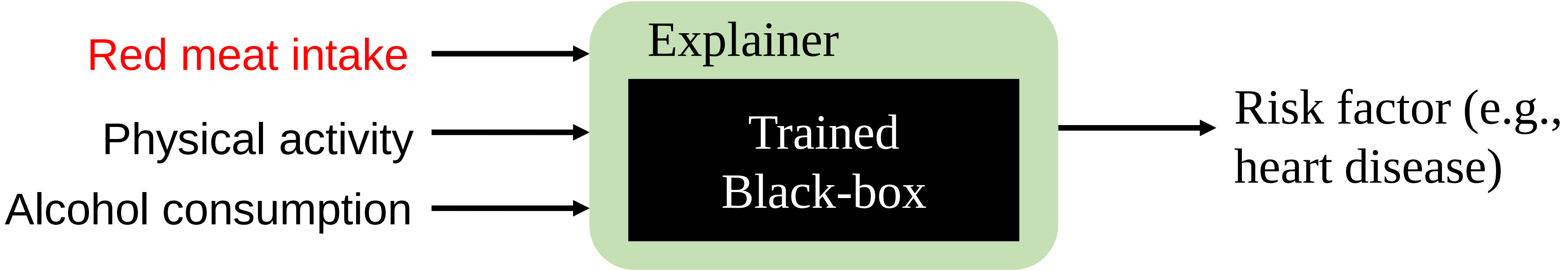
# Explaining Time Series via Contrastive and Locally Sparse Perturbations

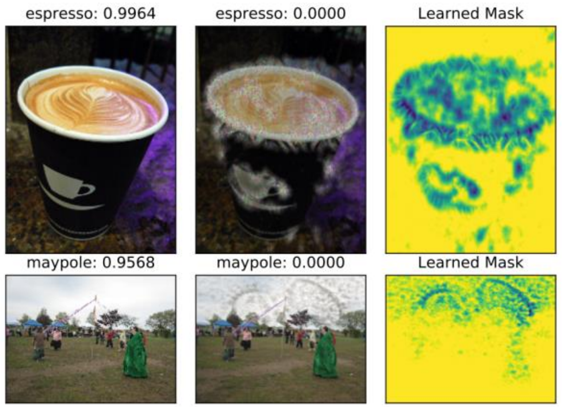

**Zichuan Liu[1,2] Yingying Zhang[2] Tianchun Wang[3] Zefan Wang[2,4] Dongsheng Luo[5]**
**Mengnan Du[6] Min Wu[7] Yi Wang[8] Chunlin Chen[1] Lunting Fan[2] Qingsong Wen[2]**

[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

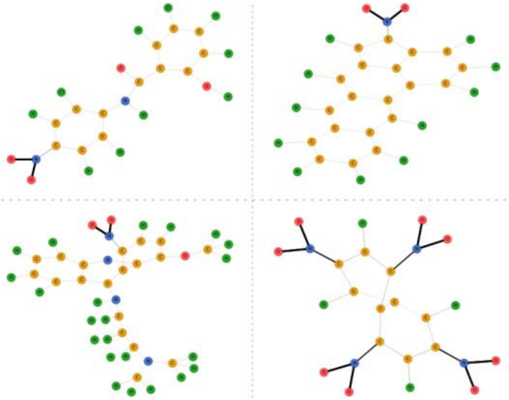
Black-box models with post-hoc explanation techniques: ***Find salient features*!**



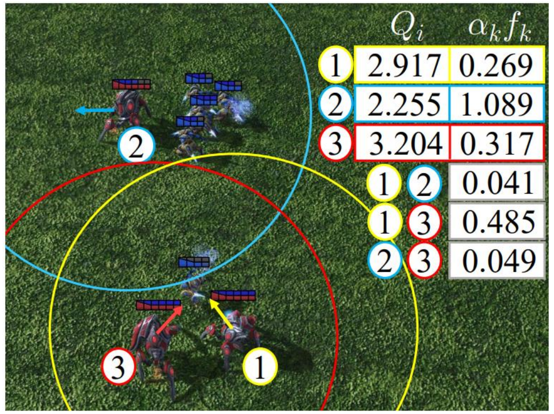

Visual Explanation

Source: Fong et al.



Graph Explanation

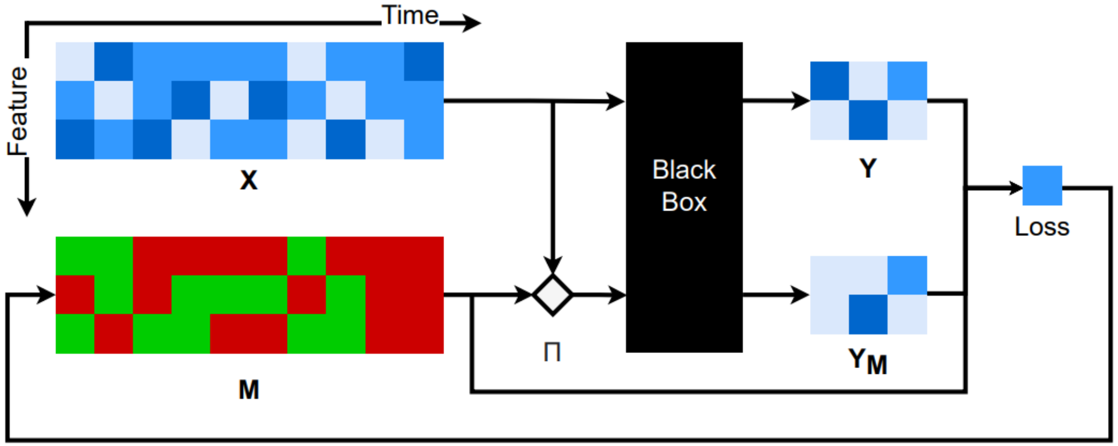
Source: Miao et al.




Game Explanation

Source: Liu et al.

# Challenges for Explaining Time Series




Dynamask, Crabbe´ et al.

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

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

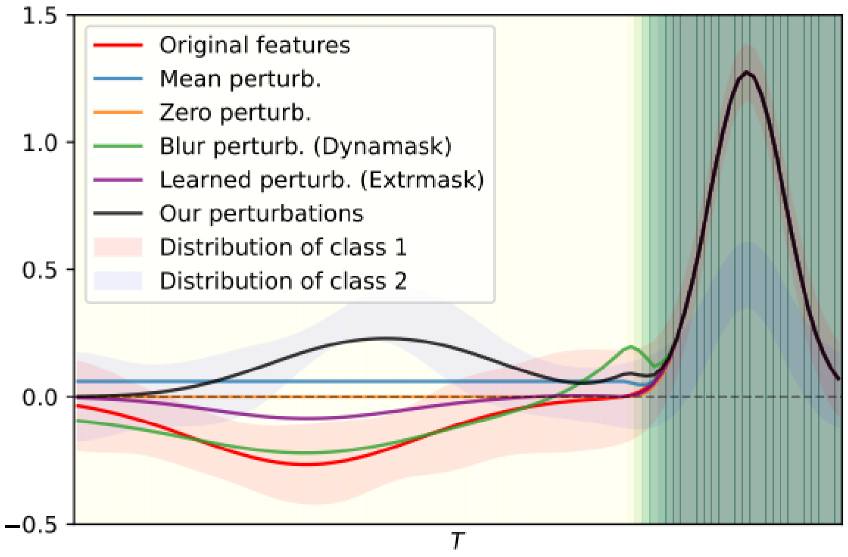
- **Fail to interpret visually**
  - Dense salient features (unlike the image and text)
  - Noisy samples in time series
- **Hard find temporal pattenrns**
  - The time series is smoothed
- **Perturbations matter**
  - Setting a more uninformative values is important
  - Give only instance-based explanations

# Existing Perturbations are Inadequate

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

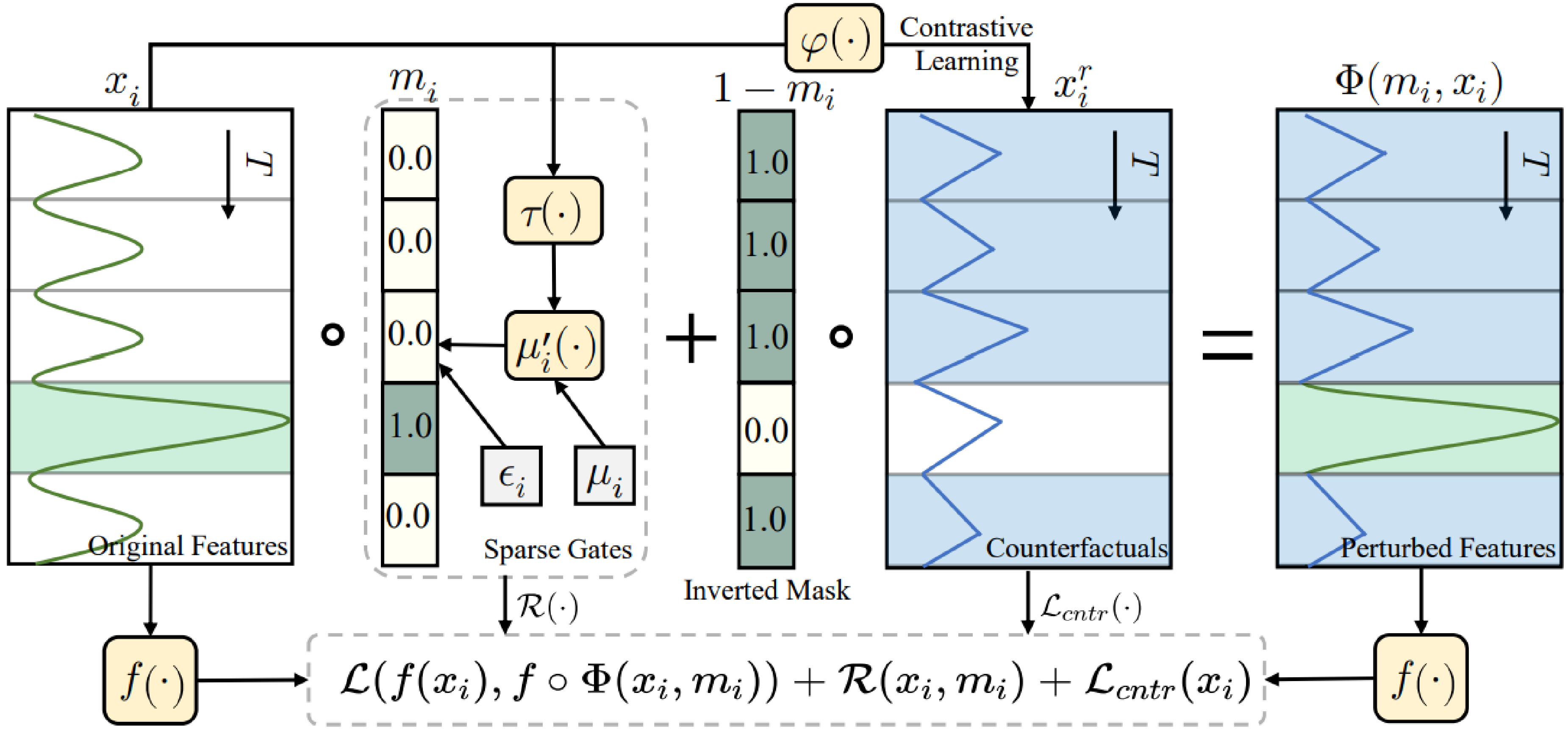
$$\text{where} \quad u = \begin{cases} 0 \\ \frac{1}{w+1}\sum_{t-w}^{t} x_i \\ \text{Gaussian blur} \\ \text{NN}(x) \\ \ldots \end{cases}$$

- ➢ 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



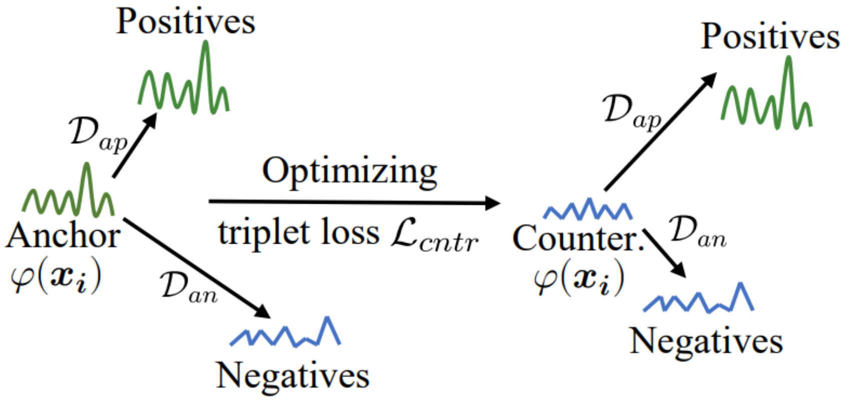

**Perturbation:** $\Phi(x, m) = m \times x + (1 - m) \times \varphi_{cntr}(x)$

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

# Two Main Contributions

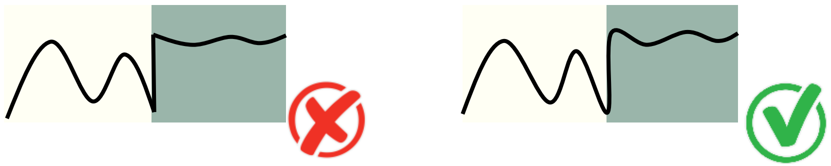
- **Learning counterfactuals from contrastive loss**
  - Step1: Find positive and negative samples
  - Step2: Optimizing via Manhattan distance

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



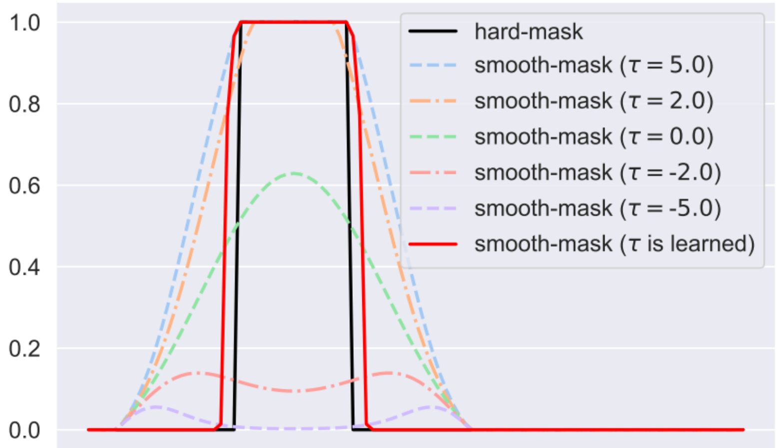

Learning counterfactuals

- **Learning sparse gates with smooth constraint**



If not smooth, predictor f may error!

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




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.

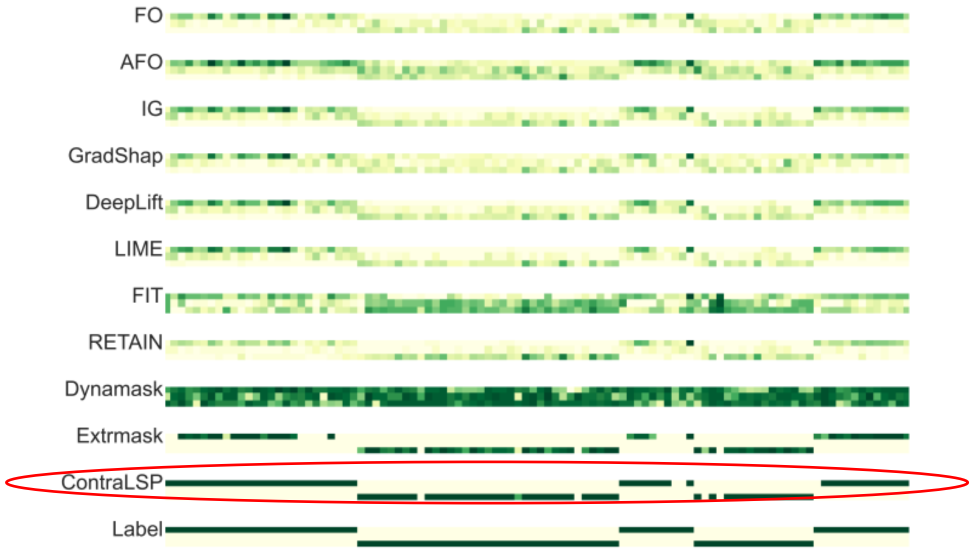
| Method | Rare-Time | | | | Rare-Time (DiffGroups) | | | |
|---|---|---|---|---|---|---|---|---|
| | AUP ↑ | AUR ↑ | $I_m/10^4$ ↑ | $S_m/10^2$ ↓ | AUP ↑ | AUR ↑ | $I_m/10^4$ ↑ | $S_m/10^2$ ↓ |
| FO | **1.00**±0.00 | 0.13±0.00 | 0.46±0.01 | 47.20±0.61 | **1.00**±0.00 | 0.16±0.00 | 0.53±0.01 | 54.89±0.70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

| Method | Rare-Observation | | | | Rare-Observation (DiffGroups) | | | |
|---|---|---|---|---|---|---|---|---|
| | AUP ↑ | AUR ↑ | $I_m/10^4$ ↑ | $S_m/10^2$ ↓ | AUP ↑ | AUR ↑ | $I_m/10^4$ ↑ | $S_m/10^2$ ↓ |
| FO | **1.00**±0.00 | 0.13±0.00 | 0.46±0.00 | 47.39±0.16 | **1.00**±0.00 | 0.14±0.00 | 0.50±0.01 | 52.13±0.96 |
| AFO | **1.00**±0.00 | 0.16±0.00 | 0.55±0.01 | 56.81±0.39 | **1.00**±0.00 | 0.16±0.01 | 0.54±0.02 | 56.92±1.24 |
| IG | **1.00**±0.00 | 0.13±0.00 | 0.46±0.00 | 47.82±0.15 | **1.00**±0.00 | 0.13±0.00 | 0.47±0.00 | 49.90±0.88 |
| SVS | **1.00**±0.00 | 0.13±0.00 | 0.46±0.00 | 47.39±0.16 | **1.00**±0.00 | 0.13±0.00 | 0.47±0.01 | 49.53±0.84 |
| Dynamask | 0.97±0.00 | 0.65±0.00 | 8.32±0.06 | 22.87±0.58 | 0.98±0.00 | 0.52±0.01 | 6.12±0.10 | 30.88±0.70 |
| Extrmask | **1.00**±0.00 | 0.76±0.00 | 13.25±0.07 | 9.55±0.39 | **1.00**±0.00 | 0.70±0.04 | 10.40±0.54 | 32.81±0.88 |
| ContraLSP | **1.00**±0.00 | **1.00**±0.00 | **20.68**±0.03 | **0.32**±0.16 | **1.00**±0.00 | **0.99**±0.00 | **20.51**±0.07 | **0.57**±0.20 |


## 2. Black-box Classification

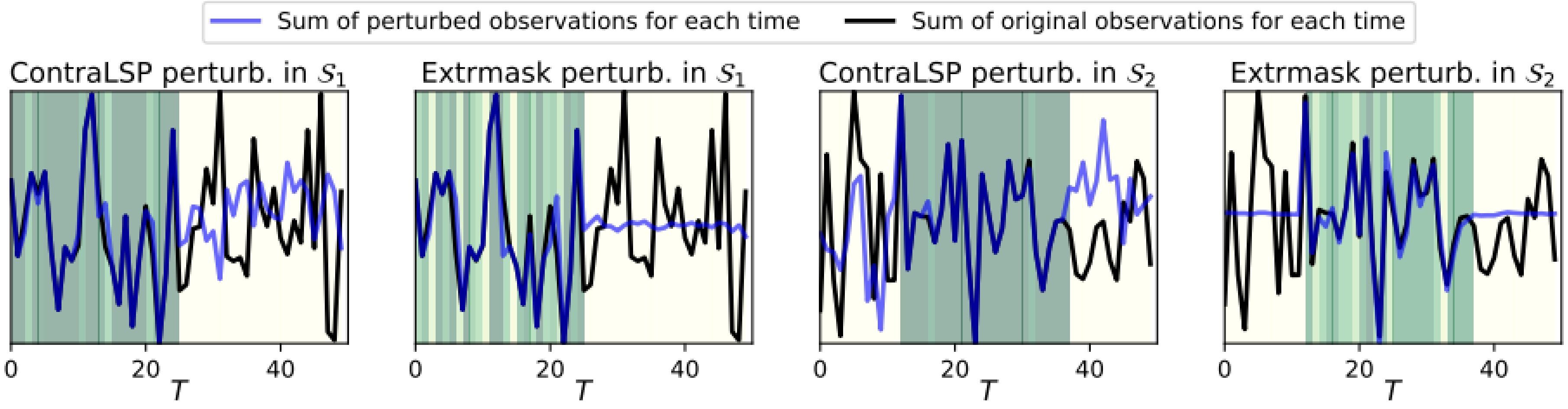
Table 2: Performance on Switch Feature and State data.

| Method | Switch-Feature | | | | State | | | |
|---|---|---|---|---|---|---|---|---|
| | AUP ↑ | AUR ↑ | $I_m/10^4$ ↑ | $S_m/10^3$ ↓ | AUP ↑ | AUR ↑ | $I_m/10^4$ ↑ | $S_m/10^3$ ↓ |
| FO | 0.89±0.03 | 0.37±0.02 | 1.86±0.14 | 15.60±0.28 | 0.90±0.05 | 0.30±0.01 | 2.73±0.15 | 28.07±0.54 |
| AFO | 0.82±0.06 | 0.41±0.02 | 2.00±0.14 | 17.32±0.29 | 0.84±0.08 | 0.36±0.03 | 3.16±0.27 | 34.03±1.10 |
| IG | 0.91±0.02 | 0.44±0.03 | 2.21±0.17 | 16.87±0.52 | 0.93±0.02 | 0.34±0.03 | 3.17±0.28 | 30.19±1.22 |
| GradShap | 0.88±0.02 | 0.38±0.02 | 1.92±0.13 | 15.85±0.40 | 0.88±0.06 | 0.30±0.02 | 2.76±0.20 | 28.18±0.96 |
| DeepLift | 0.91±0.02 | 0.44±0.02 | 2.23±0.16 | 16.86±0.52 | 0.93±0.02 | 0.35±0.03 | 3.20±0.27 | 30.21±1.19 |
| LIME | 0.94±0.02 | 0.40±0.02 | 2.01±0.13 | 16.09±0.58 | **0.95**±0.02 | 0.32±0.03 | 2.94±0.26 | 28.55±1.53 |
| FIT | 0.48±0.03 | 0.43±0.02 | 1.99±0.11 | 17.16±0.50 | 0.45±0.02 | 0.59±0.02 | 7.92±0.40 | 33.59±0.17 |
| Retain | 0.93±0.01 | 0.33±0.04 | 1.54±0.20 | 15.08±1.13 | 0.52±0.16 | 0.21±0.02 | 1.56±0.24 | 25.01±0.57 |
| Dynamask | 0.35±0.00 | 0.77±0.02 | 5.22±0.26 | 12.85±0.53 | 0.36±0.01 | 0.79±0.01 | 10.59±0.20 | 25.11±0.40 |
| Extrmask | 0.97±0.01 | 0.65±0.05 | 8.45±0.51 | 6.90±1.44 | 0.87±0.01 | 0.77±0.01 | 29.71±1.39 | 7.54±0.46 |
| ContraLSP | **0.98**±0.00 | **0.80**±0.03 | **24.23**±1.27 | **0.91**±0.26 | 0.90±0.03 | **0.81**±0.01 | **50.09**±0.78 | **0.50**±0.05 |



# Synthetic Experiments (with label)

- Counterfactual information



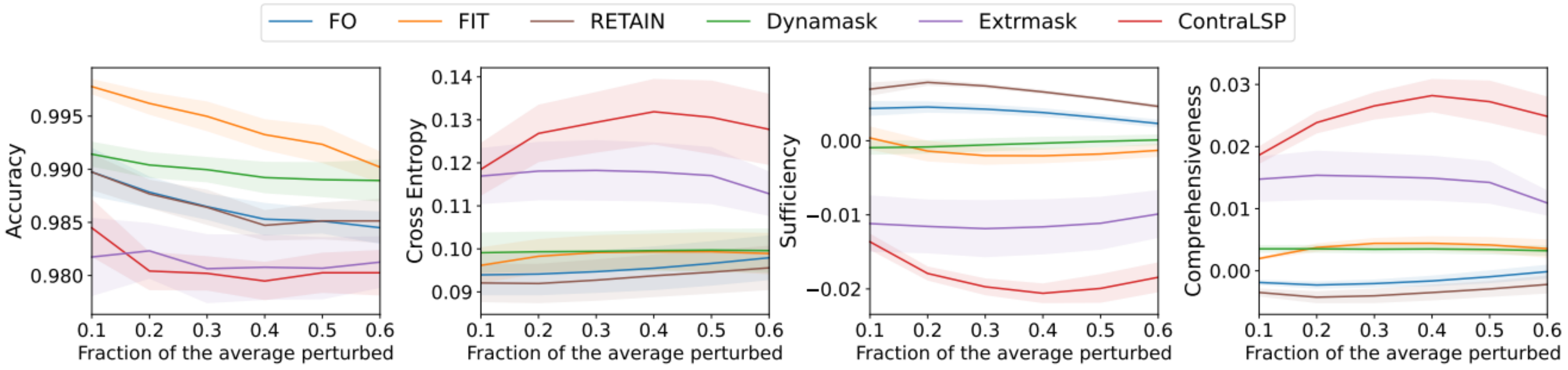

- Distribution analysis of perturbations

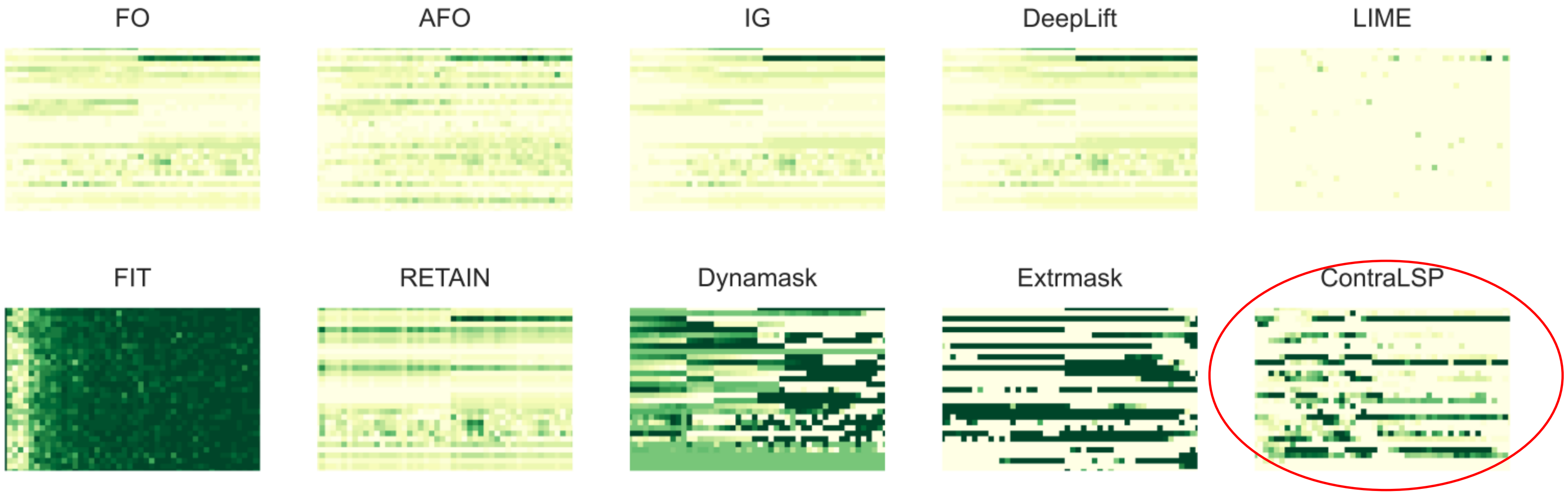
Table 12: Difference between the distribution of different perturbations and the original distribution.

| 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!**