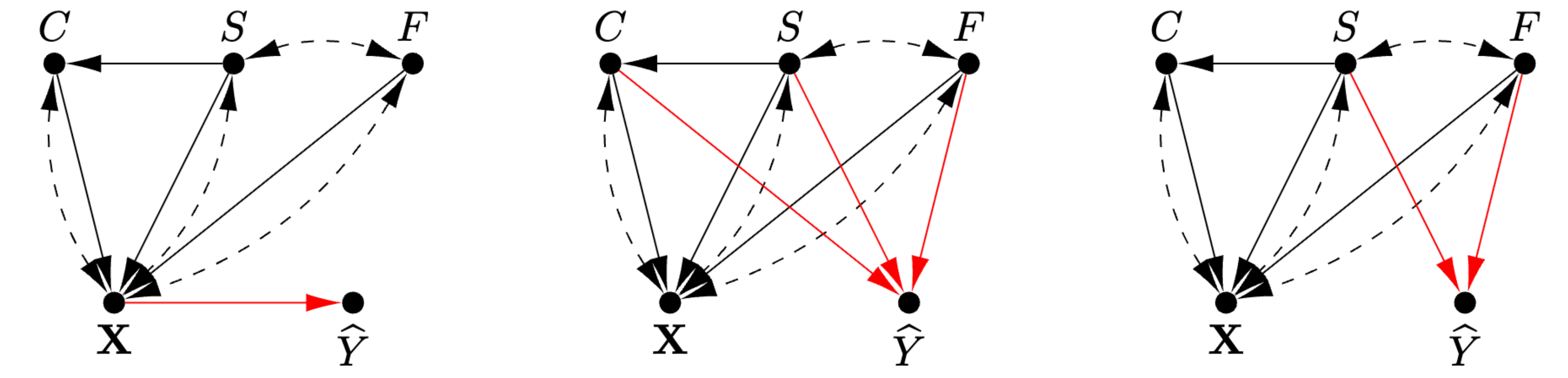
From Black-box to Causal-box: Towards Building More Interpretable Models

Inwoo Hwang Yushu Pan Elias Bareinboim

Background & Motivation



(a) Blackbox prediction (BP). (b) Concept-based prediction (CP). (c) Generalized CP (GCP).

- $\mathbf{X}$: input image (human face), $\widehat{Y}$: label prediction (attractiveness)
- C: high cheekbones, S: smiling, F: gender
- Standard black-box models and concept-based models are effective at predicting labels based on statistical correlations in the data.
- **Counterfactual question**: *"What if they had smiled?"* — $P(\widehat{Y}_{s'} \mid \mathbf{X})$
- Existing models cannot answer their own counterfactual questions.

Graphical Criterion

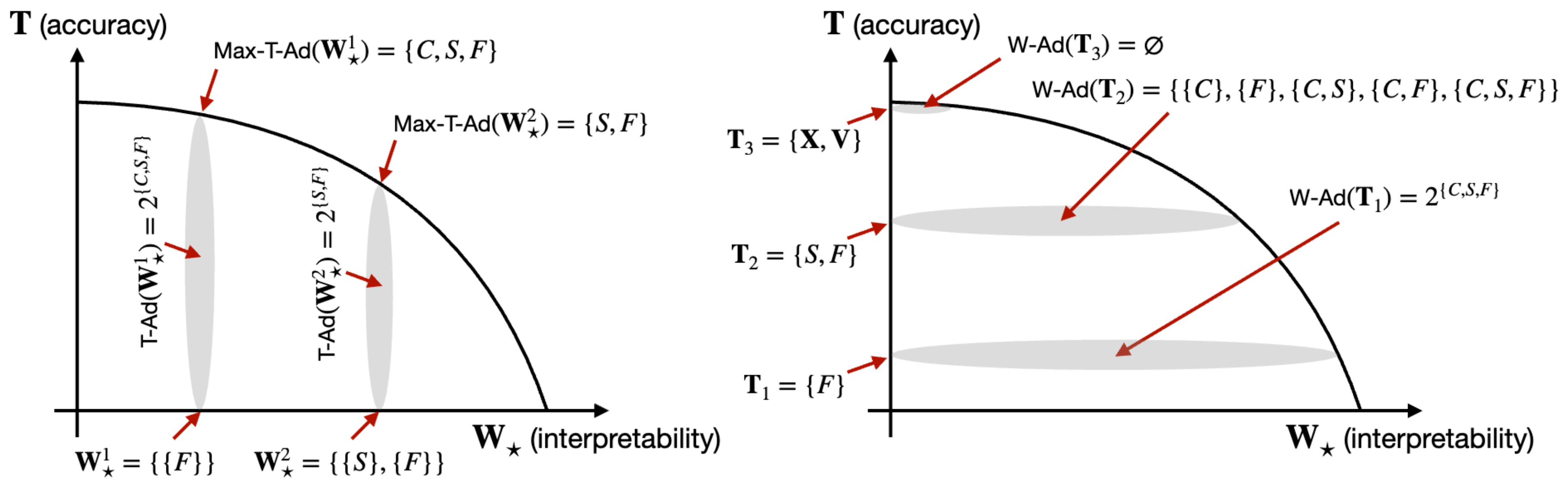
- $\mathbf{T}$: predictor of the label — $\mathbf{T}$ = (a) $\mathbf{X}$, (b) $\{C, S, F\}$, (c) $\{S, F\}$
- $\mathbf{W}$: features involved in counterfactual question (e.g., $\mathbf{W} = \{S\}$)
- **Question**: For which type of models can we evaluate a counterfactual question $P(\widehat{Y}_{\mathbf{w}} \mid \mathbf{X})$?

- **[Theorem]** A model is causally interpretable w.r.t. a query $Q(\mathbf{W}) = P(\widehat{Y}_{\mathbf{w}} \mid \mathbf{X})$ if and only if $\mathbf{T} \subseteq \mathbf{W} \cup ND(\mathbf{W})$.
- **[Implication]** Blackbox models are never causally interpretable.
- **[Implication]** For concept-based models, $\mathbf{T}$ should not include the descendants of $\mathbf{W}$. (We do not need to know the full causal graph!)

Can we understand the model's counterfactual predictions under hypothetical *"What if"* questions?

Answer: Depends on the model architecture!

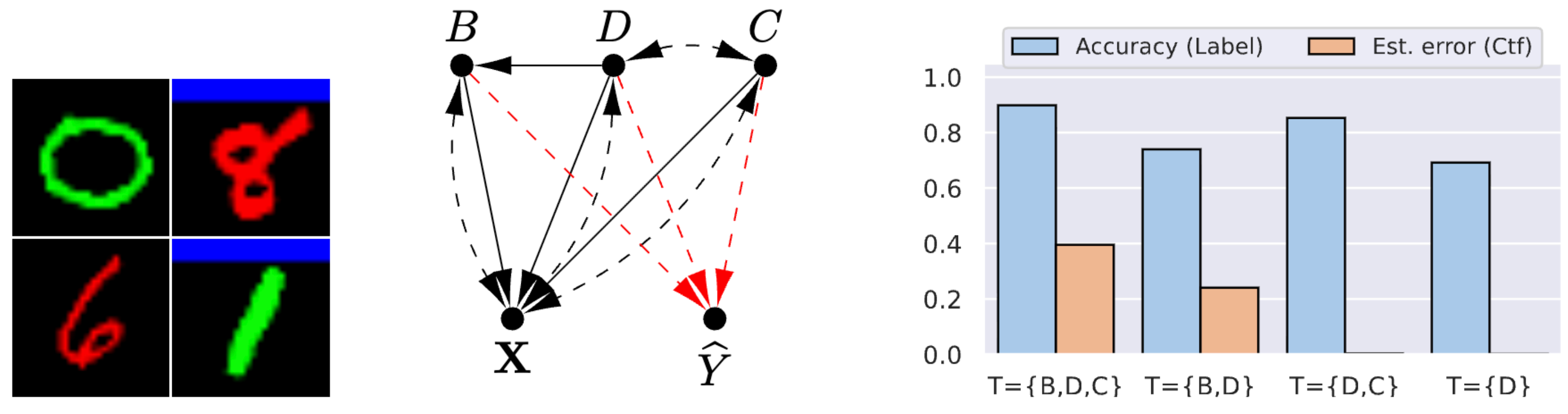
Accuracy-Interpretability Trade-off


$\mathbf{T}$ (accuracy)
Max-T-Ad($\mathbf{W}^1_\star$) = $\{C, S, F\}$
Max-T-Ad($\mathbf{W}^2_\star$) = $\{S, F\}$
T-Ad($\mathbf{W}^1_\star$) = $2^{\{C,S,F\}}$
T-Ad($\mathbf{W}^2_\star$) = $2^{\{S,F\}}$
$\mathbf{W}^1_\star = \{\{F\}\}$
$\mathbf{W}^2_\star = \{\{S\}, \{F\}\}$
$\mathbf{W}_\star$ (interpretability)
$\mathbf{T}$ (accuracy)
W-Ad($\mathbf{T}_3$) = $\varnothing$
W-Ad($\mathbf{T}_2$) = $\{\{C\}, \{F\}, \{C,S\}, \{C,F\}, \{C,S,F\}\}$
W-Ad($\mathbf{T}_1$) = $2^{\{C,S,F\}}$
$\mathbf{T}_3 = \{\mathbf{X}, \mathbf{V}\}$
$\mathbf{T}_2 = \{S, F\}$
$\mathbf{T}_1 = \{F\}$
$\mathbf{W}_\star$ (interpretability)

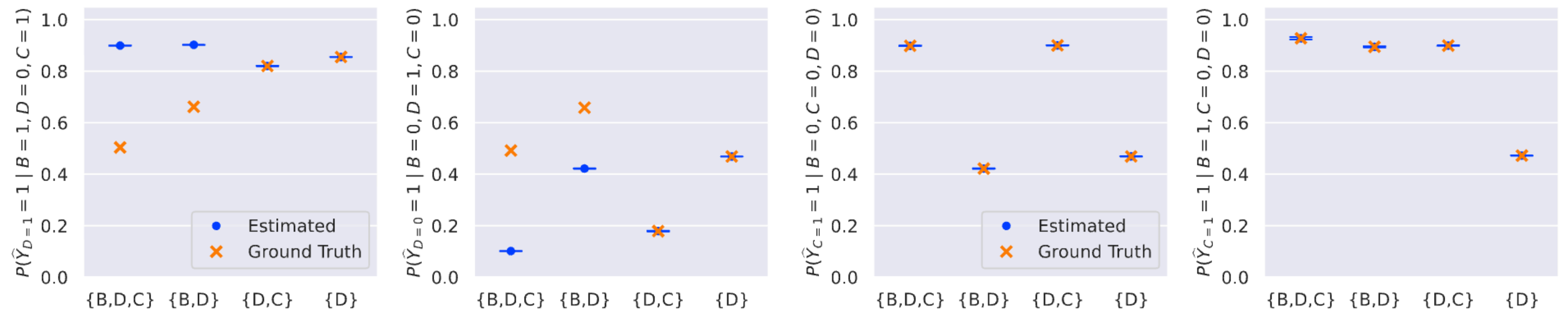
- Predictive power **decreases** as we want the models to answer **more** counterfactual queries.
- Counterfactuals that can be evaluated from the model **decrease** as the predictive power **increases**.

Experiment

Bar MNIST

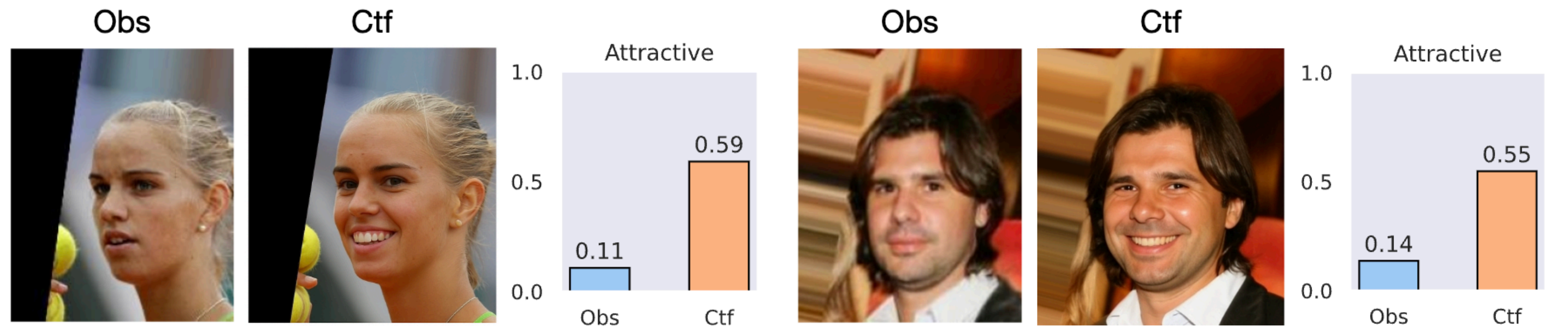


(a) (b) (c)



(a) Changing digit. (b) Changing color.

CelebA



- We examine how a model makes predictions under the counterfactual conditions *"Would the person look attractive had they smiled?"*, for causally interpretable models.

Conclusion

- Standard black-box and concept-based models cannot answer their own counterfactual "what-if" questions, a fundamental limitation we prove formally.
- We introduce the **first causal framework for building interpretable-by-design models**, revealing a precise trade-off between interpretability and predictive accuracy.

39th Conference on Neural Information Processing Systems (NeurIPS 2025), San Diego, USA