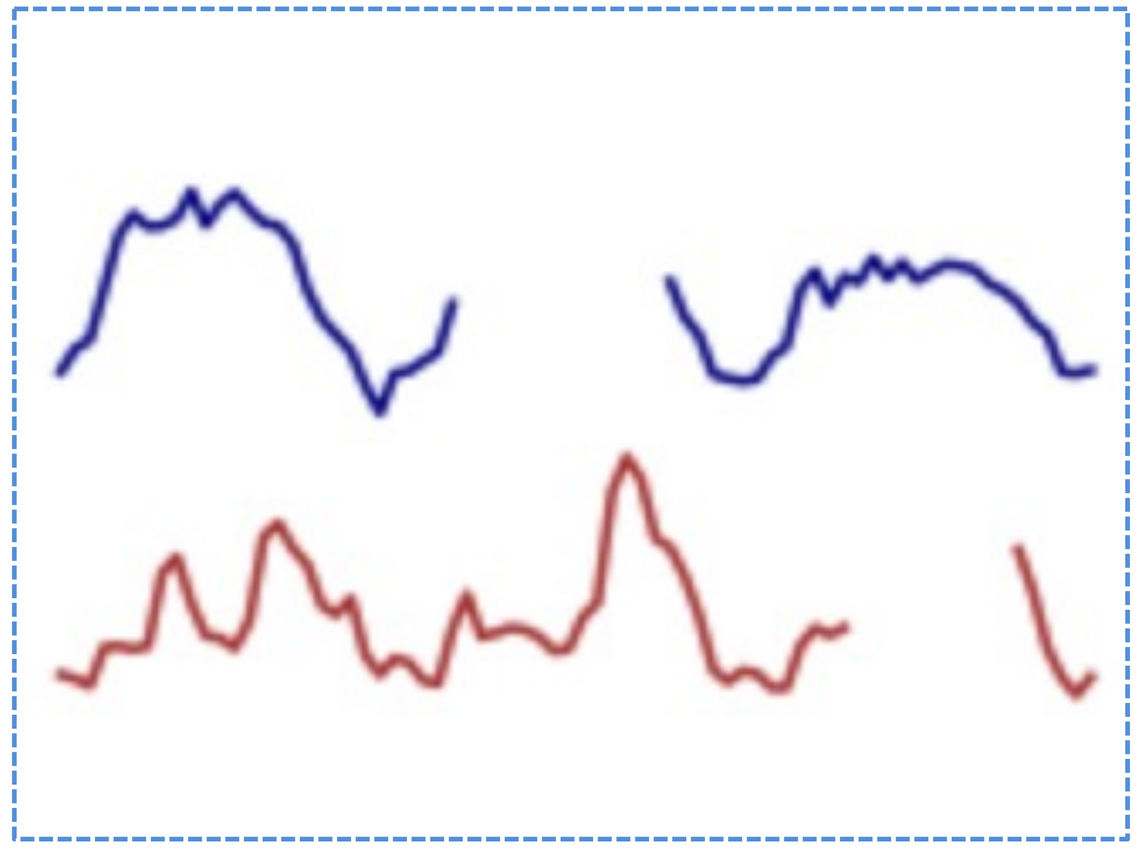
CSDI: Conditional score-based diffusion models for probabilistic time series imputation

Yusuke Tashiro[123], Jiaming Song[1], Yang Song[1], Stefano Ermon[1]

1. Stanford University
2. Mitsubishi UFJ Trust Investment Technology Institute (MTEC)
3. Japan Digital Design

Motivation: multivariate time series imputation

- Multivariate time series appear in many applications
  - e.g., Healthcare, finance, meteorology
- Time series data often contain missing values
  - could be harmful for downstream tasks



- Many imputation methods have been developed
  - imputation based on deep learning have shown good performance
    - use autoregressive models (e.g., RNNs)
  - still challenging to capture temporal and feature dependencies

Previous study: score-based diffusion models

- Gradually converts (denoises) noise to image

(cf. Ho et al. (2020))

- Score-based diffusion models achieved SOTA sample quality in many domains (Image, audio, graph, etc.)
  - some studies applied models to imputation tasks, but…

Previous study: imputation by score-based models

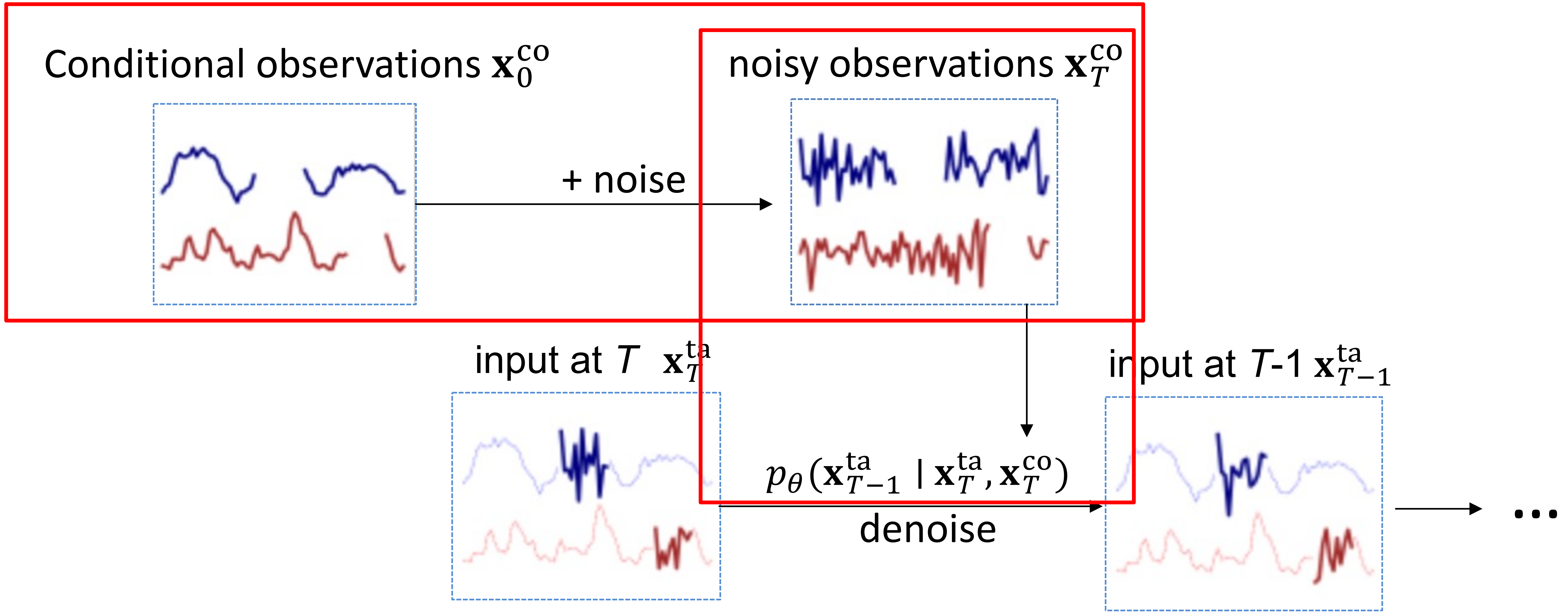
- Imputation task:

Conditional observations $\mathbf{x}_0^{\mathrm{co}}$ Imputation targets $\mathbf{x}_0^{\mathrm{ta}}$

- Approach in previous studies
  1. Train a score-based model (for unconditional generation)
  2. approximate conditional distribution by using the model

Previous study: imputation by score-based models

- Approximation at step T:



  – <u>Problem: added noise can reduce information</u>

Proposed method

- CSDI (**C**onditional **S**core-based **D**iffusion models for probabilistic time series **I**mputation)
  – explicitly utilize conditional observations $\mathbf{x}_0^{\mathrm{co}}$

Model

- Extend DDPM (denoising diffusion probabilistic models, Ho et al. (2020)) to conditional
  - DDPM considers the following diffusion model

forward process: $$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) := \mathcal{N}\left(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right)$$

reverse process: $$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t)\mathbf{I}).$$

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\alpha_t}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right), \ \sigma_\theta(\mathbf{x}_t, t) = \tilde{\beta}_t^{1/2}$$

($\alpha_t, \beta_t, \tilde{\beta}_t$: non-trainable scalar functions)

  - model can be trained by solving the optimization problem

$$\min_\theta \mathcal{L}(\theta) := \min_\theta \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)||_2^2$$

$$\text{where } \mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + (1-\alpha_t)\boldsymbol{\epsilon}.$$

Model

- Extend DDPM (denoising diffusion probabilistic models, Ho et al. (2020)) to conditional
  - CSDI considers the following diffusion model

forward process: $q(\mathbf{x}_t^{\text{ta}} \mid \mathbf{x}_{t-1}^{\text{ta}}) := \mathcal{N}\left(\sqrt{1-\beta_t}\mathbf{x}_{t-1}^{\text{ta}}, \beta_t \mathbf{I}\right)$

reverse process: $p_\theta(\mathbf{x}_{t-1}^{\text{ta}} \mid \mathbf{x}_t^{\text{ta}}, \mathbf{x}_0^{\text{co}}) := \mathcal{N}(\mathbf{x}_{t-1}^{\text{ta}}; \boldsymbol{\mu}_\theta(\mathbf{x}_t^{\text{ta}}, t \mid \mathbf{x}_0^{\text{co}}), \sigma_\theta(\mathbf{x}_t^{\text{ta}}, t \mid \mathbf{x}_0^{\text{co}})\mathbf{I}).$

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t^{\text{ta}}, t \mid \mathbf{x}_0^{\text{co}}) = \frac{1}{\alpha_t}\left(\mathbf{x}_t^{\text{ta}} - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{\text{ta}}, t \mid \mathbf{x}_0^{\text{co}})\right), \quad \sigma_\theta(\mathbf{x}_t^{\text{ta}}, t \mid \mathbf{x}_0^{\text{co}}) = \tilde{\beta}_t^{1/2}$$

($\alpha_t, \beta_t, \tilde{\beta}_t$: non-trainable scalar functions)

  - model can be trained by solving the optimization problem

$$\min_\theta \mathcal{L}(\theta) := \min_\theta \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} \|(\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{\text{ta}}, t \mid \mathbf{x}_0^{\text{co}}))\|_2^2$$

where $\mathbf{x}_t^{\text{ta}} = \sqrt{\alpha_t}\mathbf{x}_0^{\text{ta}} + (1-\alpha_t)\boldsymbol{\epsilon}.$

denoising function

Training method

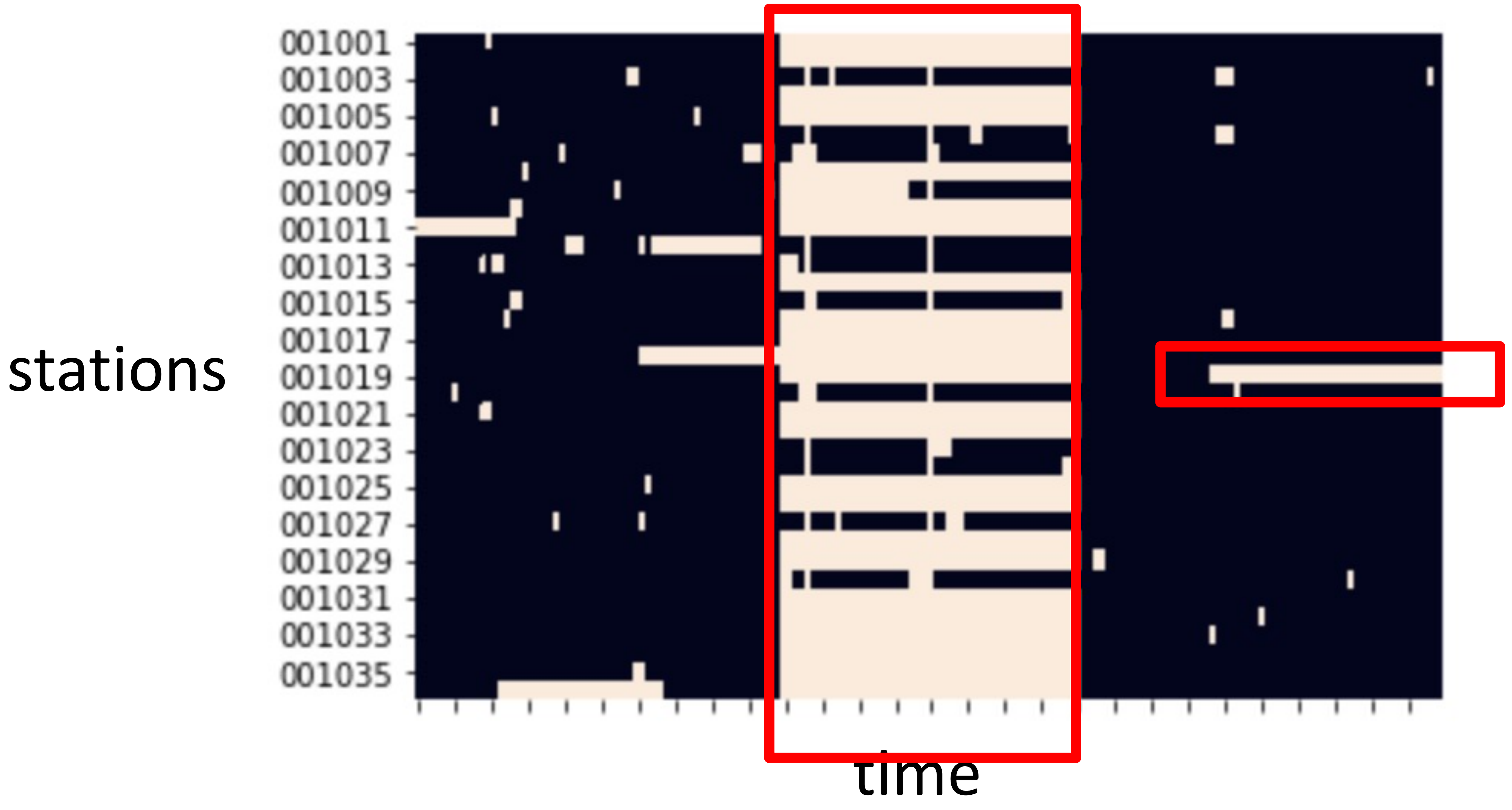
- Inspired by masked language modeling, we develop a self-supervised training method

Model architecture (denoising function)

- We adopt 2D attention mechanism to capture temporal and feature dependencies

Experiments: dataset

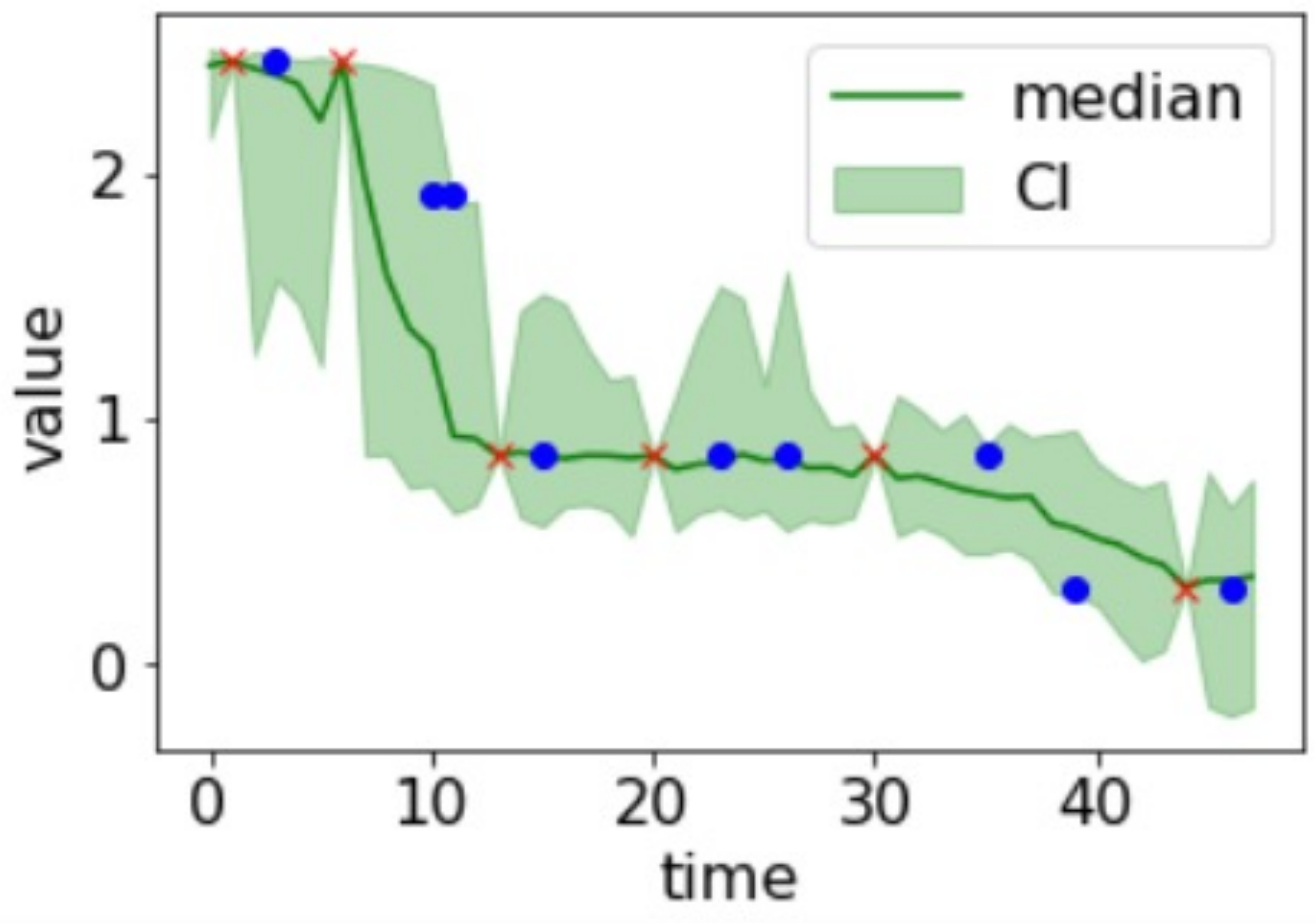
1. healthcare dataset (PhysioNet)
   - observations from ICU (35 variables for 48 hours)
   - missing pattern is random
2. air quality dataset
   - PM2.5 in Beijing (from 36 stations, 36 hours as one time series)
   - missing pattern is not random
     - sequential missing
     - block missing



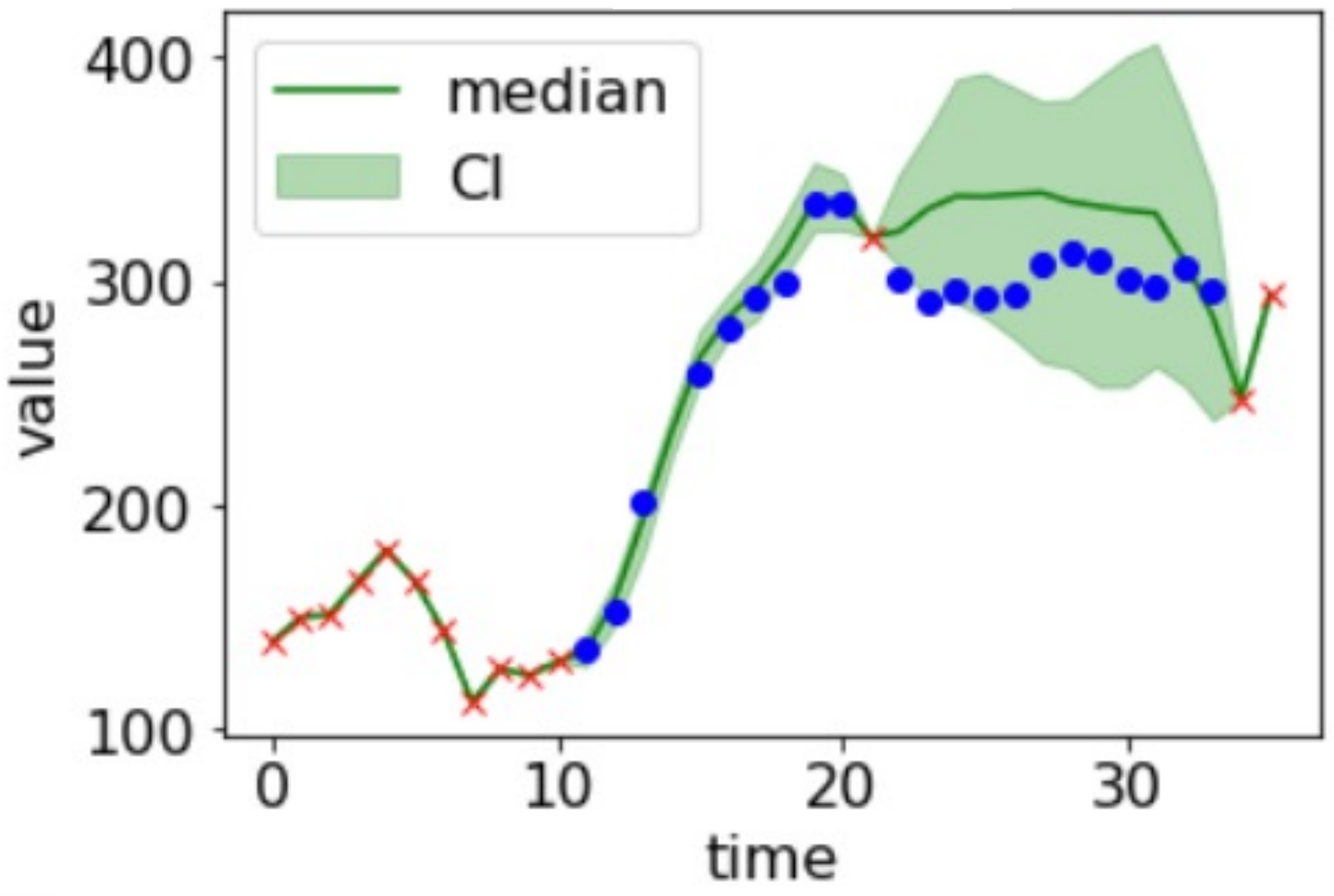
Experiment: example

- impute missing values 100 times and calculate confidence intervals
- CSDI provides reasonable probabilistic imputation
  - imputation targets (blue) are within confidence intervals (green)

healthcare



air quality



(red: conditional observations, blue: imputation targets)

Experiment: comparison with probabilistic methods

- CSDI significantly outperforms existing probabilistic methods
- CSDI outperforms imputation by unconditional score-based model

(metric: CRPS)

	healthcare			air quality
	10% missing	50% missing	90% missing	
Multitask GP [31]	0.483(–)	0.588(–)	0.964(–)	0.304(–)
GP-VAE [10]	0.574(0.003)	0.774(0.004)	0.998(0.001)	0.397(0.009)
V-RIN [32]	0.808(0.008)	0.831(0.005)	0.922(0.003)	0.526(0.025)
unconditional	0.360(0.007)	0.458(0.008)	0.671(0.007)	0.135(0.001)
CSDI (proposed)	**0.238**(**0.001**)	**0.330**(**0.002**)	**0.522**(**0.002**)	**0.108**(**0.001**)

Experiment: comparison with deterministic methods

- We use the median of samples as a point estimate
- CSDI outperforms deterministic imputation methods

(metric: MAE)

	healthcare			air quality
	10% missing	50% missing	90% missing	
V-RIN [18]	0.271(0.001)	0.365(0.002)	0.606(0.006)	25.4(0.62)
BRITS [3]	0.284(0.001)	0.368(0.002)	0.517(0.002)	14.11(0.26)
BRITS [3] (*)	0.278	–	–	11.56
GLIMA [36] (*)	0.265	–	–	10.54
RDIS [6]	0.319(0.002)	0.419(0.002)	0.631(0.002)	22.11(0.35)
unconditional	0.326(0.008)	0.417(0.010)	0.625(0.010)	12.13(0.07)
CSDI (proposed)	**0.217(0.001)**	**0.301(0.002)**	**0.481(0.003)**	**9.60(0.04)**

Experiments: multivariate time series forecasting

- We can apply CSDI to probabilistic forecasting
  – Consider future values as missing values
  – CSDI achieves competitive performance (outperforms baselines on 3 of 5 datasets)

(metric: CRPS-sum)

	solar	electricity	traffic	taxi	wiki
GP-copula [27]	0.337(0.024)	0.024(0.002)	0.078(0.002)	0.208(0.183)	0.086(0.004)
TransMAF [25]	0.301(0.014)	0.021(0.000)	0.056(0.001)	0.179(0.002)	0.063(0.003)
TLAE [20]	**0.124(0.033)**	0.040(0.002)	0.069(0.001)	0.130(0.006)	0.241(0.001)
TimeGrad [24]	0.287(0.020)	0.021(0.001)	0.044(0.006)	**0.114(0.020)**	0.049(0.002)
CSDI (proposed)	0.298(0.004)	**0.017(0.000)**	**0.020(0.001)**	0.123(0.003)	**0.047(0.003)**

Summary

- CSDI utilizes conditional score-based models for probabilistic time series imputation

- Future directions
  - fast sampling
  - application to downstream tasks
  - extension to other domains