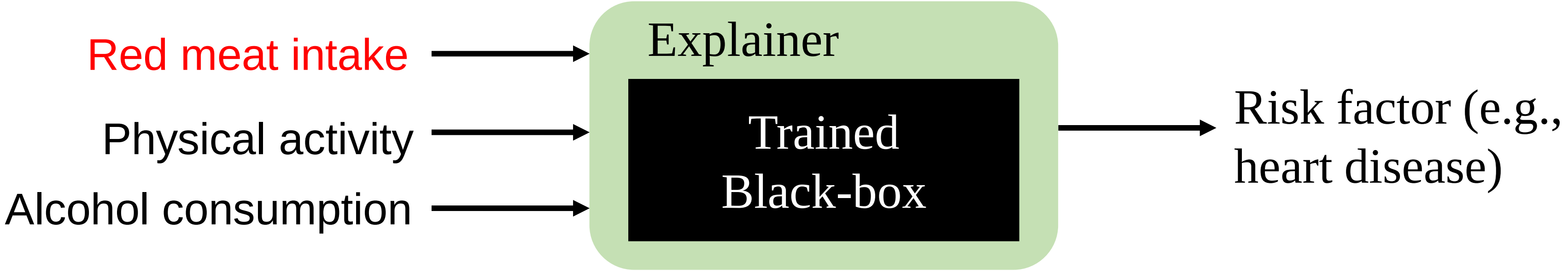
Explaining Time Series via Contrastive and Locally Sparse Perturbations

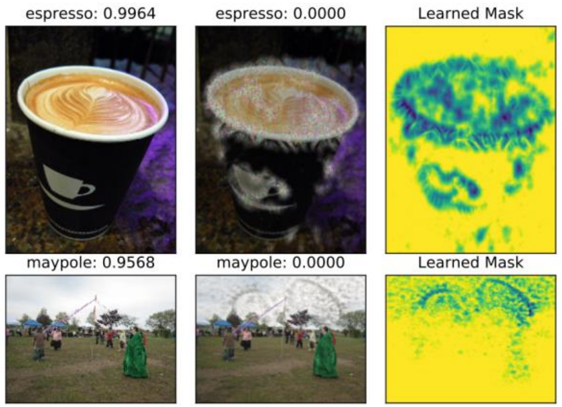
Zichuan Liu[1,2] Yingying Zhang[2] Tianchun Wang[3] Zefan Wang[2,4] Dongsheng Luo[5] Mengnan Du[6] Min Wu[7] Yi Wang[8] Chunlin Chen[1] Lunting Fan[2] Qingsong Wen[2]

[1]Nanjing University, [2]Ailibaba Group,
[3]Pennsylvania State University, [4]Tsinghua University,
[5]Florida International University,
[6]New Jersey Institute of Technology,
[7]A*STAR, [8]The University of Hong Kong

Background

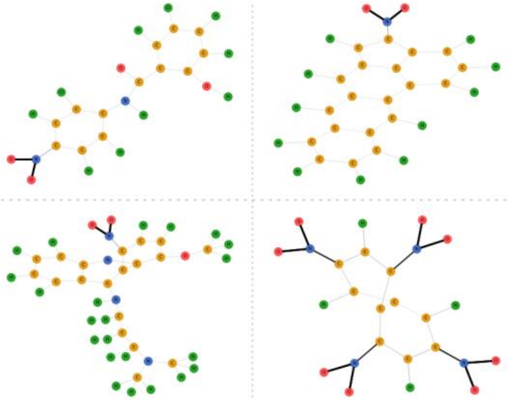
Black-box models with post-hoc explanation techniques: ***Find salient features*!**





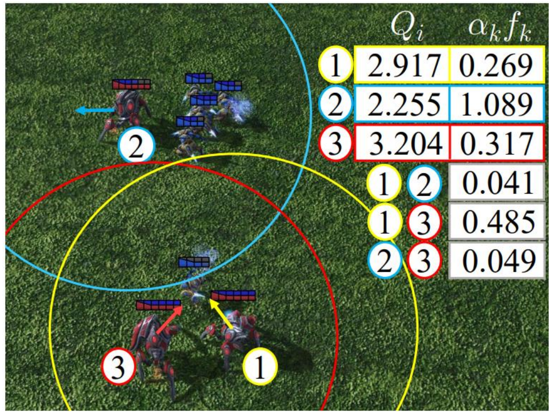
Visual Explanation

Source: Fong et al.



Graph Explanation

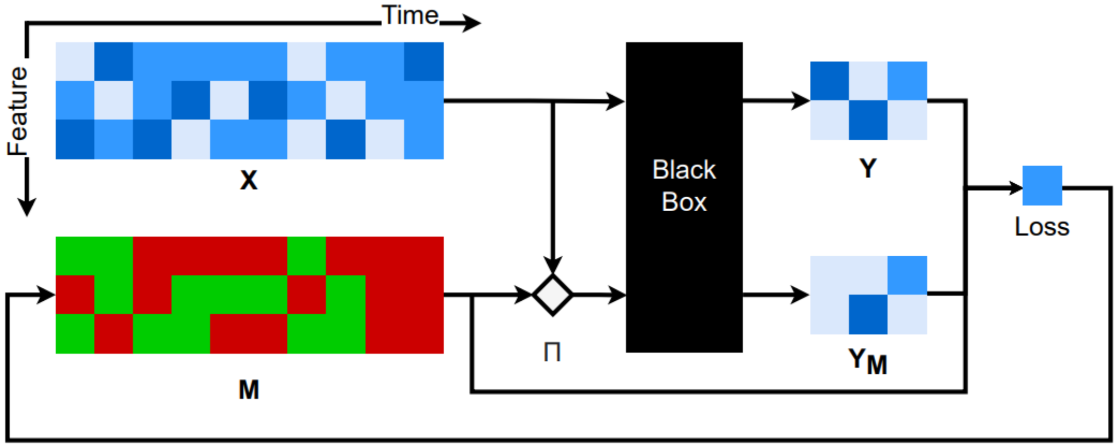
Source: Miao et al.



Game Explanation

Source: Liu et al.

Challenges for Explaining Time Series



Dynamask, Crabbe´ et al.

$$\Phi(x, m) = m \times x + (1 - m) \times u$$

$$\arg\min \underbrace{\mathcal{L}(f(x), f \circ \Phi(x, m))}_{\text{label consistency}} + \underbrace{\mathcal{R}(m)}_{\text{regular}} + \underbrace{\mathcal{A}(m)}_{\text{smooth}}$$

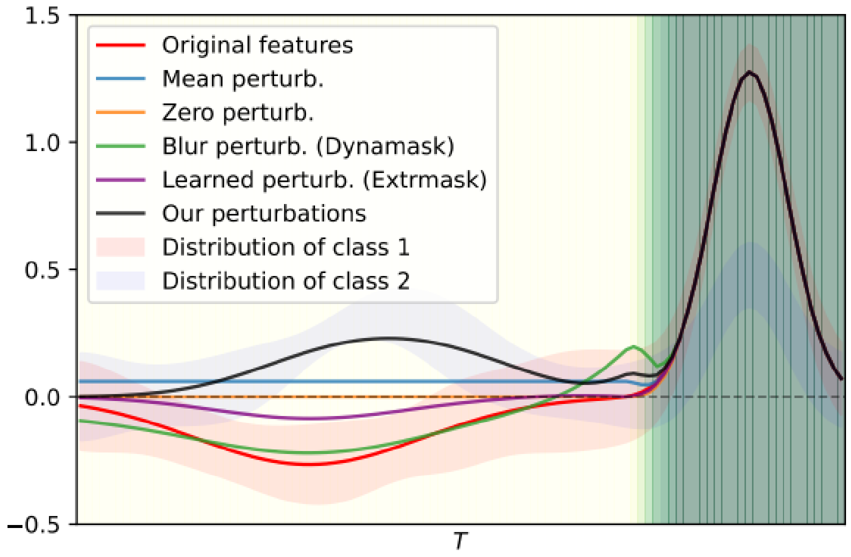
- **Fail to interpret visually**
  - Dense salient features (unlike the image and text)
  - Noisy samples in time series
- **Hard find temporal pattenrns**
  - The time series is smoothed
- **Perturbations matter**
  - Setting a more uninformative values is important
  - Give only instance-based explanations

Existing Perturbations are Inadequate

$$\Phi(x, m) = m \times x + (1 - m) \times u$$

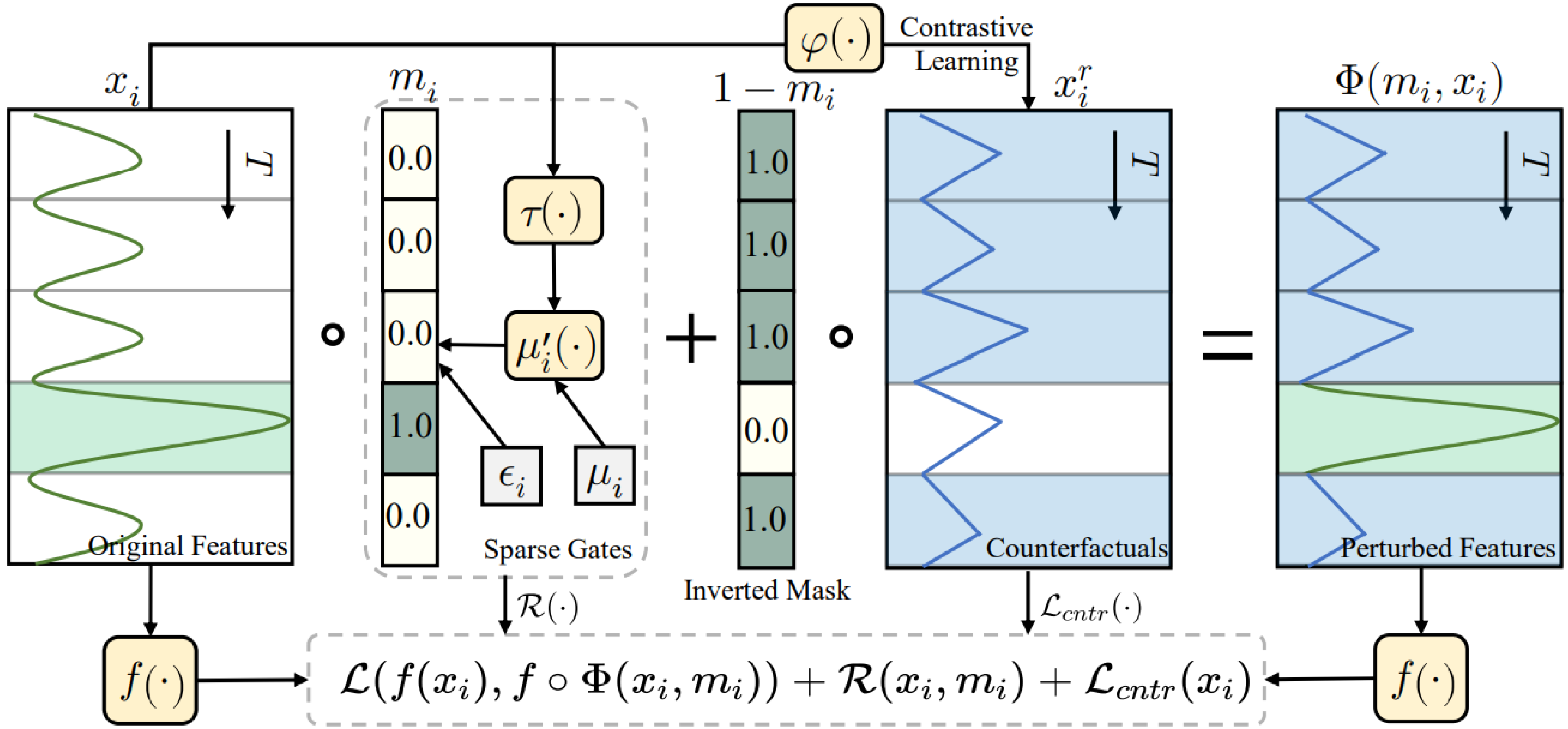
$$\text{where} \quad u = \begin{cases} 0 \\ \frac{1}{w+1} \sum_{t-w}^{t} x_i \\ \text{Gaussian blur} \\ \text{NN}(x) \\ \dots \end{cases}$$

- ➢ Those perturbations may ***out of distribution*** or ***label leakage***
- ➢ Cannot relate temporal patterns ***across samples***



Illustrating different styles of perturbation. Other perturbations could be either not uninformative or not in-domain, while ours is counterfactual that is toward the distribution of negative samples.

ContraLSP Architecture



Perturbation: $\Phi(x, m) = m \times x + (1 - m) \times \varphi_{cntr}(x)$

How to learn the ***uninformative*** $\varphi_{cntr}(x)$ and ***sparse mask m***?

Two Main Contributions (1)

- **Learning counterfactuals from contrastive loss**

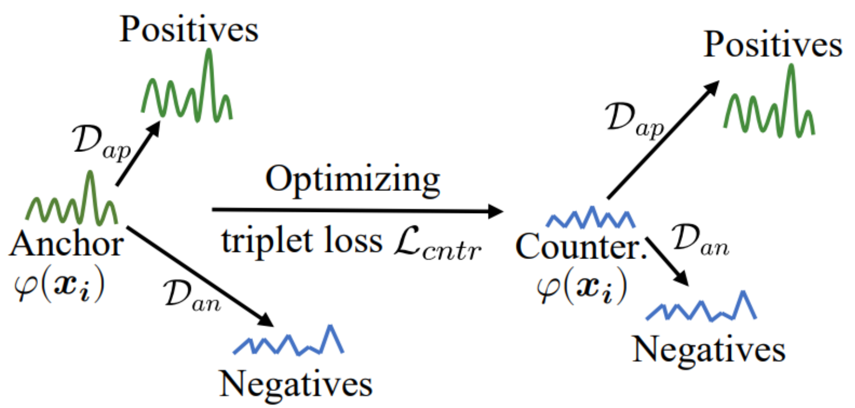
  - Step1: Find positive and negative samples

$$\left(\boldsymbol{x}_i^r, \{\boldsymbol{x}_{i,k}^{r^+}\}_{k=1}^{K^+}, \{\boldsymbol{x}_{i,k}^{r^-}\}_{k=1}^{K^-}\right)$$

Where
$$\begin{cases} \mathcal{D}_{an} = \frac{1}{K^-}\sum_{k=1}^{K^-} |\boldsymbol{x}_i^r - \boldsymbol{x}_{i,k}^{r^-}| \\ \mathcal{D}_{ap} = \frac{1}{K^+}\sum_{k=1}^{K^+} |\boldsymbol{x}_i^r - \boldsymbol{x}_{i,k}^{r^+}| \end{cases}$$

  - Step2: Optimizing via Manhattan distance

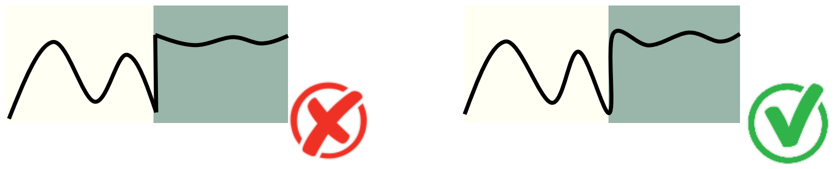
$$\mathcal{L}_{cntr}(\boldsymbol{x}_i) = \max(0, \mathcal{D}_{an} - \mathcal{D}_{ap} - b) + \|\boldsymbol{x}_i^r\|_1 \, ,$$



Learning counterfactuals

Two Main Contributions (2)

- **Learning sparse gates with smooth constraint**



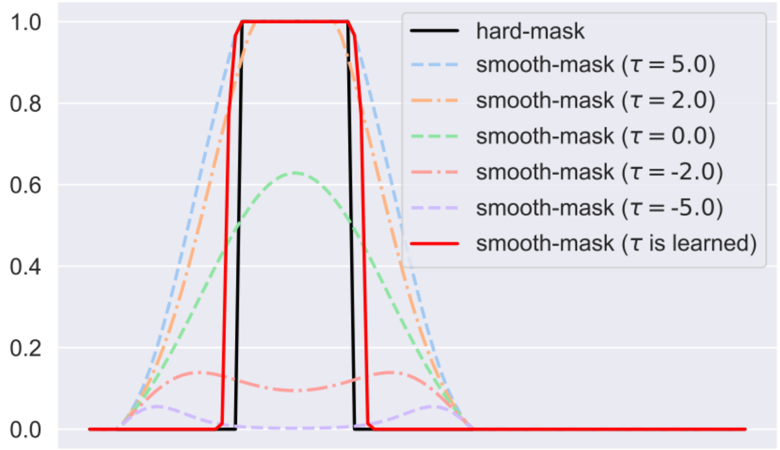
If not smooth, predictor f may error!

- Sparse gates:

$$\boldsymbol{\mu}_i' = \boldsymbol{\mu}_i \odot \sigma(\tau_{\theta_2}(\boldsymbol{x}_i)\boldsymbol{\mu}_i) = \frac{\boldsymbol{\mu}_i}{1 + e^{-\tau_{\theta_2}(\boldsymbol{x}_i)\boldsymbol{\mu}_i}},$$

- L_0-regularization:

$$\mathcal{R}(\boldsymbol{x}_i, \boldsymbol{m}_i) = \|\boldsymbol{m}_i\|_0 = \sum_{t=1}^{T}\sum_{d=1}^{D}\left(\frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{\boldsymbol{\mu}_i'[t,d]}{\sqrt{2}\delta}\right)\right),$$



Binary-skewed masks

Synthetic Experiments (with label)

1. White-box Regression

Table 1: Performance on Rare-Time and Rare-Observation experiments w/o different groups.

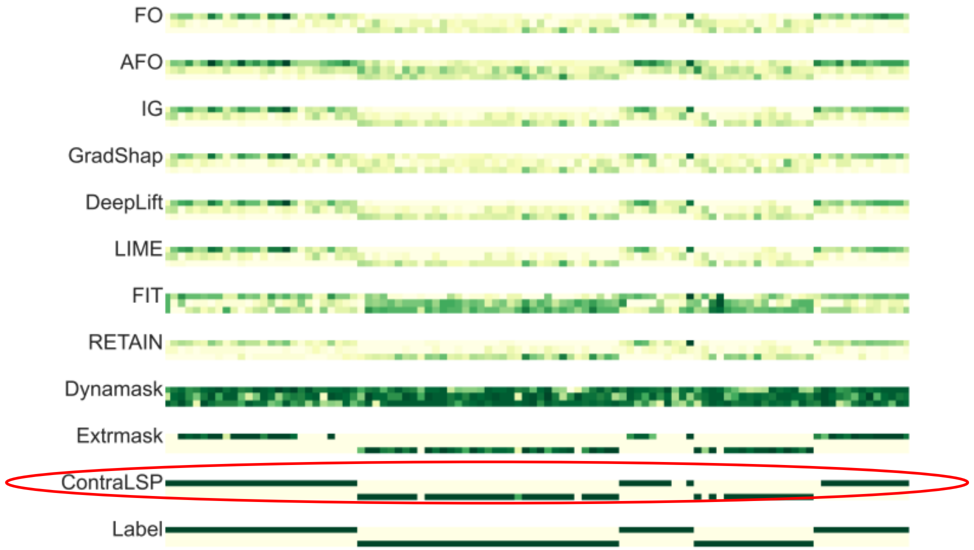
METHOD	RARE-TIME AUP ↑	RARE-TIME AUR ↑	RARE-TIME $I_m/10^4$ ↑	RARE-TIME $S_m/10^2$ ↓	RARE-TIME (DIFFGROUPS) AUP ↑	RARE-TIME (DIFFGROUPS) AUR ↑	RARE-TIME (DIFFGROUPS) $I_m/10^4$ ↑	RARE-TIME (DIFFGROUPS) $S_m/10^2$ ↓
FO	$\mathbf{1.00}_{\pm0.00}$	$0.13_{\pm0.00}$	$0.46_{\pm0.01}$	$47.20_{\pm0.61}$	$\mathbf{1.00}_{\pm0.00}$	$0.16_{\pm0.00}$	$0.53_{\pm0.01}$	$54.89_{\pm0.70}$
AFO	$\mathbf{1.00}_{\pm0.00}$	$0.15_{\pm0.01}$	$0.51_{\pm0.01}$	$55.60_{\pm0.85}$	$\mathbf{1.00}_{\pm0.00}$	$0.16_{\pm0.00}$	$0.54_{\pm0.01}$	$57.76_{\pm0.72}$
IG	$\mathbf{1.00}_{\pm0.00}$	$0.13_{\pm0.00}$	$0.46_{\pm0.01}$	$47.61_{\pm0.62}$	$\mathbf{1.00}_{\pm0.00}$	$0.15_{\pm0.00}$	$0.53_{\pm0.01}$	$54.62_{\pm0.85}$
SVS	$\mathbf{1.00}_{\pm0.00}$	$0.13_{\pm0.00}$	$0.47_{\pm0.01}$	$47.20_{\pm0.61}$	$\mathbf{1.00}_{\pm0.00}$	$0.15_{\pm0.00}$	$0.52_{\pm0.02}$	$54.28_{\pm0.84}$
DYNAMASK	$\underline{0.99}_{\pm0.01}$	$0.67_{\pm0.02}$	$8.68_{\pm0.11}$	$37.24_{\pm0.48}$	$\underline{0.99}_{\pm0.01}$	$0.51_{\pm0.00}$	$5.75_{\pm0.13}$	$47.33_{\pm1.02}$
EXTRMASK	$\mathbf{1.00}_{\pm0.00}$	$\underline{0.88}_{\pm0.00}$	$\underline{16.40}_{\pm0.13}$	$\underline{13.10}_{\pm0.78}$	$\mathbf{1.00}_{\pm0.00}$	$\underline{0.83}_{\pm0.03}$	$\underline{13.37}_{\pm0.78}$	$\underline{27.44}_{\pm3.68}$
CONTRALSP	$\mathbf{1.00}_{\pm0.00}$	$\mathbf{0.97}_{\pm0.01}$	$\mathbf{19.51}_{\pm0.30}$	$\mathbf{4.65}_{\pm0.71}$	$\mathbf{1.00}_{\pm0.00}$	$\mathbf{0.94}_{\pm0.01}$	$\mathbf{18.92}_{\pm0.37}$	$\mathbf{4.40}_{\pm0.60}$

METHOD	RARE-OBSERVATION AUP ↑	RARE-OBSERVATION AUR ↑	RARE-OBSERVATION $I_m/10^4$ ↑	RARE-OBSERVATION $S_m/10^2$ ↓	RARE-OBSERVATION (DIFFGROUPS) AUP ↑	RARE-OBSERVATION (DIFFGROUPS) AUR ↑	RARE-OBSERVATION (DIFFGROUPS) $I_m/10^4$ ↑	RARE-OBSERVATION (DIFFGROUPS) $S_m/10^2$ ↓
FO	$\mathbf{1.00}_{\pm0.00}$	$0.13_{\pm0.00}$	$0.46_{\pm0.00}$	$47.39_{\pm0.16}$	$\mathbf{1.00}_{\pm0.00}$	$0.14_{\pm0.00}$	$0.50_{\pm0.01}$	$52.13_{\pm0.96}$
AFO	$\mathbf{1.00}_{\pm0.00}$	$0.16_{\pm0.00}$	$0.55_{\pm0.01}$	$56.81_{\pm0.39}$	$\mathbf{1.00}_{\pm0.00}$	$0.16_{\pm0.01}$	$0.54_{\pm0.02}$	$56.92_{\pm1.24}$
IG	$\mathbf{1.00}_{\pm0.00}$	$0.13_{\pm0.00}$	$0.46_{\pm0.00}$	$47.82_{\pm0.15}$	$\mathbf{1.00}_{\pm0.00}$	$0.13_{\pm0.00}$	$0.47_{\pm0.00}$	$49.90_{\pm0.88}$
SVS	$\mathbf{1.00}_{\pm0.00}$	$0.13_{\pm0.00}$	$0.46_{\pm0.00}$	$47.39_{\pm0.16}$	$\mathbf{1.00}_{\pm0.00}$	$0.13_{\pm0.00}$	$0.47_{\pm0.01}$	$49.53_{\pm0.84}$
DYNAMASK	$\underline{0.97}_{\pm0.00}$	$0.65_{\pm0.00}$	$8.32_{\pm0.06}$	$22.87_{\pm0.58}$	$\underline{0.98}_{\pm0.00}$	$0.52_{\pm0.01}$	$6.12_{\pm0.10}$	$\underline{30.88}_{\pm0.70}$
EXTRMASK	$\mathbf{1.00}_{\pm0.00}$	$\underline{0.76}_{\pm0.00}$	$\underline{13.25}_{\pm0.07}$	$\underline{9.55}_{\pm0.39}$	$\mathbf{1.00}_{\pm0.00}$	$\underline{0.70}_{\pm0.04}$	$\underline{10.40}_{\pm0.54}$	$32.81_{\pm0.88}$
CONTRALSP	$\mathbf{1.00}_{\pm0.00}$	$\mathbf{1.00}_{\pm0.00}$	$\mathbf{20.68}_{\pm0.03}$	$\mathbf{0.32}_{\pm0.16}$	$\mathbf{1.00}_{\pm0.00}$	$\mathbf{0.99}_{\pm0.00}$	$\mathbf{20.51}_{\pm0.07}$	$\mathbf{0.57}_{\pm0.20}$

2. Black-box Classification

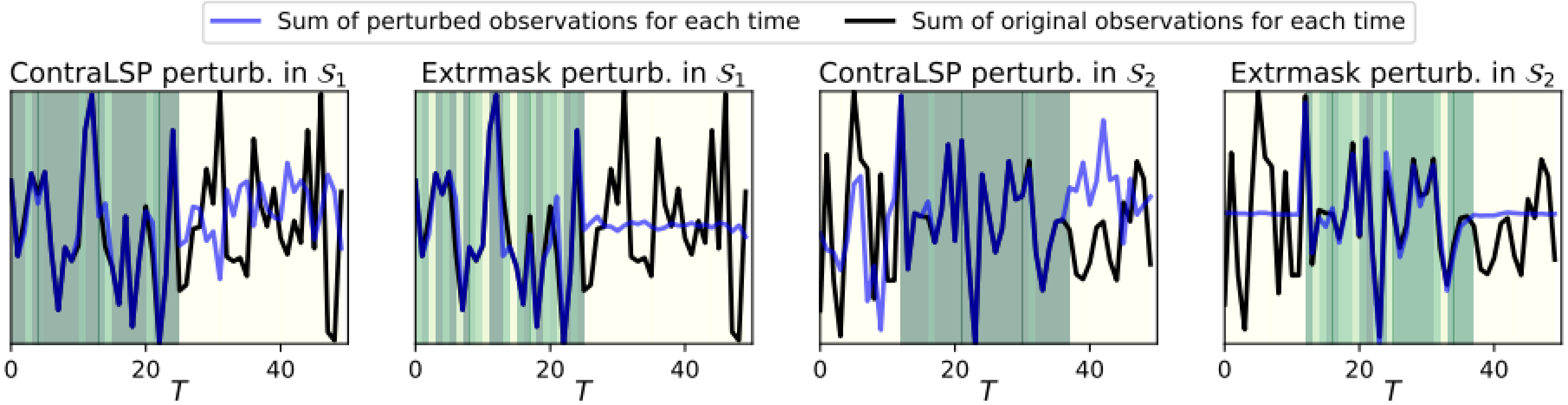
Table 2: Performance on Switch Feature and State data.

METHOD	SWITCH-FEATURE AUP ↑	SWITCH-FEATURE AUR ↑	SWITCH-FEATURE $I_m/10^4$ ↑	SWITCH-FEATURE $S_m/10^3$ ↓	STATE AUP ↑	STATE AUR ↑	STATE $I_m/10^4$ ↑	STATE $S_m/10^3$ ↓
FO	$0.89_{\pm0.03}$	$0.37_{\pm0.02}$	$1.86_{\pm0.14}$	$15.60_{\pm0.28}$	$0.90_{\pm0.05}$	$0.30_{\pm0.01}$	$2.73_{\pm0.15}$	$28.07_{\pm0.54}$
AFO	$0.82_{\pm0.06}$	$0.41_{\pm0.02}$	$2.00_{\pm0.14}$	$17.32_{\pm0.29}$	$0.84_{\pm0.08}$	$0.36_{\pm0.03}$	$3.16_{\pm0.27}$	$34.03_{\pm1.10}$
IG	$0.91_{\pm0.02}$	$0.44_{\pm0.03}$	$2.21_{\pm0.17}$	$16.87_{\pm0.52}$	$\underline{0.93}_{\pm0.02}$	$0.34_{\pm0.03}$	$3.17_{\pm0.28}$	$30.19_{\pm1.22}$
GRADSHAP	$0.88_{\pm0.02}$	$0.38_{\pm0.02}$	$1.92_{\pm0.13}$	$15.85_{\pm0.40}$	$0.88_{\pm0.06}$	$0.30_{\pm0.02}$	$2.76_{\pm0.20}$	$28.18_{\pm0.96}$
DEEPLIFT	$0.91_{\pm0.02}$	$0.44_{\pm0.02}$	$2.23_{\pm0.16}$	$16.86_{\pm0.52}$	$\underline{0.93}_{\pm0.02}$	$0.35_{\pm0.03}$	$3.20_{\pm0.27}$	$30.21_{\pm1.19}$
LIME	$0.94_{\pm0.02}$	$0.40_{\pm0.02}$	$2.01_{\pm0.13}$	$16.09_{\pm0.58}$	$\mathbf{0.95}_{\pm0.02}$	$0.32_{\pm0.03}$	$2.94_{\pm0.26}$	$28.55_{\pm1.53}$
FIT	$0.48_{\pm0.03}$	$0.43_{\pm0.02}$	$1.99_{\pm0.11}$	$17.16_{\pm0.50}$	$0.45_{\pm0.02}$	$0.59_{\pm0.02}$	$7.92_{\pm0.40}$	$33.59_{\pm0.17}$
RETAIN	$0.93_{\pm0.01}$	$0.33_{\pm0.04}$	$1.54_{\pm0.20}$	$15.08_{\pm1.13}$	$0.52_{\pm0.16}$	$0.21_{\pm0.02}$	$1.56_{\pm0.24}$	$25.01_{\pm0.57}$
DYNAMASK	$0.35_{\pm0.00}$	$\underline{0.77}_{\pm0.02}$	$5.22_{\pm0.26}$	$12.85_{\pm0.53}$	$0.36_{\pm0.01}$	$\underline{0.79}_{\pm0.01}$	$10.59_{\pm0.20}$	$25.11_{\pm0.40}$
EXTRMASK	$\underline{0.97}_{\pm0.01}$	$0.65_{\pm0.05}$	$\underline{8.45}_{\pm0.51}$	$\underline{6.90}_{\pm1.44}$	$0.87_{\pm0.01}$	$0.77_{\pm0.01}$	$\underline{29.71}_{\pm1.39}$	$\underline{7.54}_{\pm0.46}$
CONTRALSP	$\mathbf{0.98}_{\pm0.00}$	$\mathbf{0.80}_{\pm0.03}$	$\mathbf{24.23}_{\pm1.27}$	$\mathbf{0.91}_{\pm0.26}$	$0.90_{\pm0.03}$	$\mathbf{0.81}_{\pm0.01}$	$\mathbf{50.09}_{\pm0.78}$	$\mathbf{0.50}_{\pm0.05}$



Synthetic Experiments (with label)

- Counterfactual information



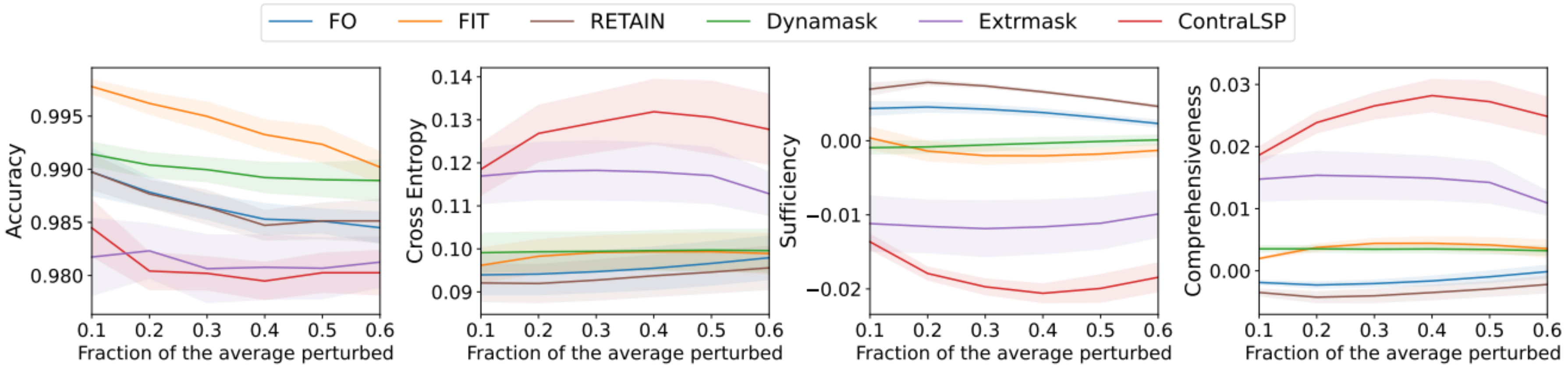
- Distribution analysis of perturbations

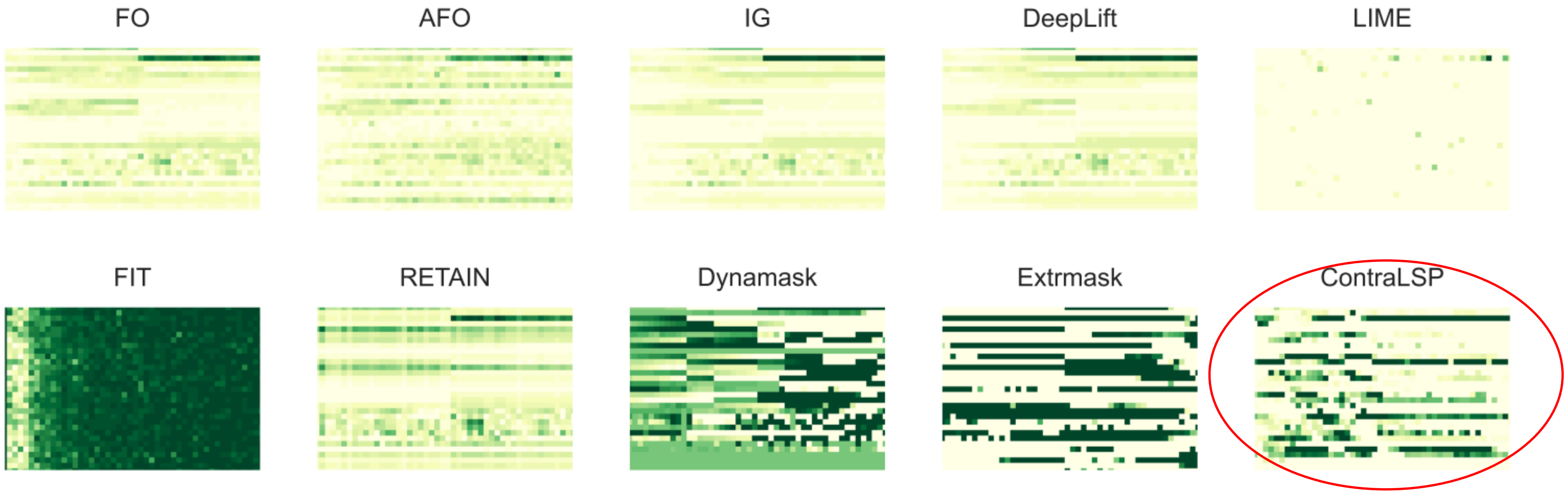
Table 12: Difference between the distribution of different perturbations and the original distribution.

Perturbation type	Rare-Time		Rare-Observation	
	KDE-score ↑	KL-divergence ↓	KDE-score ↑	KL-divergence ↓
Zero perturbation	−25.242	0.0523	−23.377	0.0421
Mean perturbation	−30.805	0.0731	−26.421	0.0589
Extrmask perturbation	−22.532	0.0219	−19.102	0.0104
ContraLSP perturbation	−23.290	0.0393	−22.732	0.0386

Real-world Experiments (without label)

3. MIMIC-III Mortality Data



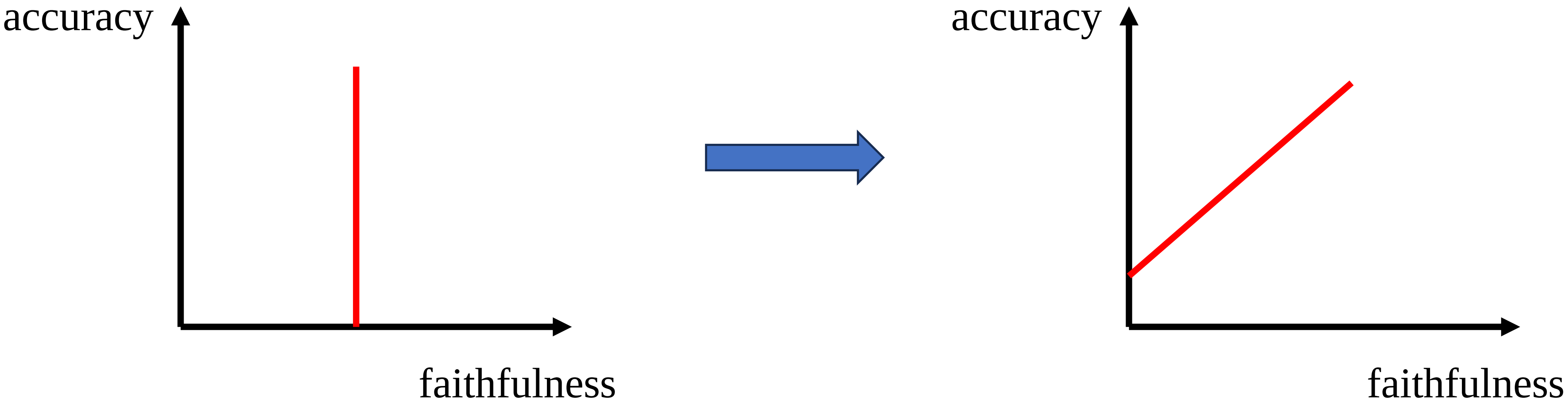


Conclusion

- We propose ContraLSP as a time series explainer, which incorporates counterfactual samples to build uninformative in-domain perturbation.
- We incorporate sample-specific sparse gates to generate more binary-skewed and smooth masks.
- The code is available at https://github.com/zichuan-liu/ContraLSP.

Future Explorations

- How to represent uncertainty when black box models are inaccurate



- Quantification of compression amplitude and parameter tuning strategy

$$\widetilde{\mathcal{L}} = \mathcal{L}_{\mathrm{LC}} + \alpha \mathcal{L}_M + \beta(\mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{dr}),$$

Thanks for your listening!

Any Questions? Please use the chat !