STRUCTURES
CLUSTER OF
EXCELLENCE

IMMERSE

Multimodal Teacher Forcing for Reconstructing Nonlinear Dynamical Systems

Manuel Brenner[1,2], Georgia Koppe*[1,4], Daniel Durstewitz*[1,2,3]

*Equal contribution
[1]Dept. of Theoretical Neuroscience, Central Institute of Mental Health, Mannheim, Germany,
[2]Faculty of Physics and Astronomy, Heidelberg University, Germany,
[3]Interdisciplinary Center for Scientific Computing, Heidelberg University, Germany
[4]Dept. of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty, Heidelberg University, Germany

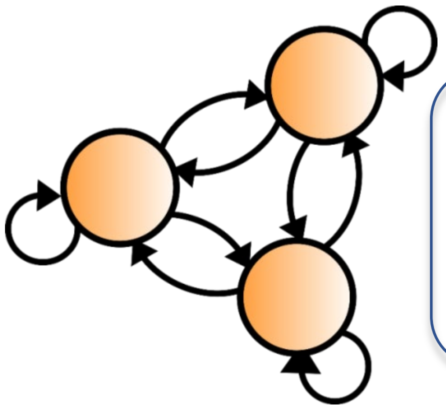
When Machine Learning meets Dynamical Systems: Theory and Application

Dynamical Systems Reconstructions (DSR)

Unknown DS

$$\dot{x}_1 = s(x_2 - x_1)$$
$$\dot{x}_2 = rx_1 - x_2 - x_1x_3$$
$$\dot{x}_3 = x_1x_2 - bx_3$$

Inference

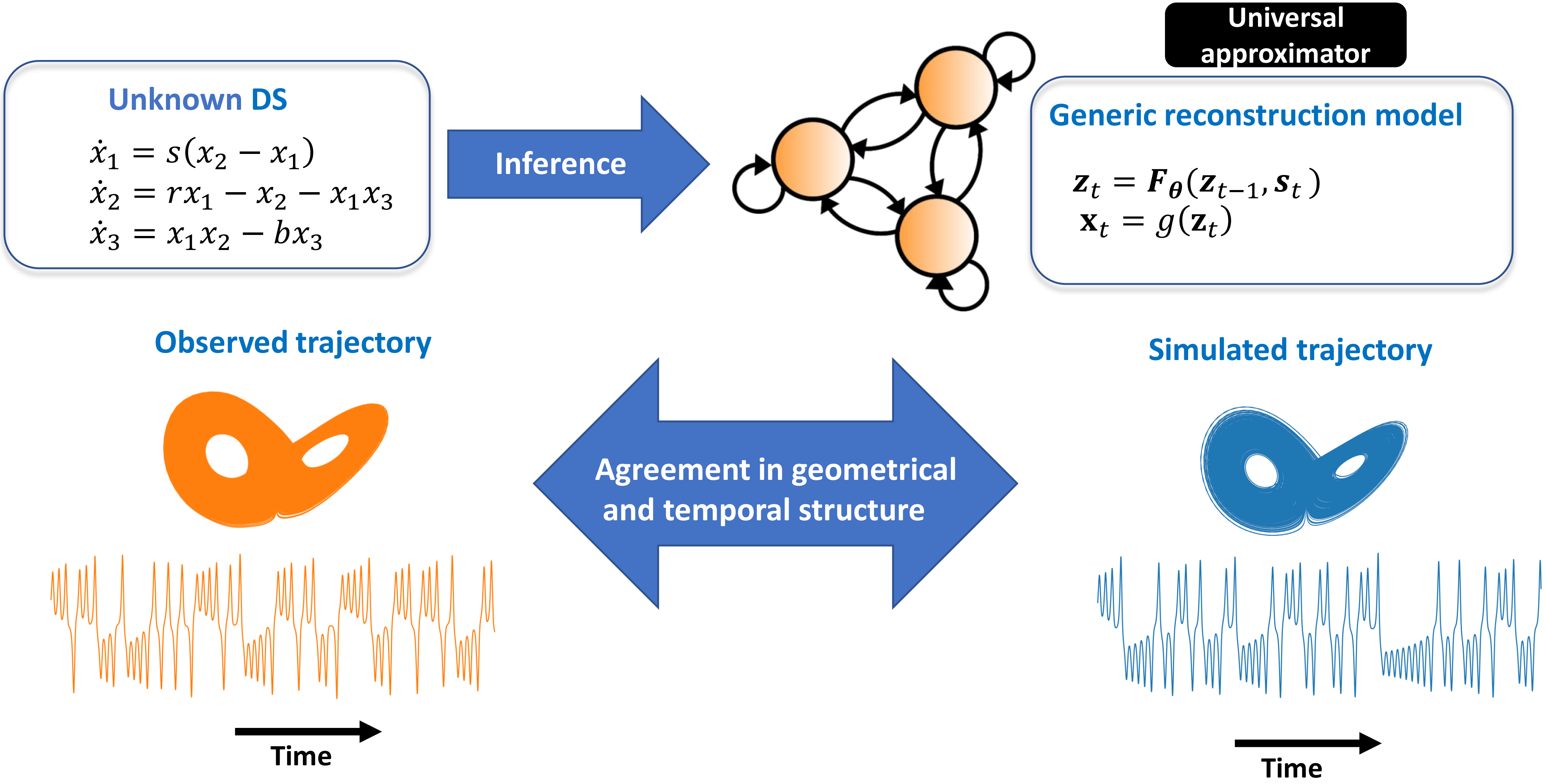


Universal approximator

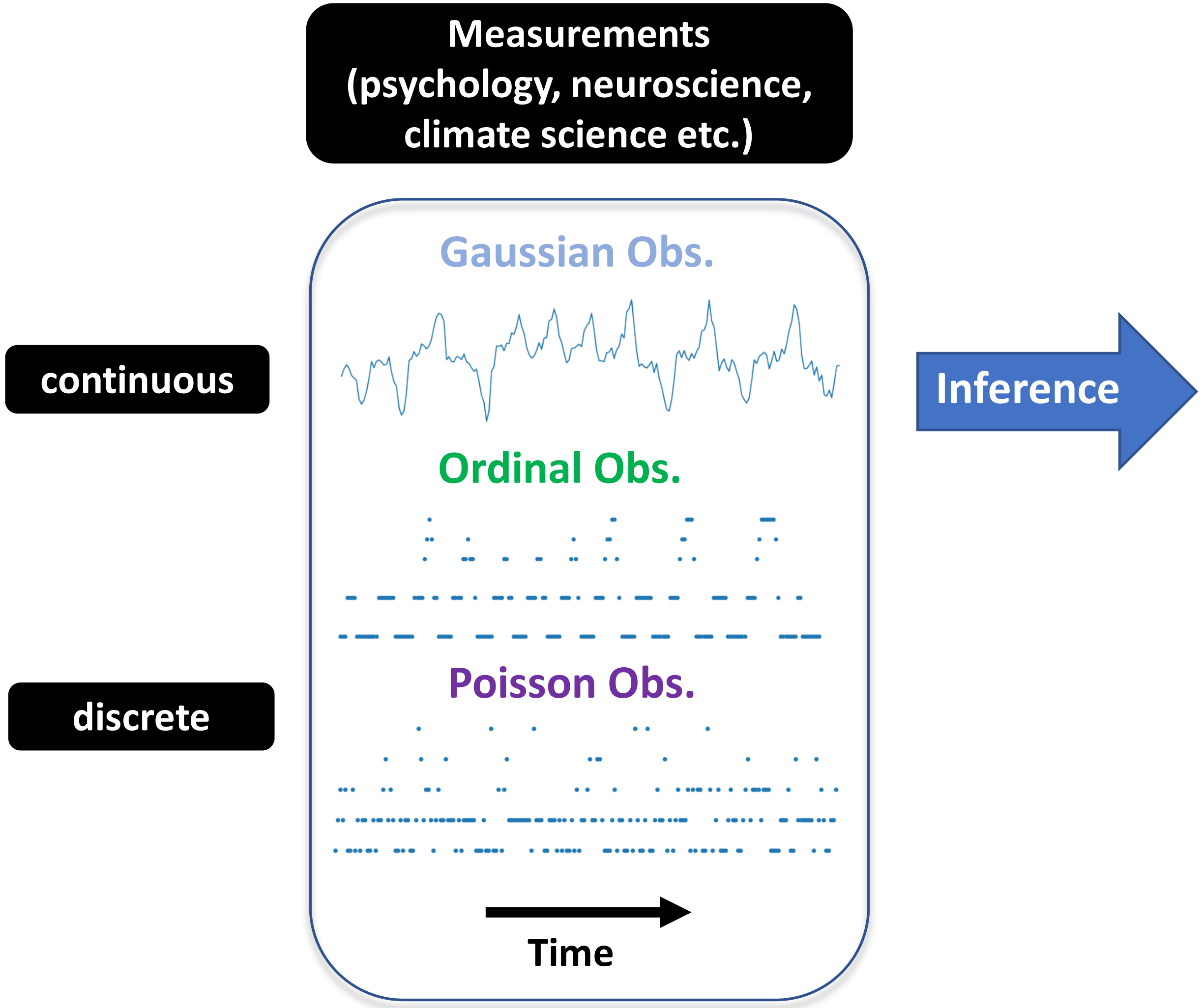
Generic reconstruction model

$$\boldsymbol{z}_t = \boldsymbol{F}_{\boldsymbol{\theta}}(\boldsymbol{z}_{t-1}, \boldsymbol{s}_t)$$
$$\mathbf{x}_t = g(\mathbf{z}_t)$$

Dynamical Systems Reconstructions (DSR)



DSR from Multimodal Time Series

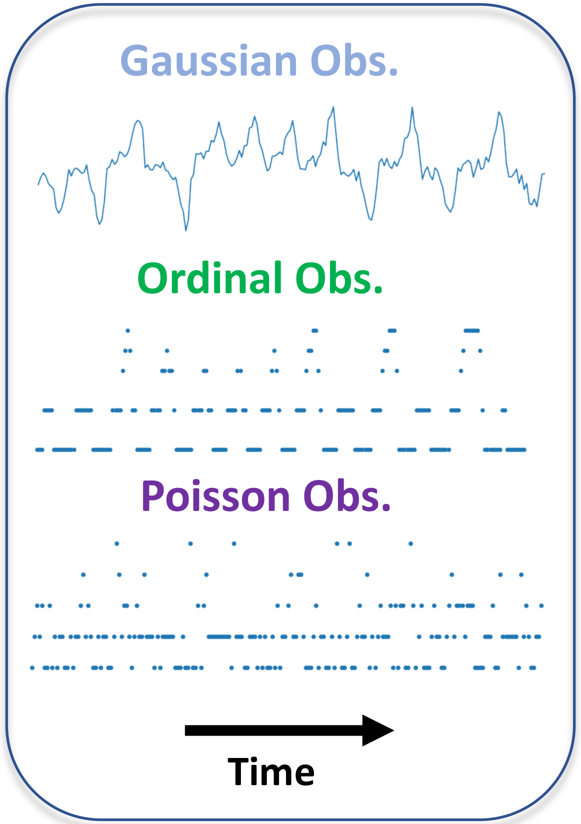


DSR from Multimodal Time Series

Measurements (psychology, neuroscience, climate science etc.)

continuous

discrete



Inference

$$\cdots \; z_{t-1} \rightarrow z_t \rightarrow z_{t+1} \; \cdots$$

dendPLRNN (Brenner, Hess et al., ICML 2022)

Math. tractability

$$\boldsymbol{z}_t = \boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{W}\phi(\boldsymbol{z}_{t-1}) + \boldsymbol{h}$$

$$\phi(\boldsymbol{z}_{t-1}) = \sum_{b=1}^{B} \alpha_b \max(0, \boldsymbol{z}_{t-1} - \boldsymbol{h}_b)$$

Different distributional assumptions

Gaussian Obs.

$$\boldsymbol{x}_t | \boldsymbol{z}_t \sim \mathcal{N}(\mathbf{B}\boldsymbol{z}_t, \boldsymbol{\Gamma})$$

Ordinal Obs.

$$\boldsymbol{o}_t | \boldsymbol{z}_t \sim \text{Ordinal}(\beta \boldsymbol{z}_t, \epsilon)$$

Poisson Obs.

$$\boldsymbol{p}_t | \boldsymbol{z}_t \sim \text{Poisson}(\lambda(\boldsymbol{z}_t))$$

Challenges of DSR

Challenge

Chaotic dynamics

↓

Divergent loss gradients
(Mikhaeil, Monfared&Durstewitz, NeurIPS 2022)

Theorem 2. *Suppose that an RNN $F_{\boldsymbol{\theta}} \in \mathcal{R}$ (parameterized by $\boldsymbol{\theta}$) has a chaotic attractor Γ^* with $\mathcal{B}_{\Gamma^*}$ as its basin of attraction. Then, for almost every orbit with $\boldsymbol{z}_1 \in \mathcal{B}_{\Gamma^*}$, (i) the Jacobians connecting temporally distal states $\boldsymbol{z}_T$ and $\boldsymbol{z}_t$ ($T \gg t$), $\frac{\partial \boldsymbol{z}_T}{\partial \boldsymbol{z}_t}$, will exponentially explode for $T \to \infty$, and (ii) the tangent vector $\frac{\partial \boldsymbol{z}_T}{\partial \boldsymbol{\theta}}$ and so the gradients of the loss function, $\frac{\partial \mathcal{L}_T}{\partial \boldsymbol{\theta}}$, will diverge as $T \to \infty$.*

Challenges of DSR

Challenge

Chaotic dynamics

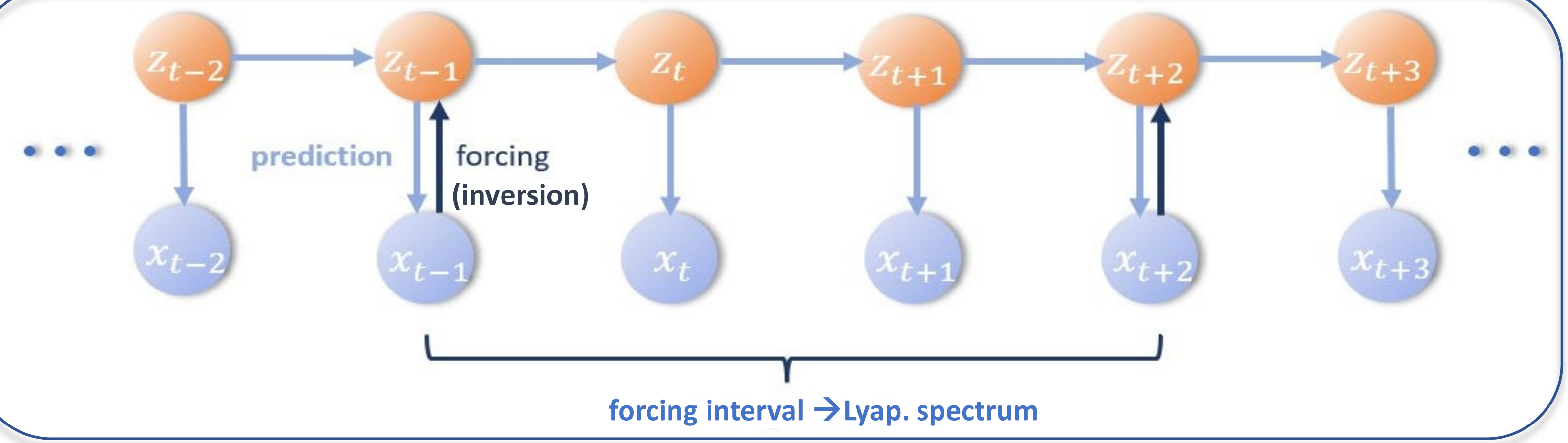
↓

Divergent loss gradients
(Mikhaeil, Monfared&Durstewitz, NeurIPS 2022)

Theorem 2. *Suppose that an RNN $F_{\theta} \in \mathcal{R}$ (parameterized by θ) has a chaotic attractor Γ^* with $\mathcal{B}_{\Gamma^*}$ as its basin of attraction. Then, for almost every orbit with $z_1 \in \mathcal{B}_{\Gamma^*}$, (i) the Jacobians connecting temporally distal states z_T and z_t $(T \gg t)$, $\frac{\partial z_T}{\partial z_t}$, will exponentially explode for $T \to \infty$, and (ii) the tangent vector $\frac{\partial z_T}{\partial \theta}$ and so the gradients of the loss function, $\frac{\partial \mathcal{L}_T}{\partial \theta}$, will diverge as $T \to \infty$.*

Proposed Solution

Sparse teacher forcing (TF)

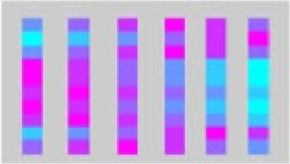


SOTA performance for DSR on unimodal normal data (Brenner, Hess et al., ICML 2022)

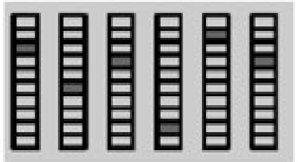
Multimodal Variational Autoencoders

Observations

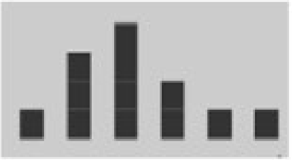
Gaussian Obs.



Ordinal Obs.



Poisson Obs.



$$\boldsymbol{Y} = \{\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}; \{\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T\}; \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_T\}\}$$

Encoder (CNN)

Approximate posterior

$$q_{\phi}(\tilde{\boldsymbol{Z}}|\boldsymbol{Y}) = \mathcal{N}(\boldsymbol{\mu}_{\phi}(\boldsymbol{Y}), \boldsymbol{\Sigma}_{\phi}(\boldsymbol{Y}))$$

Decoder

Observation Models

$$\boldsymbol{x}_t \mid \tilde{\boldsymbol{z}}_t \sim \mathcal{N}(\boldsymbol{B}\tilde{\boldsymbol{z}}_t, \boldsymbol{\Gamma})$$

$$\boldsymbol{o}_t \mid \tilde{\boldsymbol{z}}_t \sim \mathrm{Ordinal}(\boldsymbol{\beta}\tilde{\boldsymbol{z}}_t, \epsilon)$$

$$\boldsymbol{c}_t \mid \tilde{\boldsymbol{z}}_t \sim \mathrm{Poisson}(\lambda(\tilde{\boldsymbol{z}}_t))$$

ELBO Loss

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}; \boldsymbol{Y}) = -\mathbb{E}_{q_{\phi}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\tilde{\boldsymbol{Z}}) + \log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{Z}})] - \mathbb{H}_{q_{\phi}}(\tilde{\boldsymbol{Z}} \mid \boldsymbol{Y})$$

Data Likelihoods

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\tilde{\boldsymbol{Z}}) = \sum_{t=1}^{T} (\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\tilde{\boldsymbol{z}}_t) + \log p_{\boldsymbol{\theta}}(\boldsymbol{o}_t|\tilde{\boldsymbol{z}}_t) + \log p_{\boldsymbol{\theta}}(\boldsymbol{c}_t|\tilde{\boldsymbol{z}}_t))$$

Multimodal Teacher Forcing (MVAE-TF)

$$\mathcal{L} = \mathcal{L}_{MVAE} + \mathcal{L}_{con} + \mathcal{L}_{PLRNN}$$

MVAE Loss

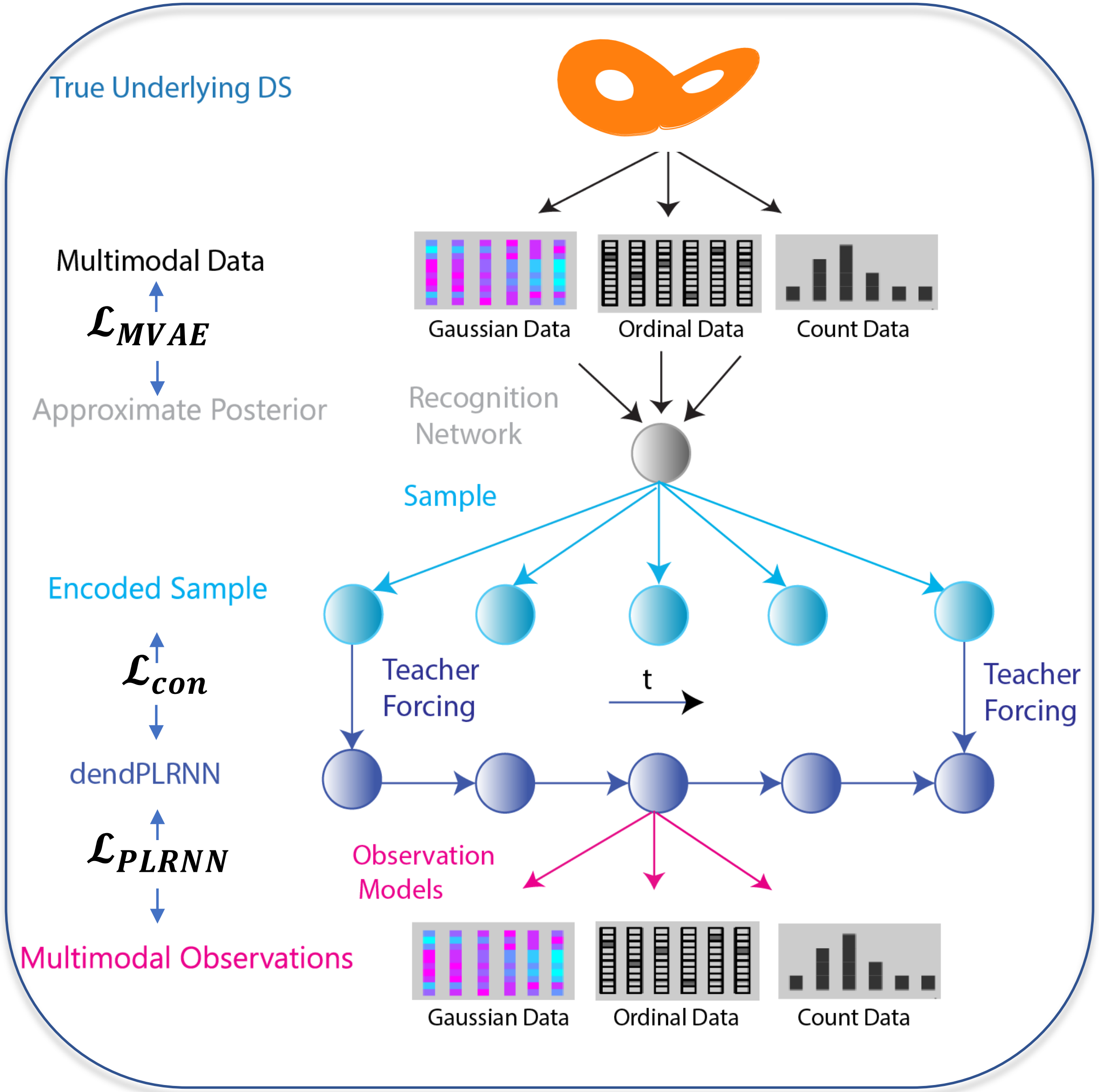
$$\mathcal{L}_{MVAE} = -\mathbb{E}_{q_\phi}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\tilde{\boldsymbol{Z}}) - \mathbb{H}_{q_\phi}(\tilde{\boldsymbol{Z}} \mid \boldsymbol{Y})$$

Consistency Loss between MVAE and PLRNN

$$\mathcal{L}_{con} = \frac{1}{2}\sum_{t=2}^{T}\left(\log|\boldsymbol{\Sigma}| + (\tilde{\boldsymbol{z}}_t - \boldsymbol{z}_{1:K,t})^\top \boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{z}}_t - \boldsymbol{z}_{1:K,t})\right)$$

Data likelihoods from PLRNN

$$\mathcal{L}_{\mathrm{PLRNN}} = -\sum_{t=1}^{T}\left(\log p_{\boldsymbol{\theta}}(\boldsymbol{x_t}|\boldsymbol{z}_{1:K,t}) + \log p_{\boldsymbol{\theta}}(\boldsymbol{o}_t|\boldsymbol{z}_{1:K,t}) + \log p_{\boldsymbol{\theta}}(\boldsymbol{c}_t|\boldsymbol{z}_{1:K,t})\right)$$



Multimodal Reconstructions

Randomly initialized obs. models

Underlying DS

$\boldsymbol{z}_t$

Noisy Gaussian Obs.

$$\boldsymbol{x}_t |\, \boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{z}_t, \boldsymbol{\Gamma})$$

Ordinal Obs.

$$\boldsymbol{o}_t |\, \boldsymbol{z}_t \sim \text{Ordinal}(\beta \boldsymbol{z}_t, \epsilon)$$

Poisson Obs.

$$\boldsymbol{p}_t |\, \boldsymbol{z}_t \sim \text{Poisson}(\lambda(\boldsymbol{z}_t))$$

Multimodal Reconstructions

Randomly initialized obs. models

Underlying DS

z_t

Noisy Gaussian Obs.

$x_t | z_t \sim \mathcal{N}(z_t, \Gamma)$

Ordinal Obs.

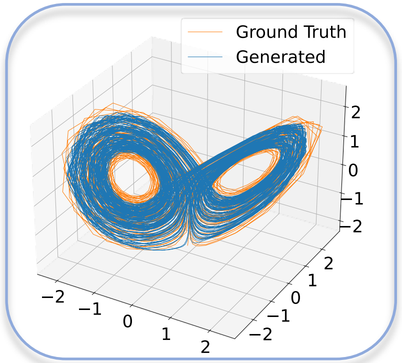
$o_t | z_t \sim \text{Ordinal}(\beta z_t, \epsilon)$

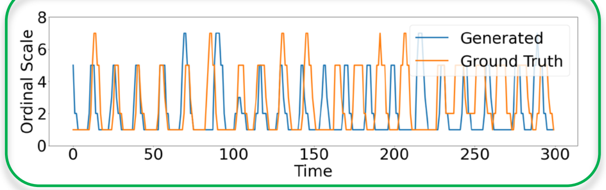
Poisson Obs.

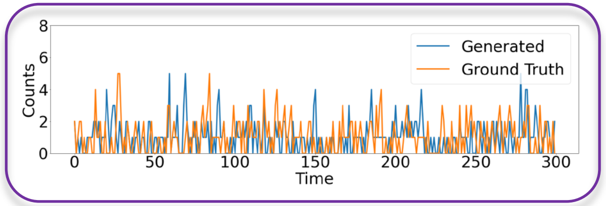
$p_t | z_t \sim \text{Poisson}(\lambda(z_t))$

Geometric Agreement

$$p_{\text{gen}}(\mathbf{x} \mid \mathbf{z})) \approx \sum_{k=1}^{K} \hat{p}_{\text{true}}^{(k)}(\mathbf{x}) \log \left(\frac{\hat{p}_{\text{true}}^{(k)}(\mathbf{x})}{\hat{p}_{\text{gen}}^{(k)}(\mathbf{x} \mid \mathbf{z})} \right)$$

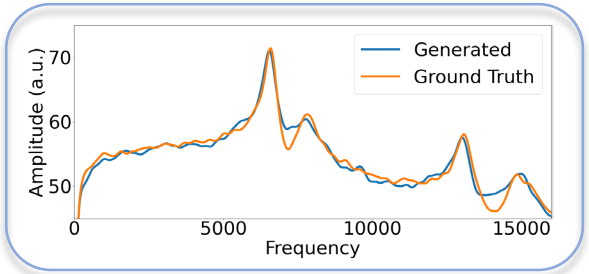


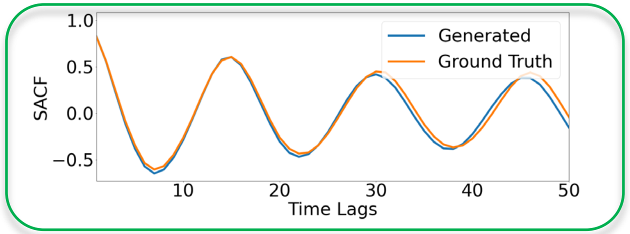


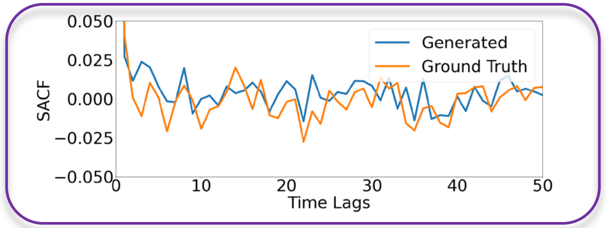


Temporal Agreement

$$H(F(\omega), G(\omega)) = \sqrt{1 - \int_{-\infty}^{\infty} \sqrt{F(\omega)G(\omega)} d\omega} \in [0,1]$$

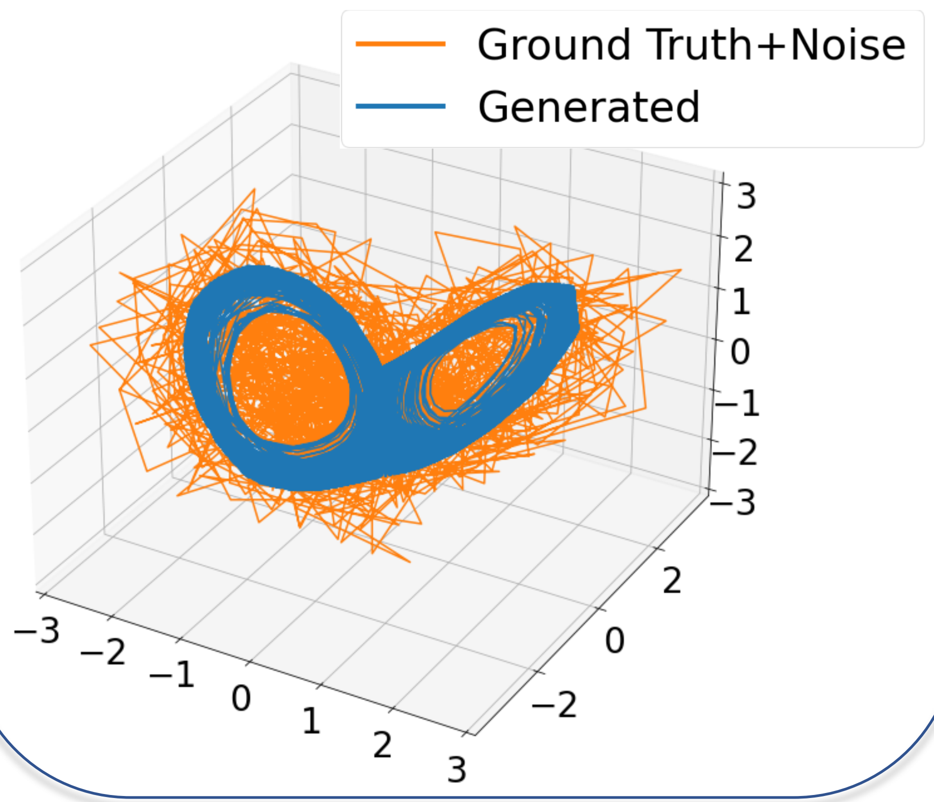




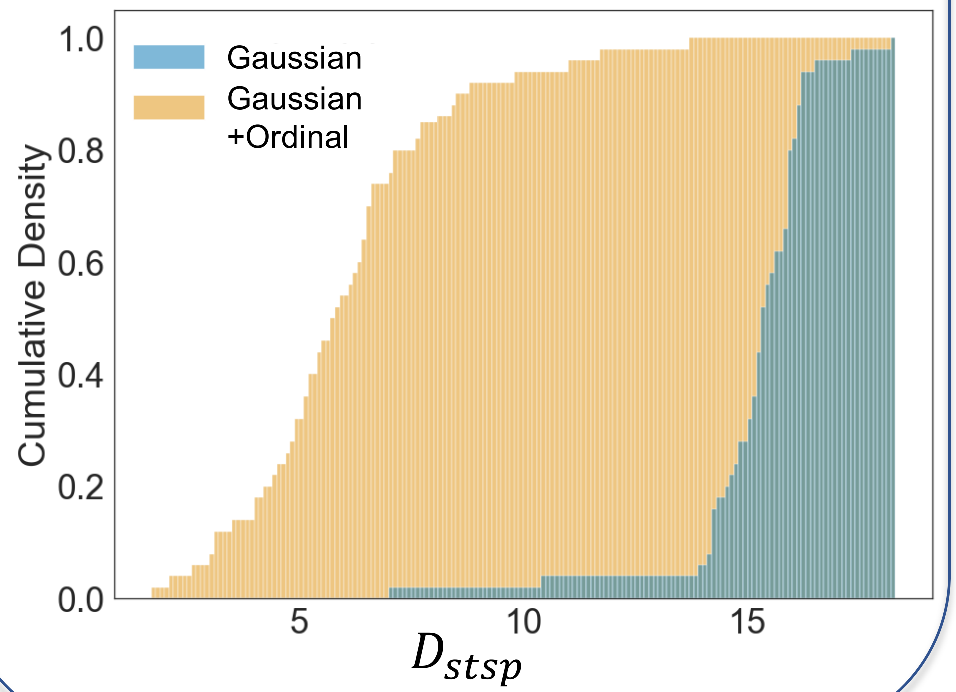


Challenging Data Settings

Reconstructed State Space

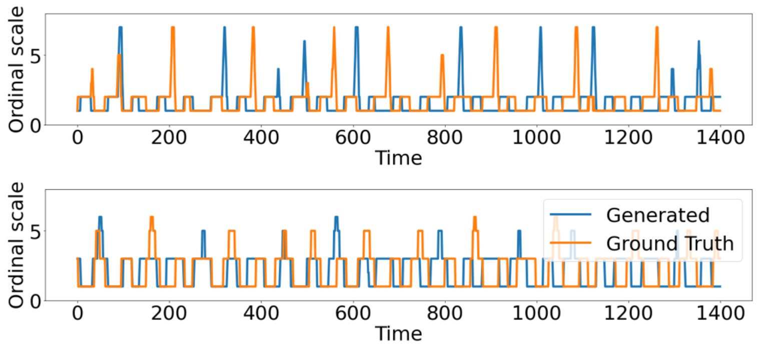


Multimodal Data Improve DSR

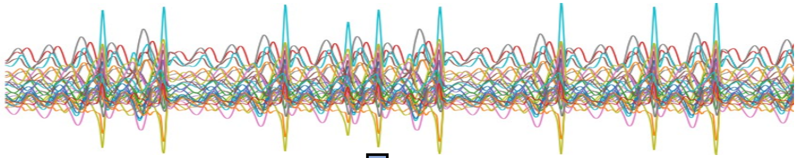


Discrete Observations&Experimental Data

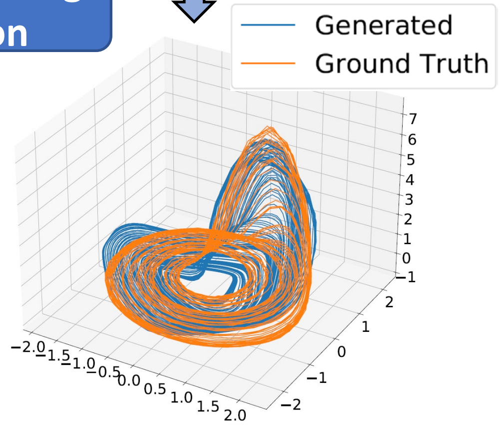
Chaotic Rössler attractor from discrete ordinal observations



Train RNN model

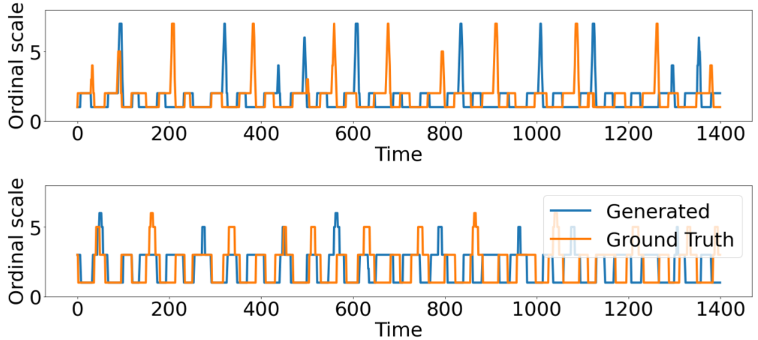


Geometry preserving transformation

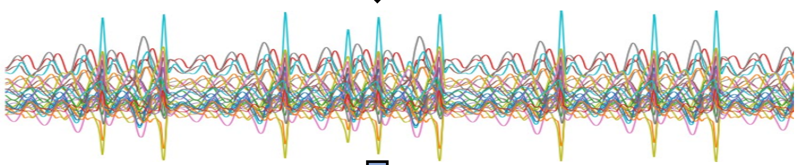


Discrete Observations&Experimental Data

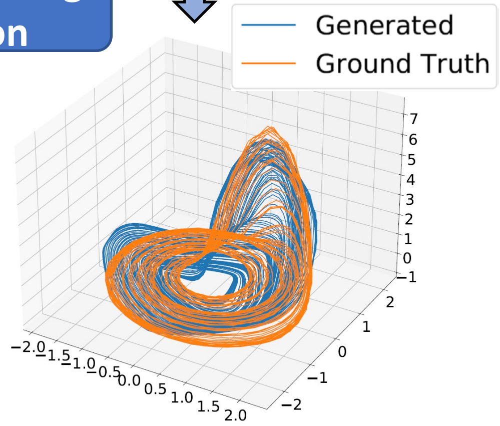
Chaotic Rössler attractor from discrete ordinal observations



Train RNN model



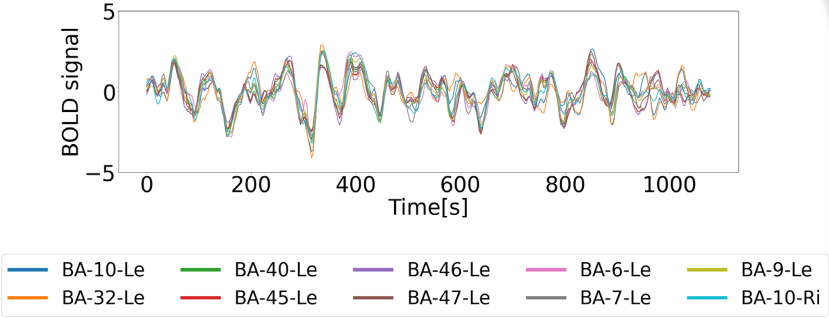
Geometry preserving transformation



Neurophysiological Data+Behavioral Data

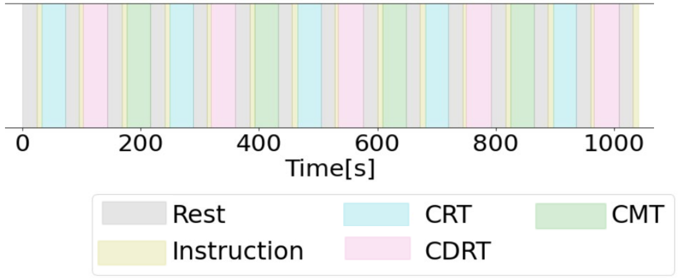
Gaussian Data

$$\boldsymbol{x}_t | \boldsymbol{z}_t \sim \mathcal{N}(\mathbf{B}\boldsymbol{z}_t, \boldsymbol{\Gamma})$$

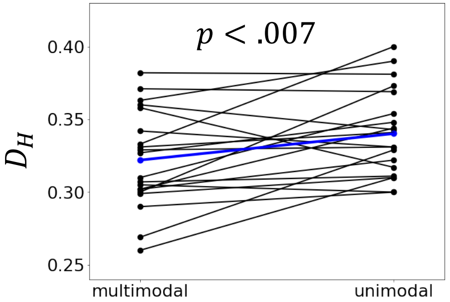


Categorical Data

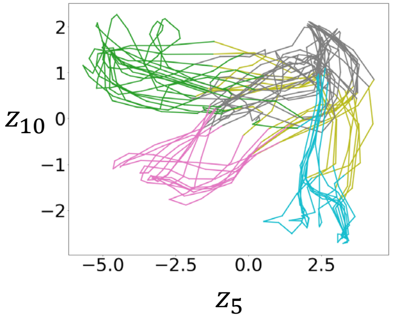
$$\boldsymbol{c}_t | \boldsymbol{z}_t \sim Cat(\boldsymbol{z}_t, \boldsymbol{\beta})$$



Improved DSR



Structured latent activity



Competitive Performance

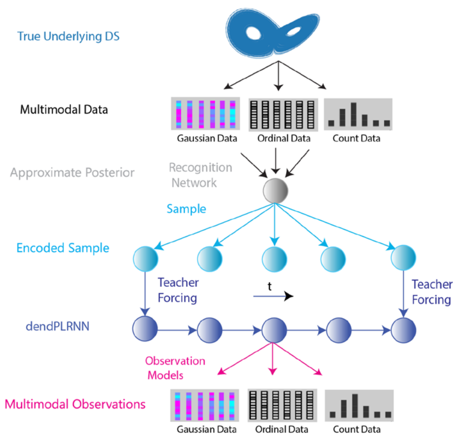
Dataset	Method	$D_{stsp}\downarrow$	$D_H\downarrow$	PE$\downarrow$	OPE$\downarrow$	SCC$\downarrow$	OACF$\downarrow$	CACF$\downarrow$
Lorenz	MVAE-TF	$\mathbf{3.4\pm0.35}$	$\mathbf{0.30\pm0.06}$	$\mathbf{1.3e{-}2\pm2e{-}4}$	$\mathbf{0.12\pm0.03}$	$\mathbf{0.07\pm0.01}$	$\mathbf{0.07\pm0.01}$	$\mathbf{6.6e{-}5\pm8.1e{-}6}$
	SVAE	11.1 ± 0.6	0.82 ± 0.05	$6.3e{-}1\pm5.1e{-}2$	0.68 ± 0.03	0.14 ± 0.01	0.18 ± 0.02	$8.5e{-}5\pm1.6e{-}5$
	GVAE-TF	4.3 ± 0.3	0.47 ± 0.07	$3.6e{-}1\pm1.5e{-}3$	X	X	X	X
	BPTT-TF	8.8 ± 1.9	0.86 ± 0.05	$4.4e{-}1\pm2.2e{-}2$	X	X	X	X
	MS	4.5 ± 1.5	0.61 ± 0.08	X	X	0.14 ± 0.04	0.11 ± 0.02	$\mathbf{6.5e{-}5\pm3.8e{-}6}$
Rössler	MVAE-TF	$\mathbf{1.45\pm0.71}$	$\mathbf{0.32\pm0.03}$	$\mathbf{1.9e{-}3\pm7.1e{-}5}$	$\mathbf{0.08\pm0.02}$	$\mathbf{0.04\pm0.004}$	$\mathbf{0.017\pm0.003}$	$\mathbf{6.5e{-}5\pm1.2e{-}5}$
	SVAE	10.7 ± 1.5	0.66 ± 0.05	$1.5e{-}1\pm3.1e{-}2$	0.24 ± 0.02	0.17 ± 0.03	0.13 ± 0.02	$1.1e{-}4\pm1.4e{-}5$
	GVAE-TF	12.1 ± 0.5	0.55 ± 0.04	$4.9e{-}2\pm3.4e{-}3$	X	X	X	X
	BPTT-TF	8.9 ± 1.4	0.64 ± 0.07	$2.8e{-}1\pm1.8e{-}3$	X	X	X	X
	MS	3.99 ± 1.1	0.59 ± 0.04	X	X	0.08 ± 0.04	0.09 ± 0.02	$1.6e{-}4\pm5.9e{-}5$
Lewis-Glass	MVAE-TF	$\mathbf{0.27\pm0.07}$	$\mathbf{0.33\pm0.02}$	$\mathbf{2.1e{-}3\pm7e{-}5}$	$\mathbf{0.11\pm0.01}$	0.12 ± 0.03	$\mathbf{0.05\pm0.02}$	$2.3e{-}4\pm2.0e{-}5$
	SVAE	2.6 ± 0.5	0.52 ± 0.03	$8.0e{-}2\pm4e{-}3$	0.26 ± 0.01	0.4 ± 0.05	0.18 ± 0.03	$7.5e{-}3\pm4.7e{-}3$
	GVAE-TF	$\mathbf{0.28\pm0.08}$	0.44 ± 0.02	$4.6e{-}3\pm4e{-}4$	X	X	X	X
	BPTT-TF	2.51 ± 0.71	0.43 ± 0.03	$2.6e{-}2\pm3e{-}3$	X	X	X	X
	MS	$\mathbf{0.33\pm0.06}$	$\mathbf{0.35\pm0.01}$	X	X	$\mathbf{0.08\pm0.01}$	$\mathbf{0.04\pm0.002}$	$\mathbf{1.9e{-}4\pm7.5e{-}6}$

Neurophysiological Data

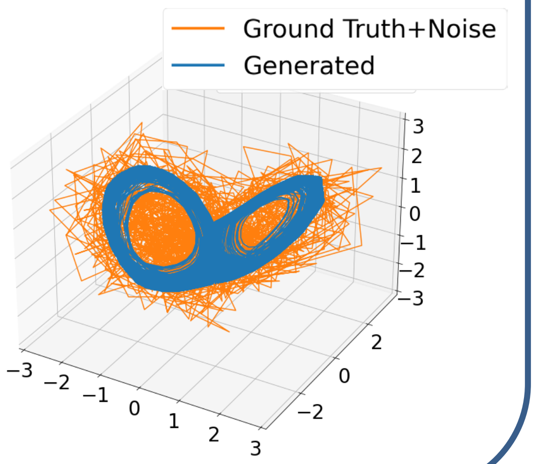
Dataset	Method	$D_{stsp}\downarrow$	$D_H\downarrow$	PE$\downarrow$
fMRI	MVAE-TF	$\mathbf{0.55\pm0.04}$	$\mathbf{0.301\pm0.007}$	$\mathbf{1.21\pm0.08}$
	SVAE	1.9 ± 0.22	0.441 ± 0.019	2.34 ± 0.12
	GVAE-TF	0.67 ± 0.06	0.335 ± 0.011	1.64 ± 0.07
	BPTT-TF	0.63 ± 0.03	0.312 ± 0.006	1.39 ± 0.05
	MS	1.06 ± 0.14	0.373 ± 0.012	X

Conclusions

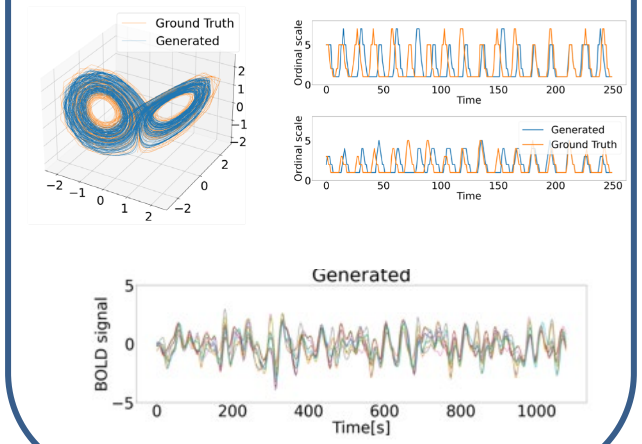
Flexible training framework for DSR from multimodal time series combining sparse TF with MVAEs



DSR from highly noisy and and non-Gaussian obs. from chaotic systems



State of the art performance on benchmark & experimental datasets



STRUCTURES CLUSTER OF EXCELLENCE

Deutsche Forschungsgemeinschaft DFG

IMMERSE

This work was funded by the European Union's Horizon 2020 research and innovation Programme under grant agreement 945263 (IMMERSE), by the German Research Foundation (DFG) within the collaborative research center TRR-265 (project A06), and by a living lab grant by the Federal Ministry of Science, Education and Culture (MWK) of the state of Baden-Württemberg, Germany (grant number 31-7547.223-7/3/2), to DD and GK.