ICLR
International Conference On Learning Representations

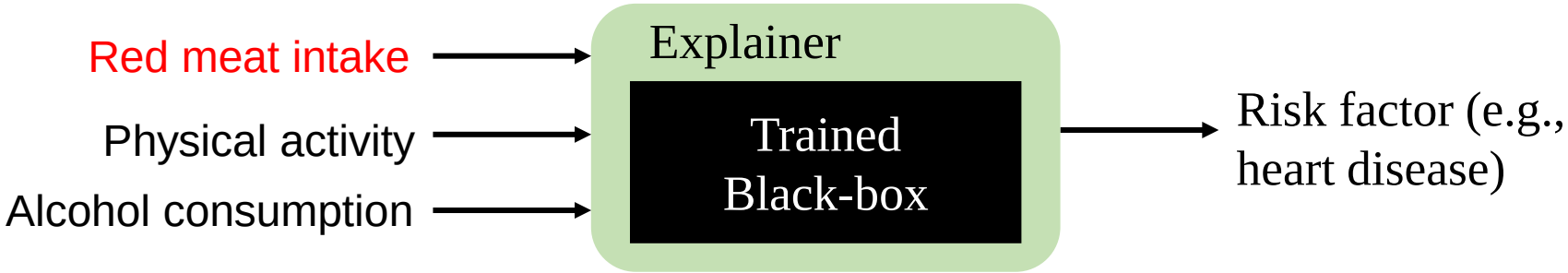
Explaining Time Series via Contrastive and Locally Sparse Perturbations

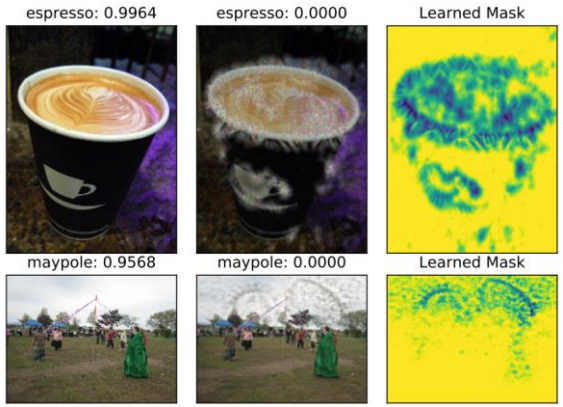
Zichuan Liu[1,2] Yingying Zhang[2] Tianchun Wang[3] Zefan Wang[2,4] Dongsheng Luo[5]
Mengnan Du[6] Min Wu[7] Yi Wang[8] Chunlin Chen[1] Lunting Fan[2] Qingsong Wen[2]

[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

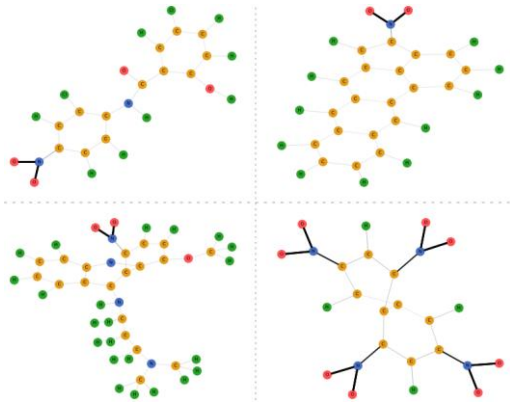
Black-box models with post-hoc explanation techniques: ***Find salient features!***





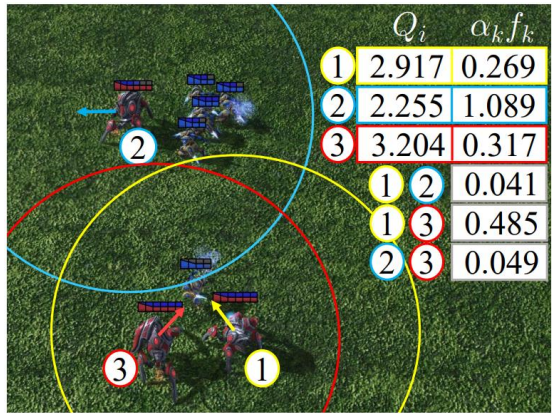
Visual Explanation

Source: Fong et al.



Graph Explanation

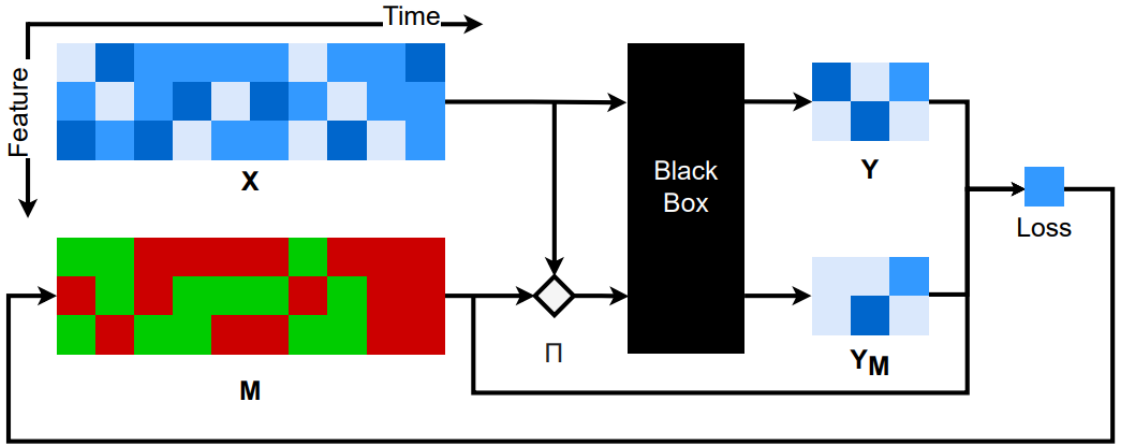
Source: Miao et al.



Game Explanation

Source: Liu et al.

Challenges for Explaning Time Series



Dynamask, [Crabbe´ et al.]

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

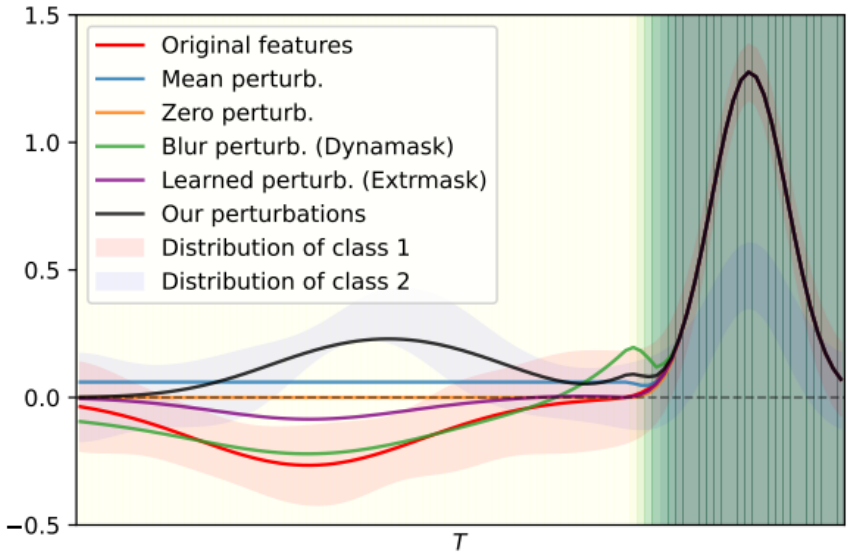
$$\arg\min \underbrace{\mathcal{L}(f(x), f \circ \Phi(x, m))}_{\text{label consistency}} + \underbrace{\mathcal{R}(m)}_{\text{regular}} + \underbrace{\mathcal{A}(m)}_{\text{smooth}}$$

- **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

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

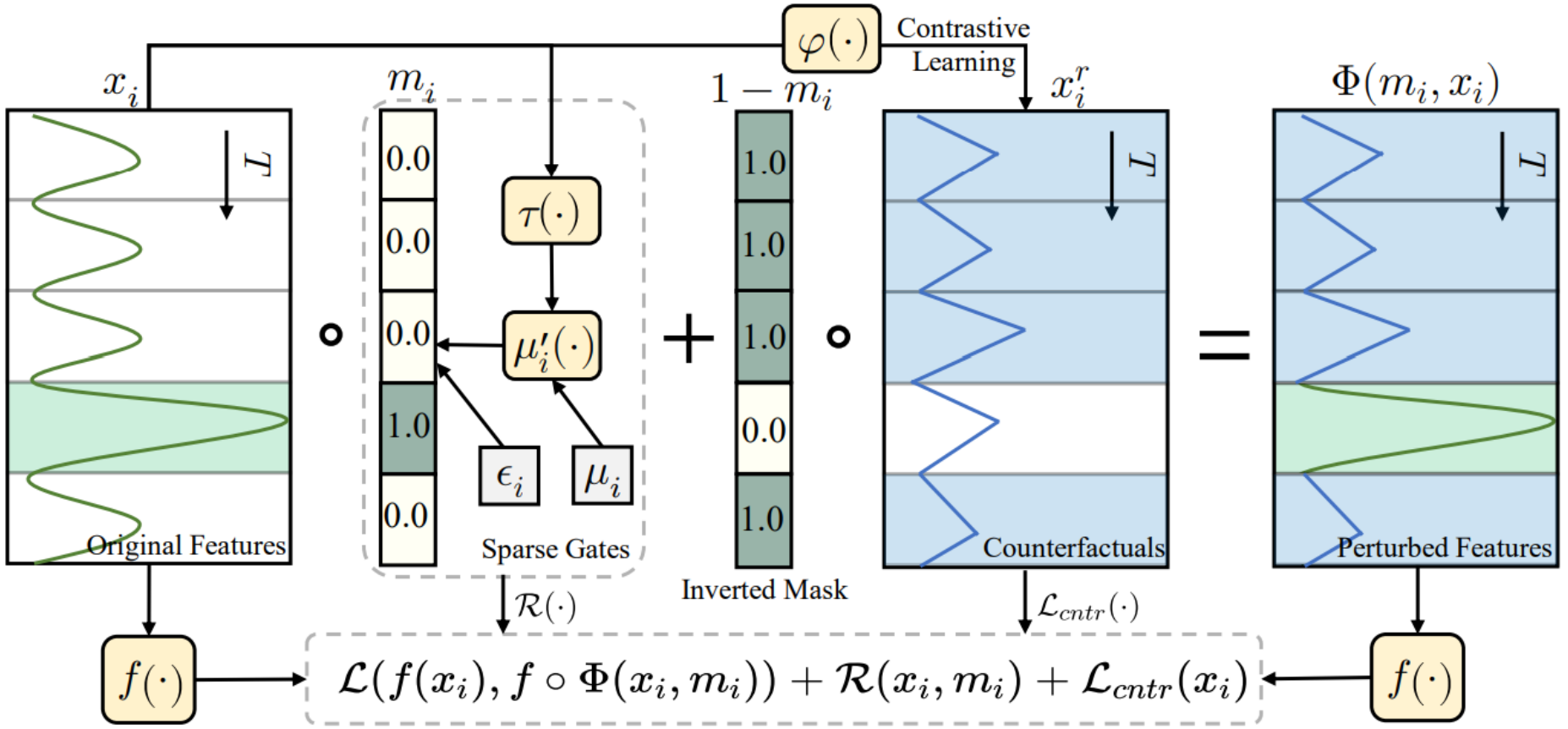
$$\text{where} \quad u = \begin{cases} 0 \\ \frac{1}{w+1} \sum_{t-w}^{t} x_i \\ \text{Gaussian blur} \\ \text{NN}(x) \\ \dots \end{cases}$$



- 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



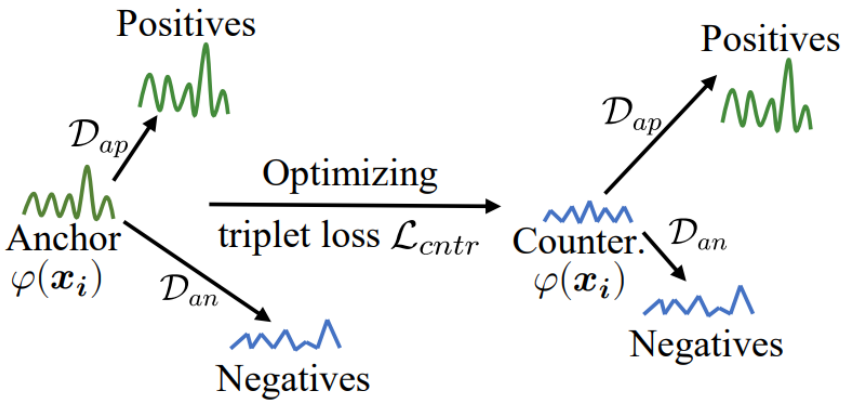
Perturbation: $\Phi(x, m) = m \times x + (1 - m) \times \varphi_{cntr}(x)$

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

Two Main Contributions

- **Learning counterfactuals from contrastive loss**
  - Step1: Find positive and negative samples
  - Step2: Optimizing via Manhattan distance

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

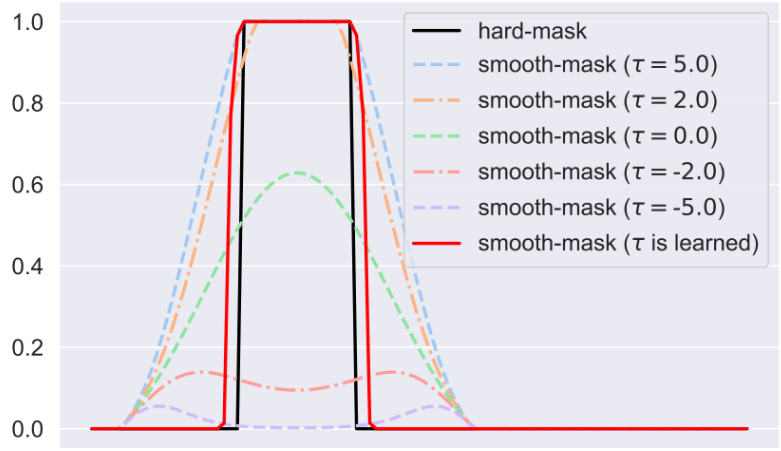


Learning counterfactuals

- **Learning sparse gates with smooth constraint**

If not smooth, predictor f may error!

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



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.

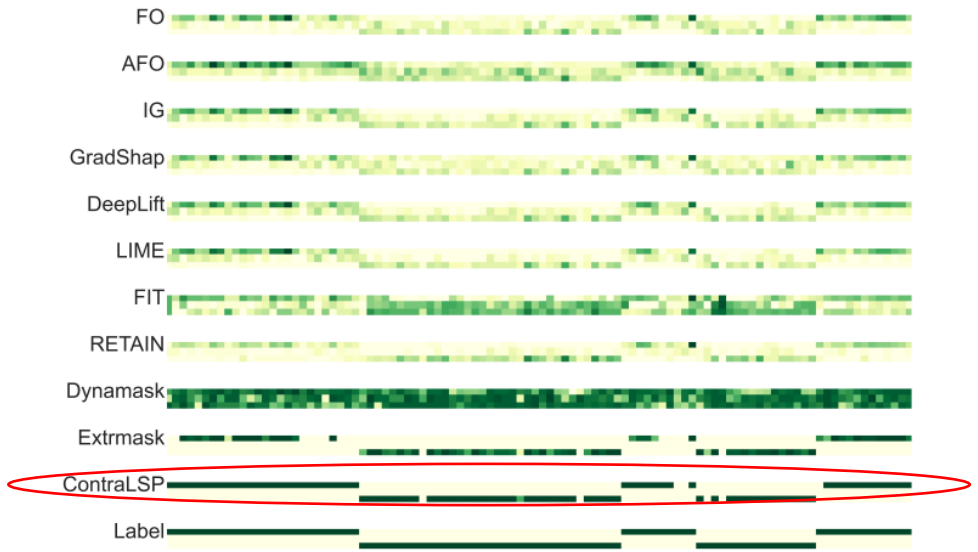
Method	Rare-Time AUP ↑	Rare-Time AUR ↑	Rare-Time $I_m/10^4$ ↑	Rare-Time $S_m/10^2$ ↓	Rare-Time (Diffgroups) AUP ↑	Rare-Time (Diffgroups) AUR ↑	Rare-Time (Diffgroups) $I_m/10^4$ ↑	Rare-Time (Diffgroups) $S_m/10^2$ ↓
FO	**1.00**$_{\pm 0.00}$	0.13$_{\pm 0.00}$	0.46$_{\pm 0.01}$	47.20$_{\pm 0.61}$	**1.00**$_{\pm 0.00}$	0.16$_{\pm 0.00}$	0.53$_{\pm 0.01}$	54.89$_{\pm 0.70}$
AFO	**1.00**$_{\pm 0.00}$	0.15$_{\pm 0.01}$	0.51$_{\pm 0.01}$	55.60$_{\pm 0.85}$	**1.00**$_{\pm 0.00}$	0.16$_{\pm 0.00}$	0.54$_{\pm 0.01}$	57.76$_{\pm 0.72}$
IG	**1.00**$_{\pm 0.00}$	0.13$_{\pm 0.00}$	0.46$_{\pm 0.01}$	47.61$_{\pm 0.62}$	**1.00**$_{\pm 0.00}$	0.15$_{\pm 0.00}$	0.53$_{\pm 0.01}$	54.62$_{\pm 0.85}$
SVS	**1.00**$_{\pm 0.00}$	0.13$_{\pm 0.00}$	0.47$_{\pm 0.01}$	47.20$_{\pm 0.61}$	**1.00**$_{\pm 0.00}$	0.15$_{\pm 0.00}$	0.52$_{\pm 0.02}$	54.28$_{\pm 0.84}$
Dynamask	0.99$_{\pm 0.01}$	0.67$_{\pm 0.02}$	8.68$_{\pm 0.11}$	37.24$_{\pm 0.48}$	0.99$_{\pm 0.01}$	0.51$_{\pm 0.00}$	5.75$_{\pm 0.13}$	47.33$_{\pm 1.02}$
Extrmask	**1.00**$_{\pm 0.00}$	0.88$_{\pm 0.00}$	16.40$_{\pm 0.13}$	13.10$_{\pm 0.78}$	**1.00**$_{\pm 0.00}$	0.83$_{\pm 0.03}$	13.37$_{\pm 0.78}$	27.44$_{\pm 3.68}$
ContraLSP	**1.00**$_{\pm 0.00}$	**0.97**$_{\pm 0.01}$	**19.51**$_{\pm 0.30}$	**4.65**$_{\pm 0.71}$	**1.00**$_{\pm 0.00}$	**0.94**$_{\pm 0.01}$	**18.92**$_{\pm 0.37}$	**4.40**$_{\pm 0.60}$

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	**1.00**$_{\pm 0.00}$	0.13$_{\pm 0.00}$	0.46$_{\pm 0.00}$	47.39$_{\pm 0.16}$	**1.00**$_{\pm 0.00}$	0.14$_{\pm 0.00}$	0.50$_{\pm 0.01}$	52.13$_{\pm 0.96}$
AFO	**1.00**$_{\pm 0.00}$	0.16$_{\pm 0.00}$	0.55$_{\pm 0.01}$	56.81$_{\pm 0.39}$	**1.00**$_{\pm 0.00}$	0.16$_{\pm 0.01}$	0.54$_{\pm 0.02}$	56.92$_{\pm 1.24}$
IG	**1.00**$_{\pm 0.00}$	0.13$_{\pm 0.00}$	0.46$_{\pm 0.00}$	47.82$_{\pm 0.15}$	**1.00**$_{\pm 0.00}$	0.13$_{\pm 0.00}$	0.47$_{\pm 0.00}$	49.90$_{\pm 0.88}$
SVS	**1.00**$_{\pm 0.00}$	0.13$_{\pm 0.00}$	0.46$_{\pm 0.00}$	47.39$_{\pm 0.16}$	**1.00**$_{\pm 0.00}$	0.13$_{\pm 0.00}$	0.47$_{\pm 0.01}$	49.53$_{\pm 0.84}$
Dynamask	0.97$_{\pm 0.00}$	0.65$_{\pm 0.00}$	8.32$_{\pm 0.06}$	22.87$_{\pm 0.58}$	0.98$_{\pm 0.00}$	0.52$_{\pm 0.01}$	6.12$_{\pm 0.10}$	30.88$_{\pm 0.70}$
Extrmask	**1.00**$_{\pm 0.00}$	0.76$_{\pm 0.00}$	13.25$_{\pm 0.07}$	9.55$_{\pm 0.39}$	**1.00**$_{\pm 0.00}$	0.70$_{\pm 0.04}$	10.40$_{\pm 0.54}$	32.81$_{\pm 0.88}$
ContraLSP	**1.00**$_{\pm 0.00}$	**1.00**$_{\pm 0.00}$	**20.68**$_{\pm 0.03}$	**0.32**$_{\pm 0.16}$	**1.00**$_{\pm 0.00}$	**0.99**$_{\pm 0.00}$	**20.51**$_{\pm 0.07}$	**0.57**$_{\pm 0.20}$

2. Black-box Classification

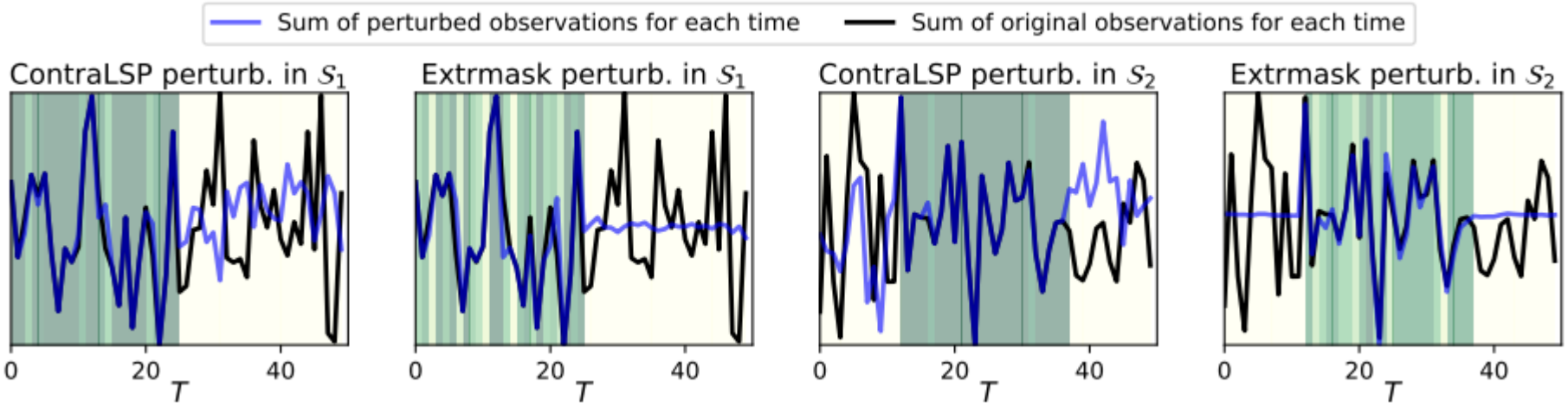
Table 2: Performance on Switch Feature and State data.

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$_{\pm 0.03}$	0.37$_{\pm 0.02}$	1.86$_{\pm 0.14}$	15.60$_{\pm 0.28}$	0.90$_{\pm 0.05}$	0.30$_{\pm 0.01}$	2.73$_{\pm 0.15}$	28.07$_{\pm 0.54}$
AFO	0.82$_{\pm 0.06}$	0.41$_{\pm 0.02}$	2.00$_{\pm 0.14}$	17.32$_{\pm 0.29}$	0.84$_{\pm 0.08}$	0.36$_{\pm 0.03}$	3.16$_{\pm 0.27}$	34.03$_{\pm 1.10}$
IG	0.91$_{\pm 0.02}$	0.44$_{\pm 0.03}$	2.21$_{\pm 0.17}$	16.87$_{\pm 0.52}$	0.93$_{\pm 0.02}$	0.34$_{\pm 0.03}$	3.17$_{\pm 0.28}$	30.19$_{\pm 1.22}$
GradShap	0.88$_{\pm 0.02}$	0.38$_{\pm 0.02}$	1.92$_{\pm 0.13}$	15.85$_{\pm 0.40}$	0.88$_{\pm 0.06}$	0.30$_{\pm 0.02}$	2.76$_{\pm 0.20}$	28.18$_{\pm 0.96}$
DeepLift	0.91$_{\pm 0.02}$	0.44$_{\pm 0.02}$	2.23$_{\pm 0.16}$	16.86$_{\pm 0.52}$	0.93$_{\pm 0.02}$	0.35$_{\pm 0.03}$	3.20$_{\pm 0.27}$	30.21$_{\pm 1.19}$
LIME	0.94$_{\pm 0.02}$	0.40$_{\pm 0.02}$	2.01$_{\pm 0.13}$	16.09$_{\pm 0.58}$	**0.95**$_{\pm 0.02}$	0.32$_{\pm 0.03}$	2.94$_{\pm 0.26}$	28.55$_{\pm 1.53}$
FIT	0.48$_{\pm 0.03}$	0.43$_{\pm 0.02}$	1.99$_{\pm 0.11}$	17.16$_{\pm 0.50}$	0.45$_{\pm 0.02}$	0.59$_{\pm 0.02}$	7.92$_{\pm 0.40}$	33.59$_{\pm 0.17}$
Retain	0.93$_{\pm 0.01}$	0.33$_{\pm 0.04}$	1.54$_{\pm 0.20}$	15.08$_{\pm 1.13}$	0.52$_{\pm 0.16}$	0.21$_{\pm 0.02}$	1.56$_{\pm 0.24}$	25.01$_{\pm 0.57}$
Dynamask	0.35$_{\pm 0.00}$	0.77$_{\pm 0.02}$	5.22$_{\pm 0.26}$	12.85$_{\pm 0.53}$	0.36$_{\pm 0.01}$	0.79$_{\pm 0.01}$	10.59$_{\pm 0.20}$	25.11$_{\pm 0.40}$
Extrmask	0.97$_{\pm 0.01}$	0.65$_{\pm 0.05}$	8.45$_{\pm 0.51}$	6.90$_{\pm 1.44}$	0.87$_{\pm 0.01}$	0.77$_{\pm 0.01}$	29.71$_{\pm 1.39}$	7.54$_{\pm 0.46}$
ContraLSP	**0.98**$_{\pm 0.00}$	**0.80**$_{\pm 0.03}$	**24.23**$_{\pm 1.27}$	**0.91**$_{\pm 0.26}$	0.90$_{\pm 0.03}$	**0.81**$_{\pm 0.01}$	**50.09**$_{\pm 0.78}$	**0.50**$_{\pm 0.05}$



Synthetic Experiments (with label)

- Counterfactual information



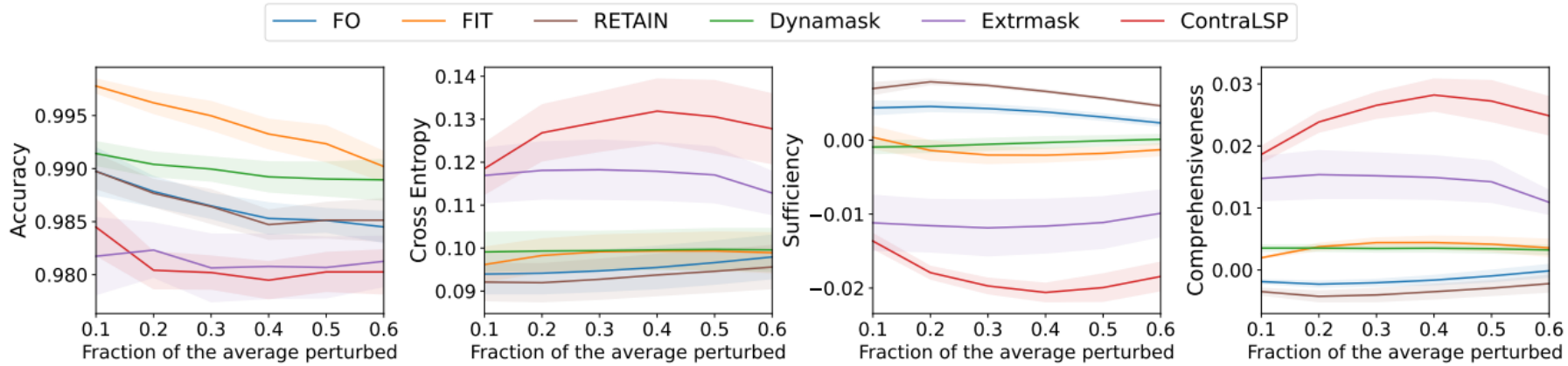
- Distribution analysis of perturbations

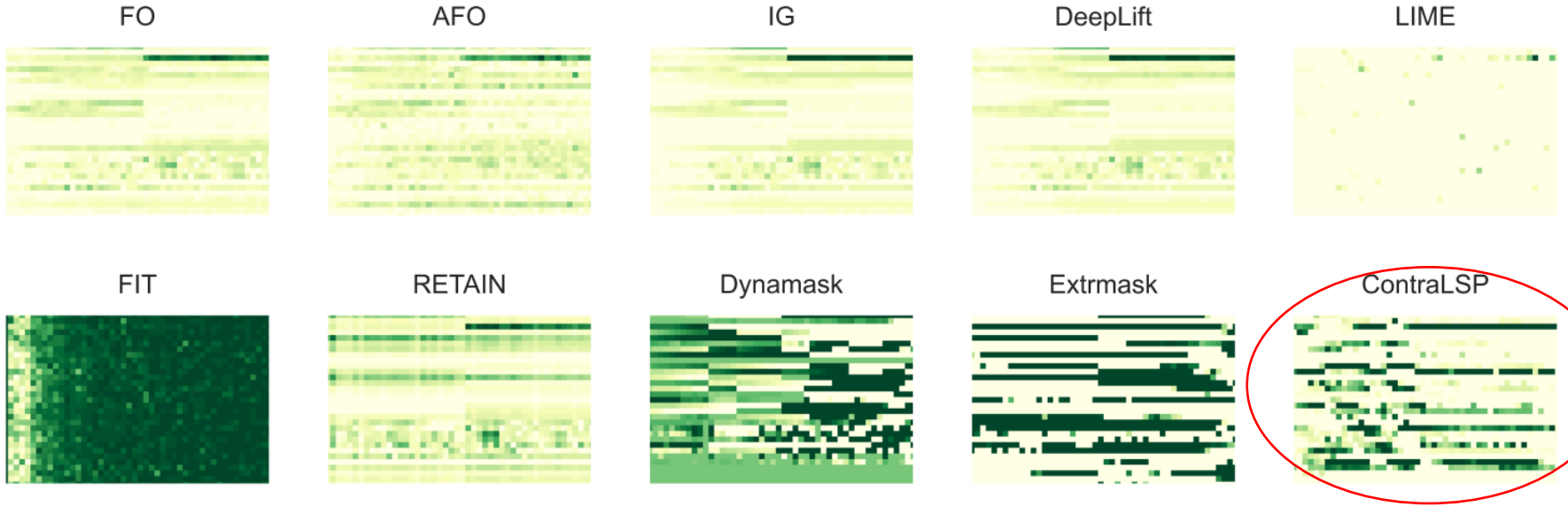
Table 12: Difference between the distribution of different perturbations and the original distribution.

Perturbation type	Rare-Time KDE-score ↑	Rare-Time KL-divergence ↓	Rare-Observation KDE-score ↑	Rare-Observation 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





Closing Remarks

- 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.

Thanks for your listening!