Appendix

A Implementation details

A.1 Neural network architectures

We use a convolutional neural network (CNN) as an inference network and a fully connected multilayer perceptron (MLP) as a generative network. The inference network convolves over the time dimension of the input data and allows for sequences of variable lengths. It consists of a number of convolutional layers that integrate information from neighboring time steps into a joint representation using a fixed receptive field (see Figure 1). The CNN outputs a tensor of size $\mathbb{R}^{T \times 3k}$, where k is the dimensionality of the latent space. Every row corresponds to a time step t and contains $3k$ parameters, which are used to predict the mean vector $\mathbf{m}_t$ as well as the diagonal and off-diagonal elements $\{b^j_{t,t}, b^j_{t,t+1}\}_{j=1:k}$ that characterize $\mathbf{B}$ at the given time step. More details about the network structures for the different experiments are provided in the following.

A.2 Healing MNIST

For the Healing MNIST data, we used a few 2D convolutional layers as preprocessors, since the single temporal states of the data are images. Those layers convolve over the image dimensions of each time step separately and generate latent features for each image. We then flatten these latent features and use them as inputs to the 1D convolution over time. The hyperparameters for the setup are given in Table S1.

Table S1: Hyperparameters used in the GP-VAE model for the experiment on Healing MNIST. Some of the parameters are only relevant in a subset of the models.

Hyperparameter	Value
Number of CNN layers in inference network	1
Number of filters per CNN layer	256
Filter size (i.e., time window size)	3
Number of feedforward layers in inference network	2
Width of feedforward layers	256
Dimensionality of latent space	256
Length scale of Cauchy kernel	2.0
Number of feedforward layers in generative network	3
Width of feedforward layers	256
Activation function of all layers	ReLU
Learning rate during training	0.001
Optimizer	Adam [Kingma and Ba, 2015]
Number of training epochs	20
Train/val/test split of data set	50,000/10,000/10,000
Dimensionality of time points	784
Length of time series	10
Tradeoff parameter β	0.8

A.3 SPRITES

Since the SPRITES data also consist of images, we use the same 2D convolutional preprocessing as in the Healing MNIST experiment. The hyperparameters are reported in Table S2.

A.4 Real medical time series data

For the Physionet 2012 data, we omitted the convolutional preprocessing, since these are not image data. We hence directly feed the data into the 1D convolutional layer over time. In this experiment, we follow the evaluation protocol from BRITS [Cao et al., 2018] and use the same set for training and evaluation. We randomly eliminate 10 % of observed measurements from the data for validation.

Table S2: Hyperparameters used in the GP-VAE model for the experiment on SPRITES. Some of the parameters are only relevant in a subset of the models.

Hyperparameter	Value
Number of CNN layers in inference network	1
Number of filters per CNN layer	32
Filter size (i.e., time window size)	3
Number of feedforward layers in inference network	2
Width of feedforward layers	256
Dimensionality of latent space	256
Length scale of Cauchy kernel	2.0
Number of feedforward layers in generative network	3
Width of feedforward layers	256
Activation function of all layers	ReLU
Learning rate during training	0.001
Optimizer	Adam [Kingma and Ba, 2015]
Number of training epochs	20
Train/val/test split of data set	8,000/1,000/1,000
Dimensionality of time points	12288
Length of time series	8
Tradeoff parameter β	0.1

The hyperparameters for this experiment are reported in Table S3.

Table S3: Hyperparameters used in the GP-VAE model for the experiment on medical time series from the Physionet data set. Some of the parameters are only relevant in a subset of the models.

Hyperparameter	Value
Number of CNN layers in inference network	1
Number of filters per CNN layer	128
Filter size (i.e., time window size)	24
Number of feedforward layers in inference network	1
Width of feedforward layers	128
Dimensionality of latent space	35
Length scale of Cauchy kernel	7.0
Number of feedforward layers in generative network	2
Width of feedforward layers	256
Activation function of all layers	ReLU
Learning rate during training	0.001
Optimizer	Adam [Kingma and Ba, 2015]
Number of training epochs	40
Train/val/test split of data set	4,000
Dimensionality of time points	35
Length of time series	48
Tradeoff parameter β	0.2

B Additional experiments

B.1 Missingness rates on Healing MNIST

To explore the influence of the missingness rates on the performance of the models, we conducted an experiment where we introduced missingness into the Healing MNIST time series at different rates between 10 % and 90 %. We compare the performance in terms of negative log likelihood (S4).

Table S4: Performance of different models on Healing MNIST data with artificial missingness and different missingness rates. We report negative log likelihood (lower is better). The reported values are means and their respective standard errors over the test set.

Missingness	VAE	HI-VAE	GP-VAE (RBF kernel)	GP-VAE (proposed)
10 %	0.0125 ± 0.0000	0.0117 ± 0.0000	0.0113 ± 0.0000	**0.0109 ± 0.0000**
20 %	0.0319 ± 0.0001	0.0275 ± 0.0001	0.0262 ± 0.0001	**0.0258 ± 0.0001**
30 %	0.0640 ± 0.0002	0.0507 ± 0.0002	0.0478 ± 0.0001	**0.0465 ± 0.0001**
40 %	0.1106 ± 0.0003	0.0803 ± 0.0003	0.0781 ± 0.0002	**0.0743 ± 0.0002**
50 %	0.1931 ± 0.0006	0.1349 ± 0.0005	0.1217 ± 0.0004	**0.1196 ± 0.0004**
60 %	0.3360 ± 0.0010	0.2218 ± 0.0008	0.1982 ± 0.0006	**0.1957 ± 0.0006**
70 %	0.6216 ± 0.0019	0.3818 ± 0.0014	0.3426 ± 0.0010	**0.3261 ± 0.0010**
80 %	1.3284 ± 0.0042	0.7613 ± 0.0025	0.6798 ± 0.0020	**0.6488 ± 0.0019**
90 %	4.0610 ± 0.0127	2.1878 ± 0.0064	1.9424 ± 0.0054	**1.8464 ± 0.0050**

It can be seen that the proposed model outperforms the other deep architectures (including the GP-VAE with an RBF kernel) for all the different missingness rates. This also highlights that the Cauchy kernel does indeed help in modeling the temporal dynamics.

B.2 Missingness mechanisms on Healing MNIST

So far, in our synthetic experiments we only looked at artificial missingness that was introduced completely at random (MCAR), that is, we uniformly sampled features to be missing independently of each other and their value. In this experiment, we explore a few more structured missingness mechanisms which are described in the following. The average missingness rate for all the different mechanisms is around 50 %.

Feature correlation (*Spatial*). We assume different features to be correlated in their missingness, that is, a feature is more likely to be missing if certain other features are missing. We implement that by defining a spatial Gaussian process with RBF kernel on the Healing MNIST images and drawing the missingness patterns as samples from this process. Neighboring pixels are therefore correlated in their missingness.

Positive temporal correlation (*Temporal*$^+$). The missingness of features is positively correlated in time, that is, if a feature is missing at one time step, it is more likely to be missing at the consecutive time step. We implement this again with a Gaussian process with RBF kernel, this time defined over time for each feature separately.

Negative temporal correlation (*Temporal*$^-$). The missingness of features is negatively correlated in time, that is, if a feature is missing at one time step, it is less likely to be missing at the consecutive time step. We implement this with a determinantal point process (DPP) over time for each feature separately.

Missingness not at random (*MNAR*). In this setting, the missingness is actually dependent on the underlying ground-truth value of the feature. In our example, white pixels in the Healing MNIST images are twice as likely to be missing than black pixels.

We assessed our model and the baselines on all these different settings and report the results in terms of likelihood in Table S5.

We observe that our proposed model outperforms all baselines in terms of likelihood and MSE on all the different

Table S5: Performance of different models on Healing MNIST data with artificial missingness and different missingness mechanisms. We report negative log likelihood (lower is better). The reported values are means and their respective standard errors over the test set.

Mechanism	VAE	HI-VAE	GP-VAE (proposed)
Spatial	0.4802 ± 0.0016	0.2259 ± 0.0010	**0.1779 ± 0.0007**
Temporal$^+$	0.1918 ± 0.0006	0.1332 ± 0.0005	**0.1206 ± 0.0004**
Temporal$^-$	0.1940 ± 0.0006	0.1349 ± 0.0005	**0.1175 ± 0.0004**
MNAR	0.4798 ± 0.0016	0.2896 ± 0.0010	**0.2606 ± 0.0008**

missingness mechanisms. We also observe that the MNAR setting seems to be the hardest one for the VAE-based models, followed by the setting with correlated features, whereas the settings with temporal correlation do not seem to be harder than the completely random ones (compare to Tab. S4). For the single imputation methods (mean and forward imputation), the MNAR setting also seems to be the hardest one, while the correlated features are not much harder than random. However, the positive temporal correlation is harder for those methods and the negative temporal correlation is even easier than random missingness.