BIT

SCHOOL OF COMPUTER SCIENCE & TECHNOLOGY

Deep Statistical Learning Laboratory

ICLR 2026

Learning for Highly Faithful Explainability

Yuhan Guo[1], Lizhong Ding[2]*, Shihao Jia, Yanyu Ren, Pengqi Li, Jiarun Fu, Changsheng Li, Ye Yuan, Guoren Wang

1: 3220231188@bit.edu.cn
2: lizhong.ding@outlook.com *: Corresponding author

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

德以明理 学以精工

Why Do We Need Explainability?

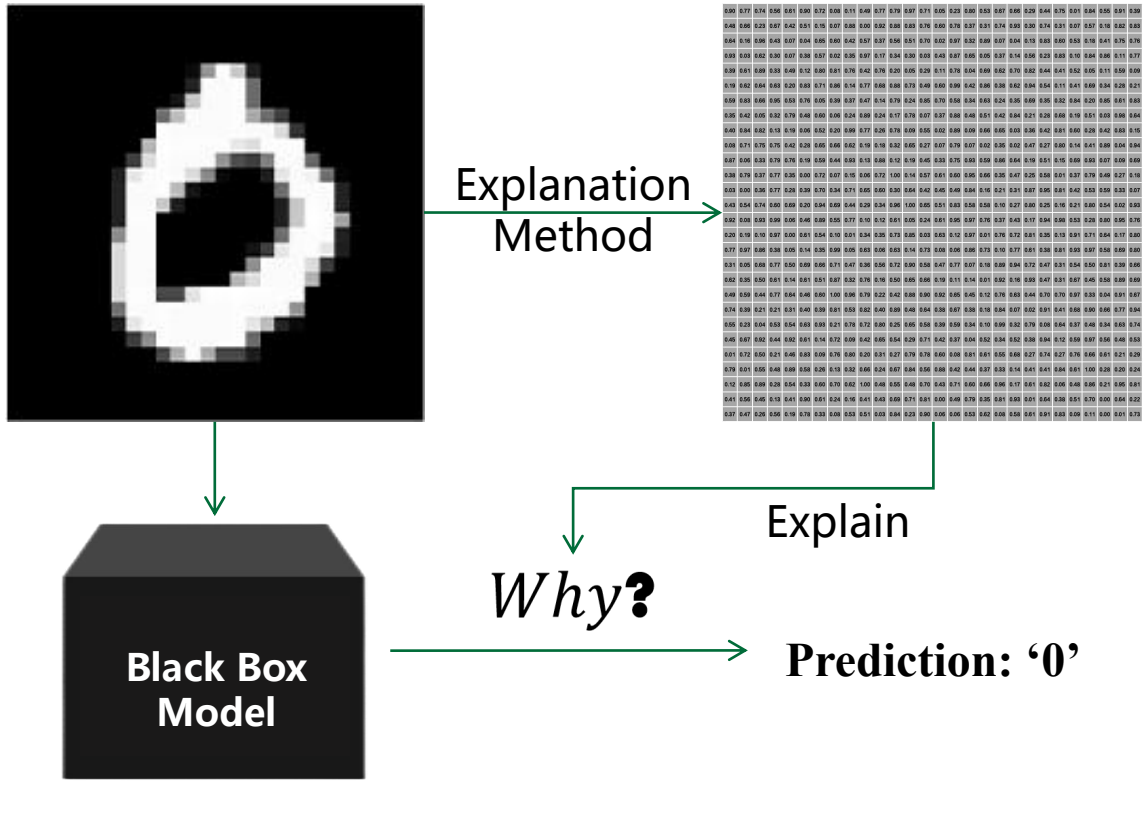
AI: Stop the cardiac pacing.

AI: Make an emergency turn.

AI: Determine as guilty.

- **Trust**: In high-stakes domains such as medicine, autonomous driving, and law, the deployment of models that **fail to provide interpretable outcomes** is precluded
- **Transparency:** The black-box nature of deep models brings severe reliability risks, which drives the rapid development of the **eXplainable Artificial Intelligence** (XAI) field

What is an explanation?

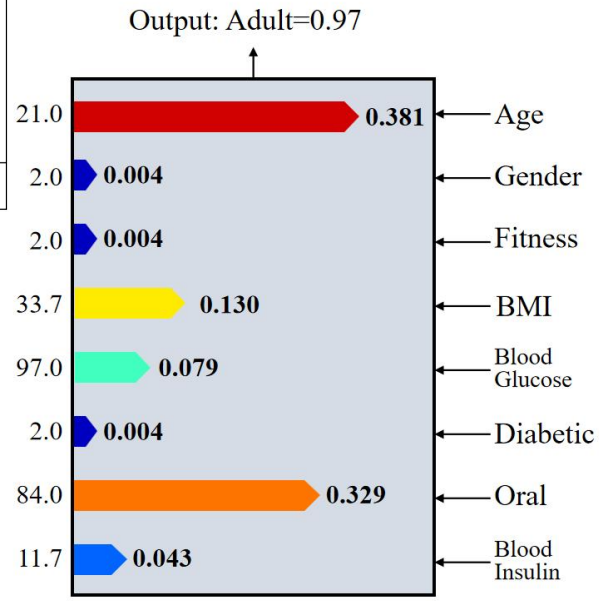


Novell's Microsoft attack completes Linux conversion: Novell Inc. has completed its conversion to Linux by launching an attack on Microsoft Corp., claiming that the company has stifled software innovation and that the market will abandon Microsoft Windows at some point in the future.

$\hat{y}_{\boldsymbol{x}}$ = 99% Sci/Tech; $\hat{y}_{\boldsymbol{x}\setminus\mathcal{A}}$ = 14%; $\hat{y}_{\boldsymbol{x}\setminus\text{L2E}}$ = 0.7%

Word Importance

Saliency Map



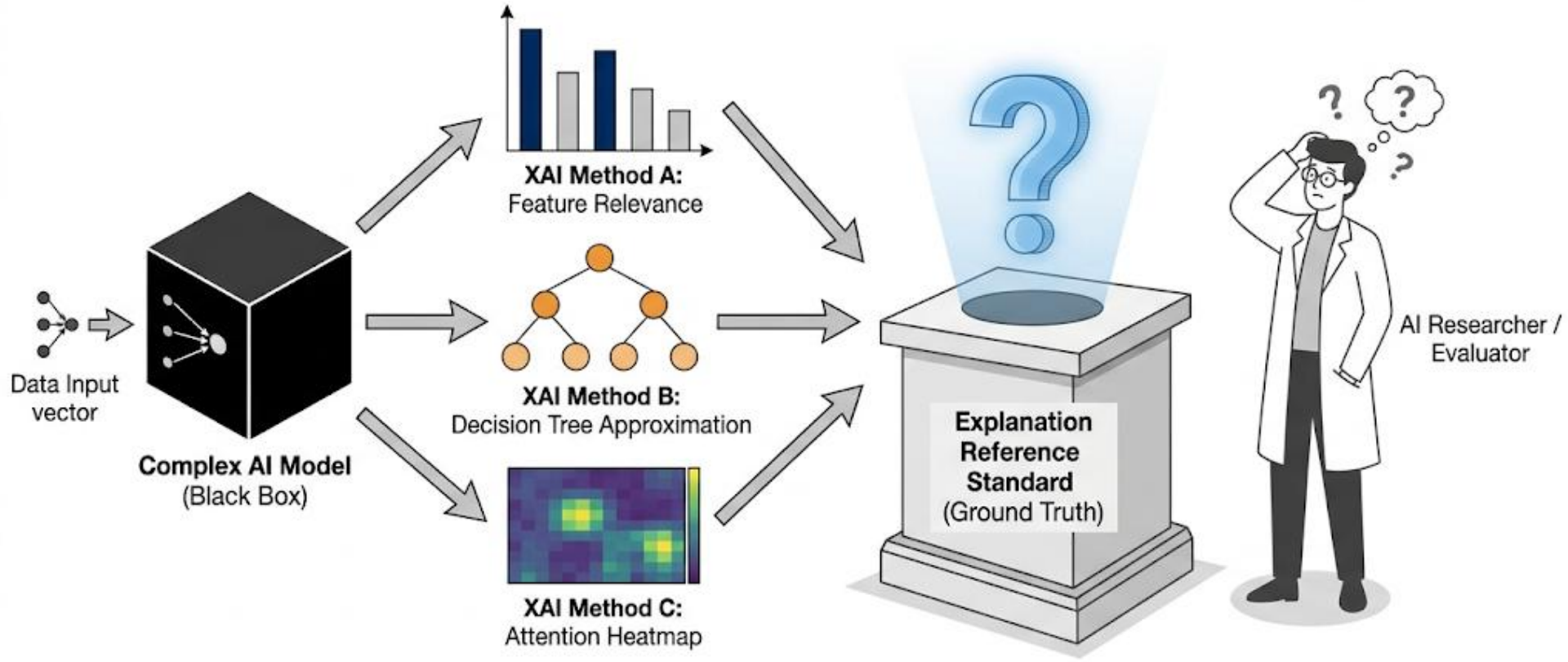
Feature Relevance

- Saliency explanation assigns an **importance score** to each input feature (e.g., pixels in image classification), indicating the degree to which the model attends to specific features

Major Challenge in XAI: Lack of Ground Truth

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

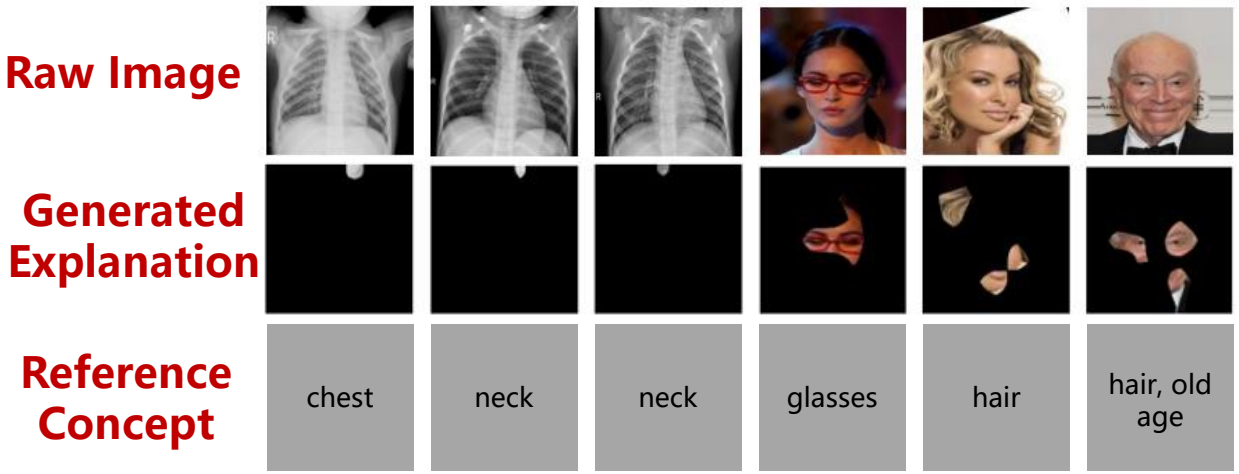
- A core challenge in the XAI field is the absence of a universally applicable **Ground Truth** concept for judging whether an explanation is correct or incorrect.



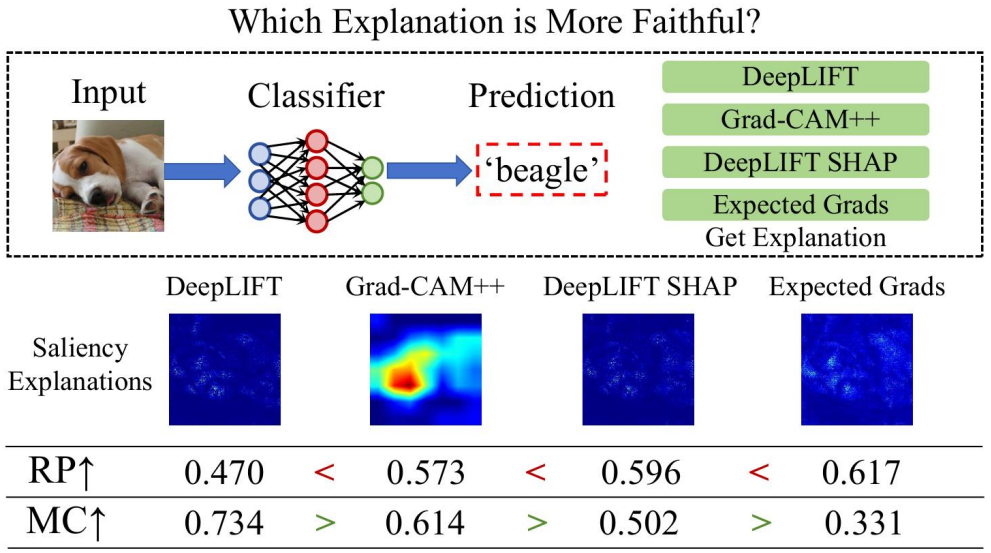
Faithfulness of Explanations

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

- **Faithfulness** evaluation metrics provides a **proxy metric** for measuring explanation quality and avoiding biases introduced by human subjective evaluation



Subjective Concept Alignment Evaluation



Objective Faithfulness Score

- The assessment of faithfulness varies across different perturbation-based experimental strategy, leading to **incompatible or conflicting outcomes** in specific scenarios

DeepFaith

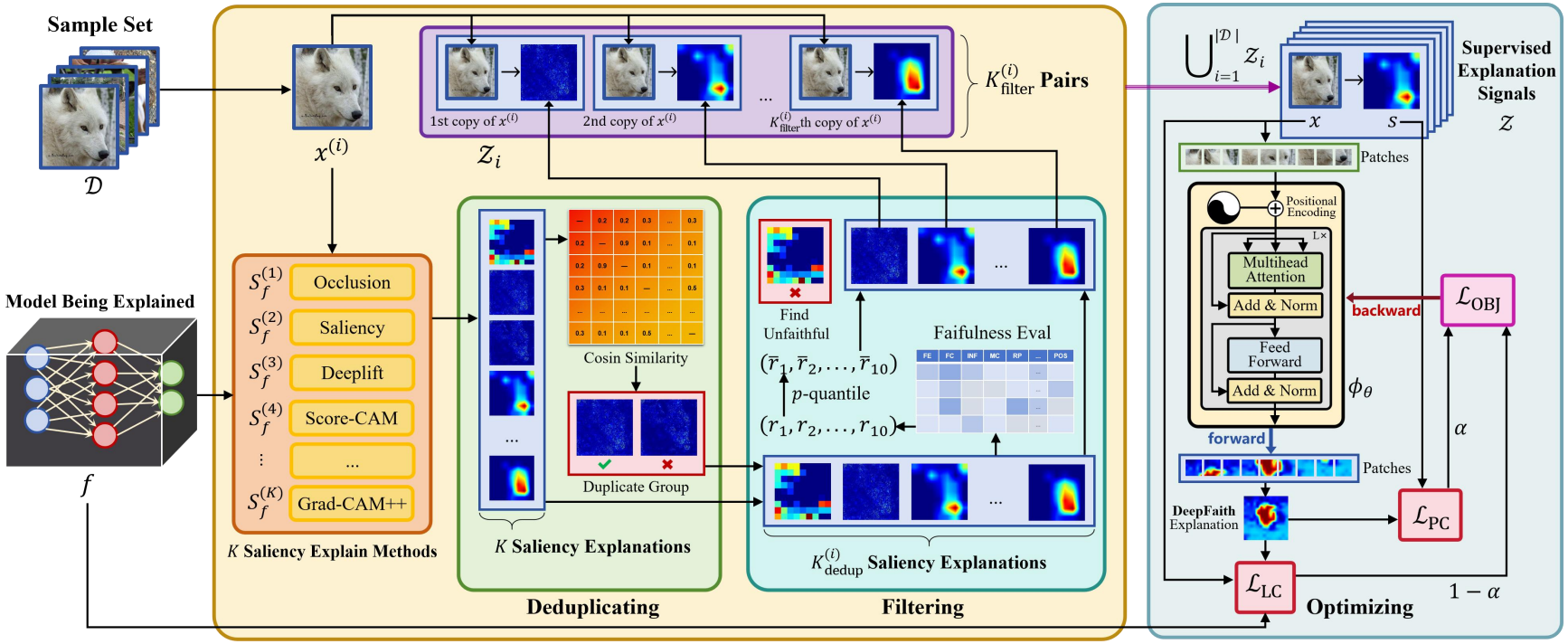
北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

Background

Conflicting faithfulness metrics and inconsistent methods impede a unified ground truth

Challenges:

1. Inconsistent formalizations of faithfulness metrics hinder unified theoretical analysis
2. Discrepant data from different methods impede consistent data utilization



DeepFaith:

Theory - Re-formalize metrics and unify objectives

Data - Curate high-quality explanations via deduplication and filtering

Approach XAI Ground Truth at all levels

Formulation and Theoretical Unification

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

- **DeepFaith** unifies 10 faithfulness evaluation metrics under **the same theoretical framework** by extracting shared functional functions.

Faithfulness Metric	Input	Formula	Output	
Faithfulness Correlation (FC) ↑	$s; x, f$	$\tau\left[\left(\sum_{i\in\mathcal{I}} s_i\right)_{\mathcal{I}\subseteq[n]}, \left(\Delta\left[f(x), f(x\setminus\mathcal{I})\right]\right)_{\mathcal{I}\subseteq[n]}\right]$	$[-1,1]$	Faithfulness Metric for **Saliency Explanations** s
Faithfulness Estimate (FE) ↑	$S_f; \{x^{(i)}, \mathcal{I}_i\}_{i=1}^N, f$	$\tau\left[\left(\sum_{j\in\mathcal{I}_i} S_f(x^{(i)})_j\right)_{i=1}^N, \left(\Delta\left[f(x^{(i)}), f(x^{(i)}\setminus\mathcal{I}_i)\right]\right)_{i=1}^N\right]$	$[-1,1]$	
Monotonicity Correlation (MC) ↑	$s; x, \{\mathcal{I}_i\}_{i=1}^N, f$	$\tau[(\sum_{j\in\mathcal{I}_i} s_j)_{i=1}^N, (\Delta\left[f(x), f(x\setminus\mathcal{I}_i)\right])_{i=1}^N]$	$[-1,1]$	
Infidelity (INF) ↑	$s; x, \{\mathcal{I}_i\sim\mathcal{P}([n])\}_{i=1}^N, f$	$\tau[(\sum_{j\in\mathcal{I}_i} s_j)_{i=1}^N, (\Delta\left[f(x), f(x\setminus\mathcal{I}_i)\right])_{i=1}^N]$	$[-1,1]$	
Deletion Score (DEL) ↓	$\pi; x, f$	$\frac{1}{n}\int_{i=0^+}^{n} \Delta^{-}\left[f(x), f(x\setminus\bigcup_{j=1}^{\lceil i\rceil}\pi(j))\right]\mathrm{d}i$	$[0,1]$	Faithfulness Metric for **Permutation Explanations** π
Insertion Score (INS) ↑	$\pi; x, f$	$\frac{1}{n}\int_{i=0^+}^{n} \Delta^{-}\left[f(x), f(x^{\circ}\cup\bigcup_{j=1}^{\lceil i\rceil}\pi(j))\right]\mathrm{d}i$	$[0,1]$	
Negative Perturbation (NEG) ↑	$\pi; x, f$	$\frac{1}{t}\int_{i=0^+}^{t} \Delta^{-}\left[f(x), f(x\setminus\bigcup_{j=1}^{\lceil i\rceil}\overleftarrow{\pi}(j))\right]\mathrm{d}i$	$[0,1]$	
Positive Perturbation (POS) ↓	$\pi; x, f$	$\frac{1}{t}\int_{i=0^+}^{t} \Delta^{-}\left[f(x), f(x\setminus\bigcup_{j=1}^{\lceil i\rceil}\pi(j))\right]\mathrm{d}i$	$[0,1]$	
Region Perturbation (RP) ↑	$\Pi_f; \{x^{(i)}\}_{i=1}^N, f$	$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n+1}\sum_{j=0}^{n}\Delta\left[f(x^{(i)}), f(x^{(i)}\setminus\bigcup_{k=1}^{j}\Pi_f(x^{(i)})(k))\right]\right)$	$[0,1]$	
Iterative Removal of Features (IROF) ↑	$\Pi_f; \{x^{(i)}\}_{i=1}^N, f$	$\frac{1}{Nn}\sum_{i=1}^{N}\int_{j=0^+}^{n} 1-\Delta^{-}\left[f(x^{(i)}), f(x^{(i)}\setminus\bigcup_{k=1}^{\lceil j\rceil}\Pi_f(x^{(i)})(k))\right]\mathrm{d}j$	$[0,1]$	

Formulation and Theoretical Unification

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

Proposition 1. *Given a model f being explained and its input space $\mathcal{X}$, for a fixed correlation measure τ and perturbation effect Δ, suppose there exists a saliency explanation mapping S_f^* such that $\forall x \in \mathcal{X}$ and $\forall \{\mathcal{I}_i \subseteq [n]\}_{i=1}^N$,*

$$S_f^* = \underset{S_f}{\operatorname{argmax}}\ \tau\left[\left(\sum_{j\in\mathcal{I}_i} S_f(x)_j\right)_{i=1}^N, \left(\Delta[f(x), f(x\backslash\mathcal{I}_i)]\right)_{i=1}^N\right], \tag{1}$$

then the saliency explanations generated by S_f^* always achieve optimal faithfulness under the FC, FE, INF, *and* MC *evaluation metrics.*

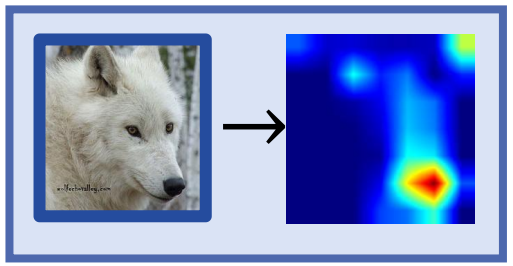
S_f^* achieves optimality across all 4 saliency explanation metrics.

Theorem 1. *Under the conditions of Proposition 1, given a fixed preservation effect Δ^- that is strictly decreasing with Δ, let $\Pi_f^*(\cdot) = \mathfrak{P}[S_f^*(\cdot)]$ denote the permutation explanation mapping induced by S_f^*, then for any sample x, $\Pi_f^*(x)$ always achieve optimal faithfulness under the* DEL, INS, NEG, POS, RP *and* IROF *evaluation metrics.*

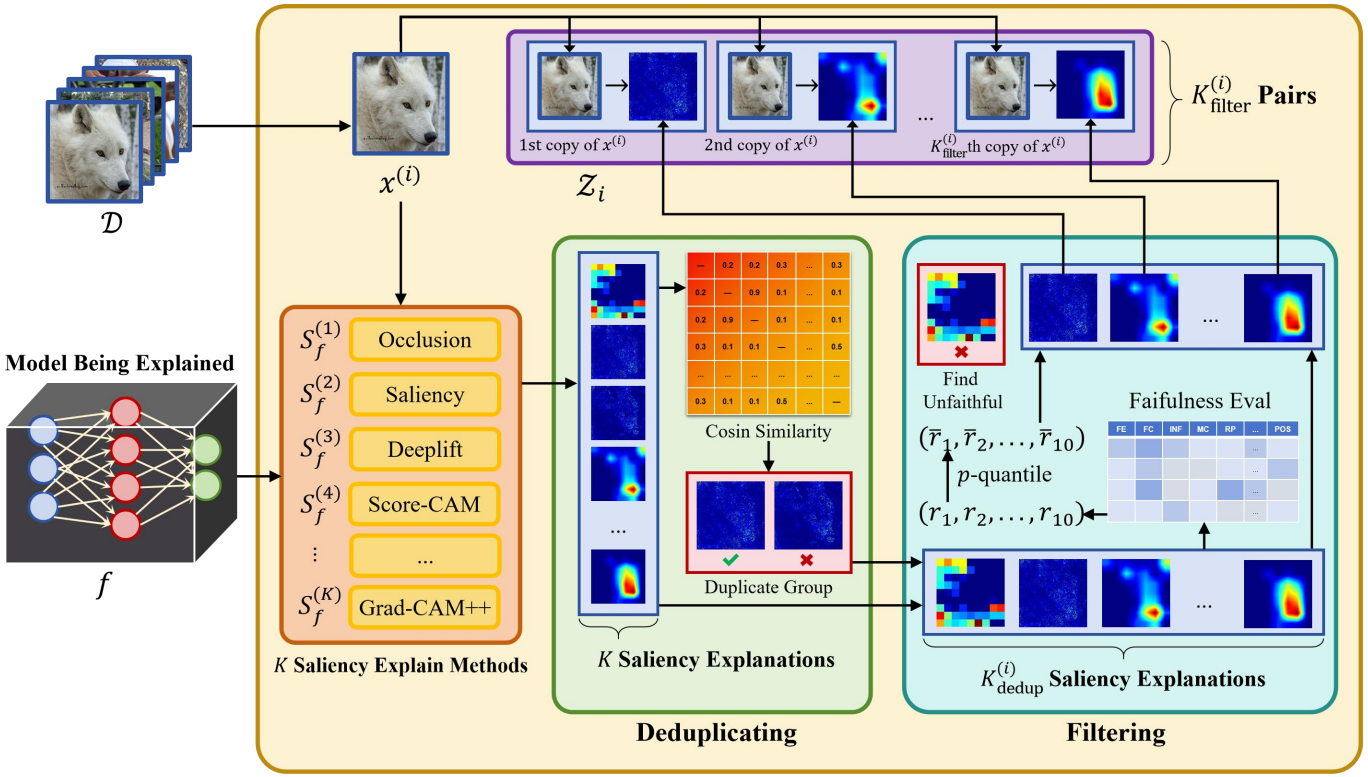
The S_f^* -induced permutation explanation also optimizes all 6 permutation metrics.

- We discover and prove that there theoretically exists a mapping S_f^* that achieves **optimality simultaneously** under 10 faithfulness metrics.

Utilization of Prior Explanation Methods



- Different explanation methods essentially reflect the **functional dependence** from input features to explanations, and such **mapping patterns are generalizable** across similar samples



- We get high-quality prior explanation signals by:

(1) Similarity-based **deduplicating**

(2) Faithfulness-based **filtering**

- An **input feature-explanation** mapping set $\mathcal{Z}$ integrates diversity and high quality

$$\mathcal{Z} = \left\{ \left(x^{(i)}, S_f^{(j)}(x^{(i)}) \right) \mid i \leq |\mathcal{D}|, j \in \left[K^{(i)}_{\text{filter}} \right] \right\}.$$

Utilization of Prior Explanation Methods

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

- We propose to fit existing explanations $\mathcal{Z}$ with a **deep neural network** ϕ_θ, further enhancing its faithfulness via the optimization objective S_f^*
- Based on the expression of S_f^*, we design an optimization objective as the loss function

$$\mathcal{L}_{\mathrm{LC}}(\phi_\theta; \mathcal{D}, f) = \frac{1}{2} - \frac{1}{2|\mathcal{D}|} \sum_{x \in \mathcal{D}} \tau \left[\left(\sum_{i \in \mathcal{I}_j} \phi_\theta(x)_i \right)_{j=1}^{k}, \left(\Delta[f(x), f(x \setminus \mathcal{I}_j)] \right)_{j=1}^{k} \right]$$

- To learn the **mapping patterns** in $\mathcal{Z}$, we design a supervised optimization objective as the loss function for training the explainer ϕ_θ

$$\mathcal{L}_{\mathrm{PC}}(\phi_\theta; \mathcal{Z}) = \frac{1}{|\mathcal{Z}|} \sum_{(x,s) \in \mathcal{Z}} (1 - \tau [\phi_\theta(x), s])$$

- We aim to achieve the **collaborative optimization** of $\mathcal{L}_{LC}$ and $\mathcal{L}_{PC}$, i.e., minimizing the following objective $\mathcal{L}_{OBJ}$

$$\mathcal{L}_{\mathrm{OBJ}}(\phi_\theta; \mathcal{D}, f, \mathcal{Z}) = \alpha \mathcal{L}_{\mathrm{PC}}(\phi_\theta; \mathcal{Z}) + (1 - \alpha) \mathcal{L}_{\mathrm{LC}}(\phi_\theta; \mathcal{D}, f).$$

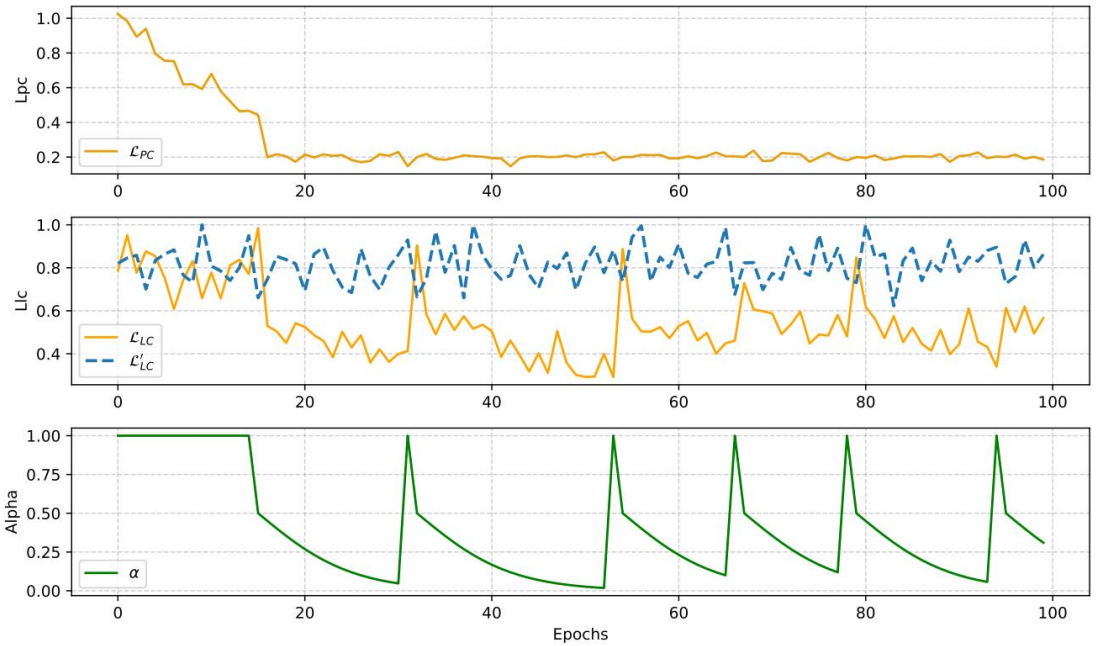
Utilization of Prior Explanation Methods

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

- We propose a **dynamic joint optimization algorithm** to achieve the collaborative optimization of $\mathcal{L}_{LC}$ and $\mathcal{L}_{PC}$

Algorithm 1 Dynamic Joint Optimization Strategy for **DeepFaith**

Input: Training dataset $\mathcal{D}$, target model f, supervised signals $\mathcal{Z}$, initial explainer parameters $\phi_\theta^{(0)}$, variance threshold ϵ, monitoring window e, scaling factor $C \geq 1$, and learning rate η
Initialize: dynamic weight $\alpha^{(0)} \leftarrow 1$, convergence flag $CF = 0$, and set $t_0 = 0$
while not converged **do**
  $\mathcal{L}_{\text{OBJ}}^{(t)} \leftarrow \alpha^{(t)}\mathcal{L}_{\text{PC}}^{(t)} + (1-\alpha^{(t)})\mathcal{L}_{\text{LC}}^{(t)}$ ▷ Compute total loss
  $\phi_\theta^{(t+1)} \leftarrow \phi_\theta^{(t)} - \eta\nabla_{\phi_\theta}\mathcal{L}_{\text{OBJ}}^{(t)}$ ▷ Update explainer parameters
  $\sigma_{\text{PC}}^2 \leftarrow \text{Var}\left(\mathcal{L}_{\text{PC}}^{(t-e+1)}, \ldots, \mathcal{L}_{\text{PC}}^{(t)}\right)$ ▷ Compute variance of $\mathcal{L}_{\text{PC}}$ over last e iterations
  if $CF = 0$ **then**
    if $\sigma_{\text{PC}}^2 < \epsilon$ **then**
      $CF \leftarrow 1, t_0 \leftarrow t$ ▷ Set Convergence Flag
    end if
  end if
  if $\sigma_{\text{PC}}^2 > C\epsilon$ **then**
    $CF \leftarrow 0, \alpha^{(t+1)} \leftarrow 1$ ▷ Reset Convergence Flag
  end if
  if $CF = 1$ **then**
    $\alpha^{(t+1)} \leftarrow 1 - \frac{1}{1+\exp(-\frac{t-t_0}{C})}$ ▷ Gradually decrease α
  end if
end while



- $\mathcal{L}_{PC}$ converges stably, while $\mathcal{L}_{LC}$ (the blue dashed line) exhibits severe oscillation when optimized independently

Explanation Faithfulness

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

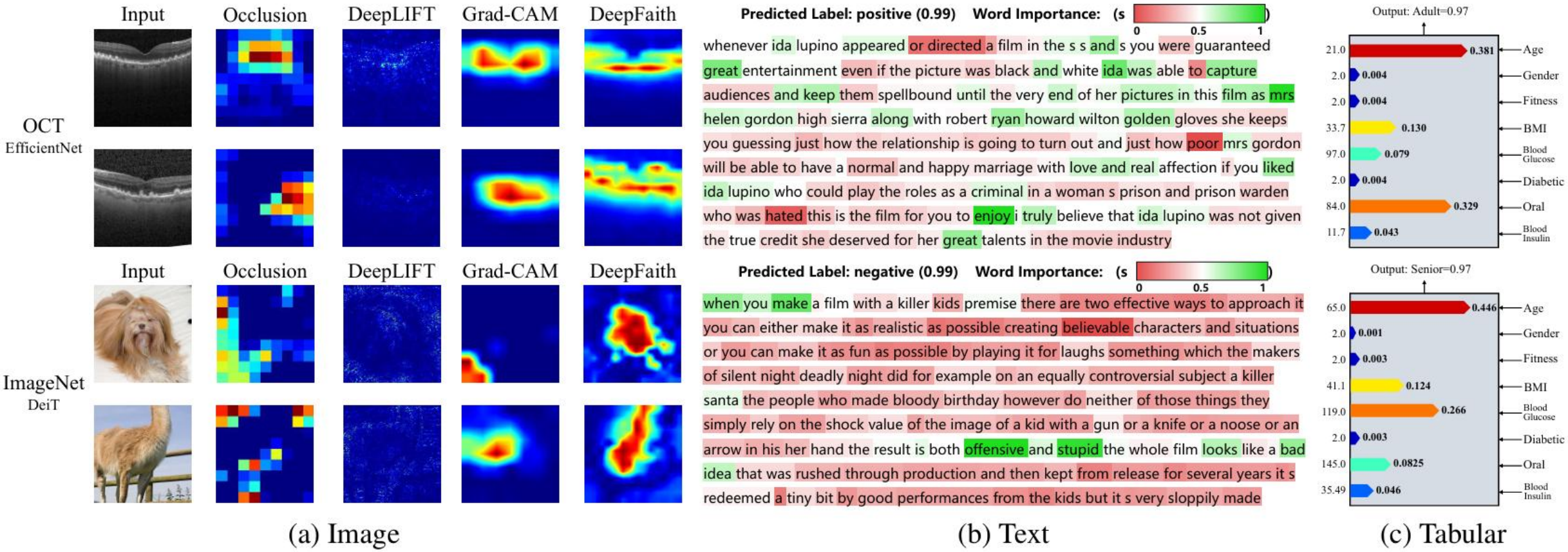
- We compare the average rankings of **DeepFaith** and baseline methods across all faithfulness evaluation metrics in **12 explanation tasks across 3 modalities**

Method	Image Tasks						Text Tasks				Tabular Tasks	
	OCT			ImageNet			IMDb		AGNews		NAP	WCD
	DeiT	EfficientNet	ResNet	DeiT	EfficientNet	ResNet	LSTM	Transformer	LSTM	Transformer	MLP	MLP
DeepFaith (ours)	**3.4**	**2.9**	**4.1**	**4.4**	**4.4**	**3.3**	**2.3**	**2.1**	**2.9**	**2.7**	**1.8**	**1.8**
Integrated Grads	7.8	7.6	4.8	6.4	7.0	5.4	3.3	5.6	4.9	5.9	2.8	5.2
Gradient SHAP	N/A	N/A	N/A	N/A	N/A	N/A	4.4	4.0	**2.9**	4.2	4.7	7.3
DeepLIFT	5.8	7.8	8.1	7.0	6.9	8.4	6.1	6.4	7.9	5.9	4.4	2.3
Saliency	13.2	11.0	12.8	10.7	11.1	10.6	5.2	5.9	4.7	5.8	2.8	4.9
Occlusion	8.5	6.5	8.4	8.9	9.6	10.9	4.6	3.6	**2.9**	**2.7**	3.3	5.9
Feature Ablation	N/A	N/A	N/A	N/A	N/A	N/A	6.4	5.1	6.6	8.5	3.5	4.5
LIME	12.3	8.1	9.9	10.7	6.6	8.5	7.7	6.8	4.6	4.5	4.7	2.7
Kernel SHAP	4.2	10.9	12.1	7.0	5.9	8.9	5.0	5.5	6.4	3.9	3.9	8.9
Input × Gradient	5.7	12.3	12.2	5.3	12.9	10.7	N/A	N/A	N/A	N/A	N/A	N/A
Guided Backprop	12.3	6.5	7.6	11.4	10.3	10.4	N/A	N/A	N/A	N/A	N/A	N/A
Grad-CAM	8.6	8.2	7.6	11.9	6.6	7.0	N/A	N/A	N/A	N/A	N/A	N/A
Score-CAM	7.0	7.9	6.0	5.8	7.2	7.1	N/A	N/A	N/A	N/A	N/A	N/A
Grad-CAM++	5.0	10.0	7.4	4.9	7.8	8.3	N/A	N/A	N/A	N/A	N/A	N/A
Expected Grads	6.9	8.0	7.1	7.1	9.5	6.4	N/A	N/A	N/A	N/A	N/A	N/A
DeepLIFT SHAP	6.9	7.3	5.6	7.5	8.7	9.3	N/A	N/A	N/A	N/A	N/A	N/A
LRP	12.0	4.5	5.4	10.2	5.0	4.5	N/A	N/A	N/A	N/A	N/A	N/A

Explanation Visualization

北京理工大學 BEIJING INSTITUTE OF TECHNOLOGY

- By learning the patterns of existing explanations, $\phi_\theta(x)$ achieves high faithfulness while **aligning with human cognition**



(a) Image (b) Text (c) Tabular