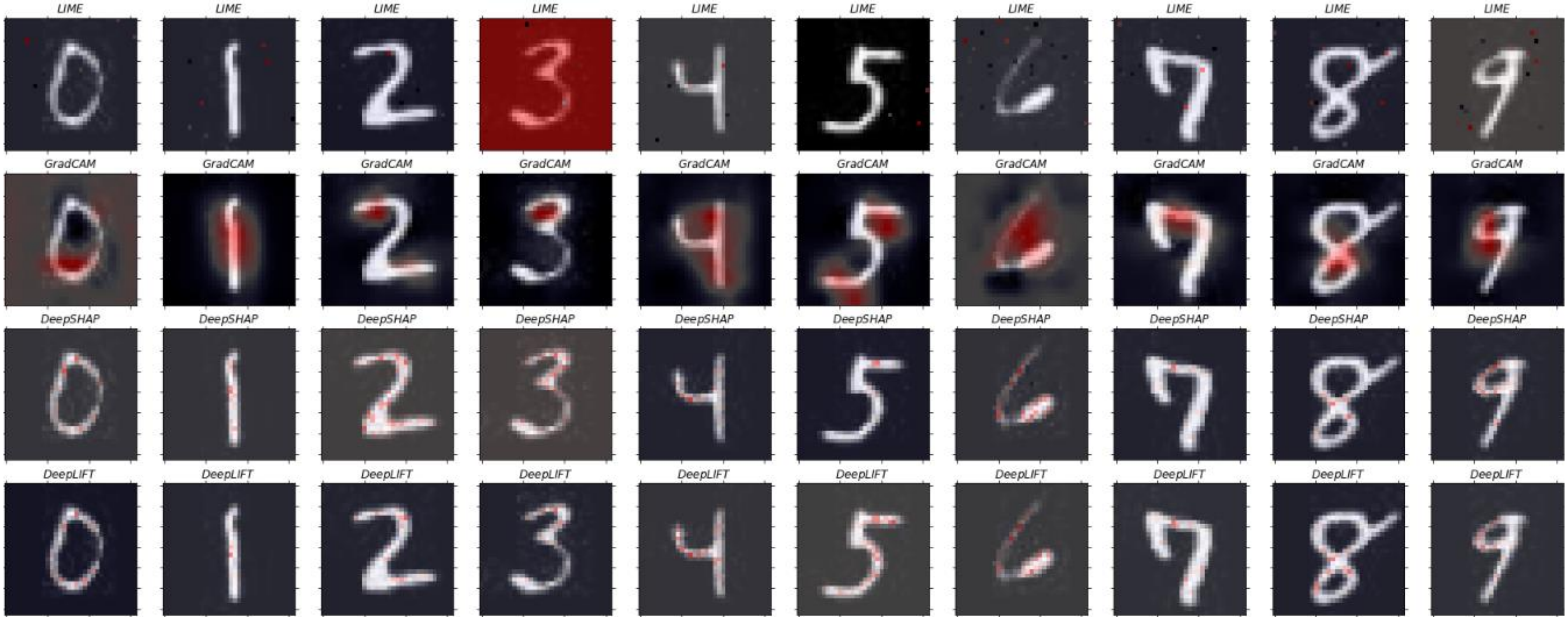
GLANCE: Global to Local Architecture Agnostic Explanations

Avinash Kori, Ben Glocker, and Francesca Toni

a.kori21@imperial.ac.uk

Motivation



Research Questions

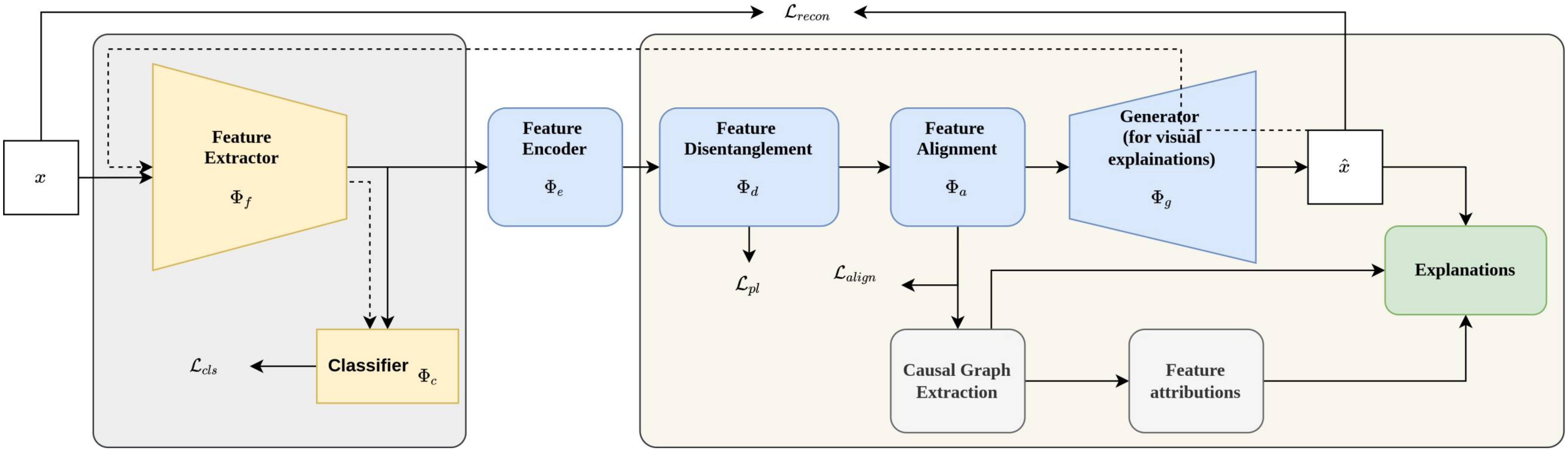
- Can we retrieve perceived causal knowledge/data generating process from the latent knowledge of the model?

- Can we develop a framework for generating human involved explanations?

- Can we simplify heuristic search in counterfactual explanations?

Our Approach

Framework



$\mathcal{L}_{recon}$
x
Feature Extractor
Φ_f
Feature Encoder
Φ_e
Feature Disentanglement
Φ_d
Feature Alignment
Φ_a
Generator (for visual explainations)
Φ_g
$\hat{x}$
Explanations
$\mathcal{L}_{pl}$
$\mathcal{L}_{align}$
$\mathcal{L}_{cls}$
Classifier
Φ_c
Causal Graph Extraction
Feature attributions

Assumptions

Assumption 1: *The latent space of the feature extractor can be split into two sets: (i) encoding the Observed Context feature and other (ii) encoding Unobserved latent features.*

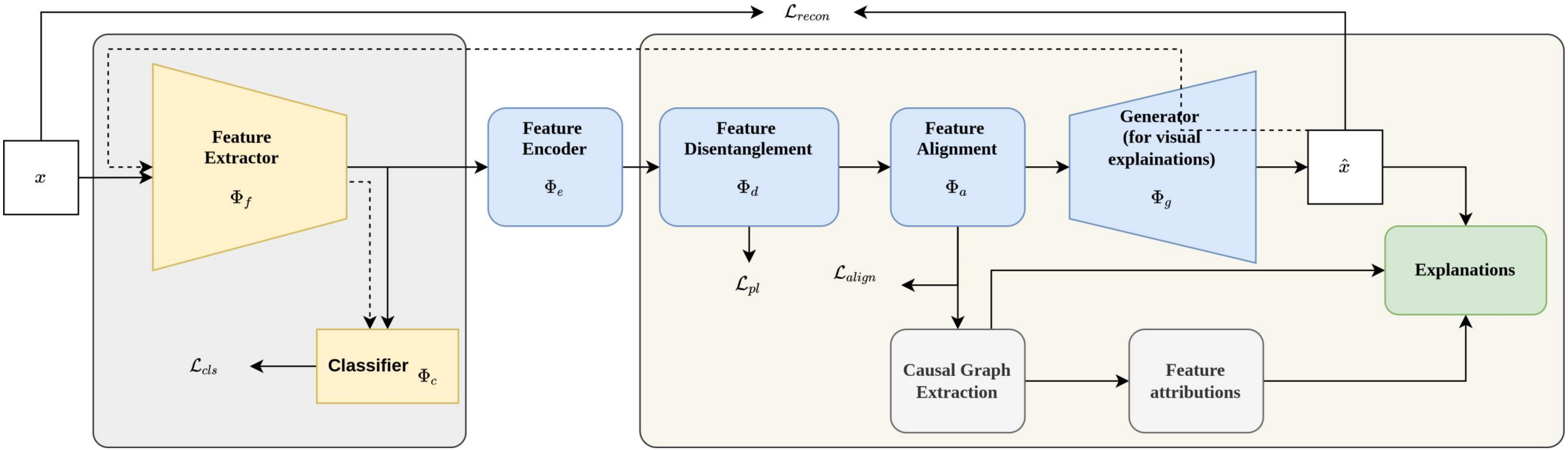
Assumption 2: *We assume all unobserved latent features to be independent of one another.*

Alignment

Properties:
(i) Subspace optimization for context features
(ii) Selective feature disentanglement

Definition 1. *The alignment of latent subspace to observed context features can be achieved by minimizing the L2 distance between the subspace of latent features and ground-truth context features. This corresponds to* $||\mathcal{E}' - \mathcal{C}||_2^2$*, constrained on* $z_i \perp z_j$*, where* $z_i, z_j \in \mathcal{E}'_u$ *and* $i \neq j$ $\left(\left\| M - \frac{\lambda_{max} \hat{U\Sigma}}{||\Sigma||_f} \right\|_2^2 \right)$

Framework



$\mathcal{L}_{recon}$
x
Feature Extractor
Φ_f
Feature Encoder
Φ_e
Feature Disentanglement
Φ_d
Feature Alignment
Φ_a
Generator (for visual explainations)
Φ_g
$\hat{x}$
Explanations
$\mathcal{L}_{pl}$
$\mathcal{L}_{align}$
Classifier
Φ_c
$\mathcal{L}_{cls}$
Causal Graph Extraction
Feature attributions

Causal Discoveries

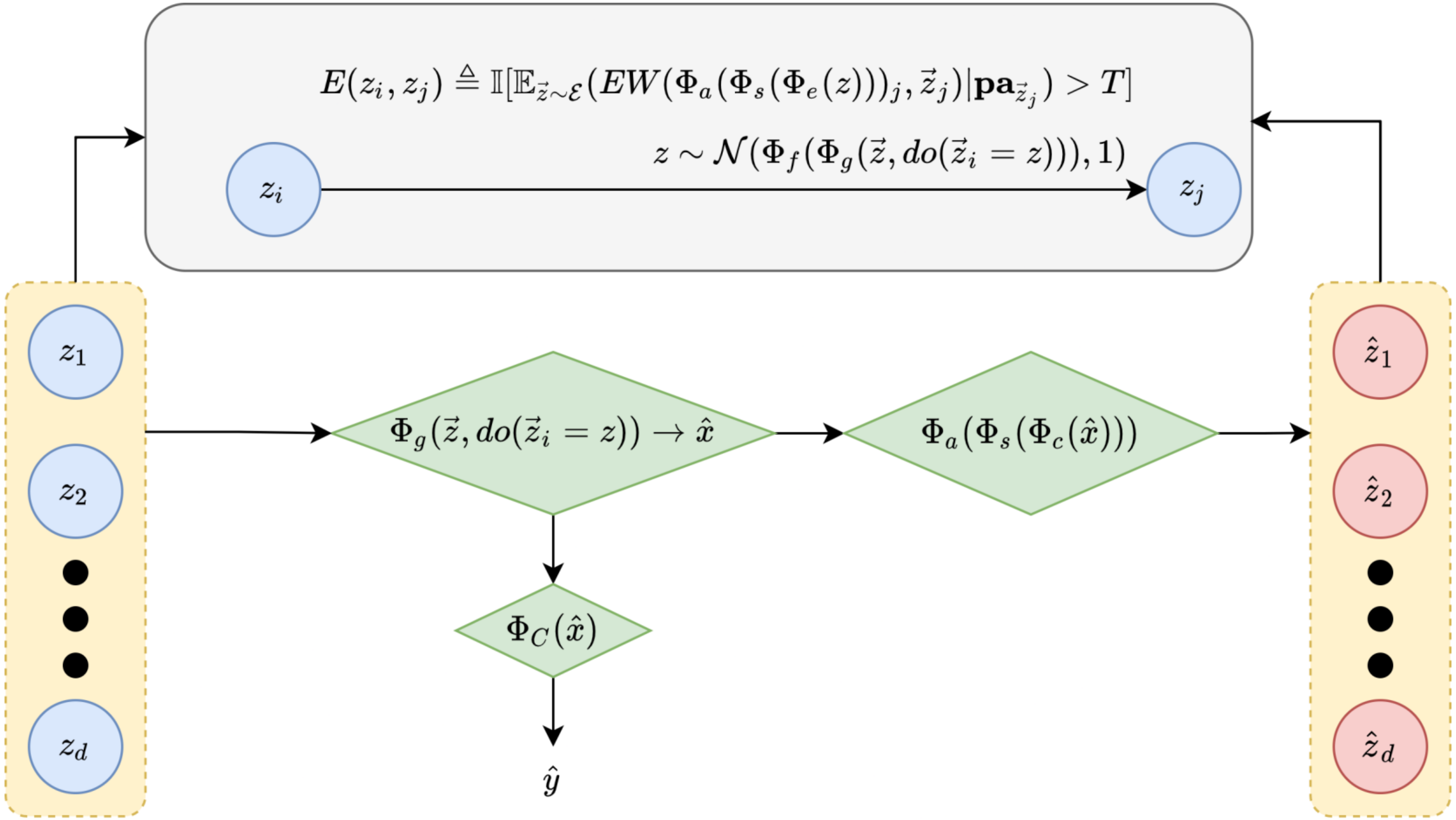
Why do we need causal Discoveries for explanations?

- Understand model perceived data generating process
- Control and decide interventional variables for generating counterfactual explanations

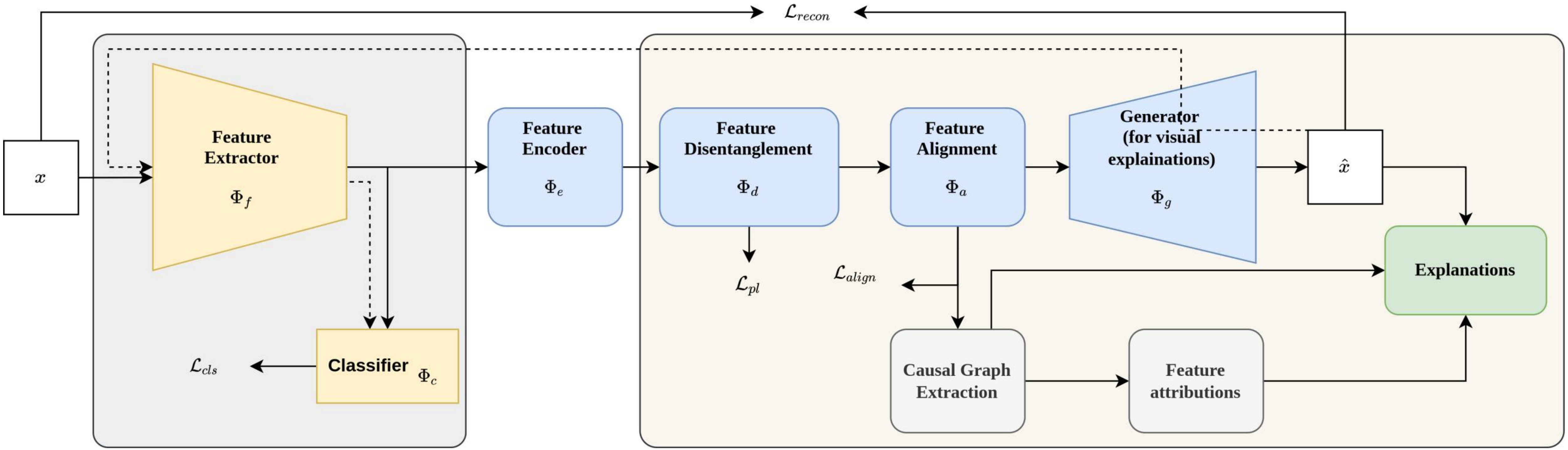
Causal Discoveries

Definition 2. *Node* n_i *and node* n_j *in a DAG are said to have a* Direct Causal Path (DCP) *if there exists an edge between* n_i *and* n_j *(either* $n_i \rightarrow n_j$ *or* $n_j \rightarrow n_i$*), and are said to have an* Indirect Causal Path (ICP) *if their exists a trail from* n_i *to* n_j *via a third node* n_k *(either* $n_i \rightarrow n_k \rightarrow n_j$ *or* $n_j \rightarrow n_k \rightarrow n_i$*). Finally, we define the* edge-weight *for an edge between* n_i *and* n_j *as:*

Causal Discoveries



Framework



$\mathcal{L}_{recon}$
x
Feature Extractor
Φ_f
Feature Encoder
Φ_e
Feature Disentanglement
Φ_d
Feature Alignment
Φ_a
Generator (for visual explainations)
Φ_g
$\hat{x}$
$\mathcal{L}_{pl}$
$\mathcal{L}_{align}$
Explanations
$\mathcal{L}_{cls}$
Classifier
Φ_c
Causal Graph Extraction
Feature attributions

Causal Discoveries

Intervention on X:
X -> Y, with $Y=y_x$
X -> Z, with $Z=z_x$

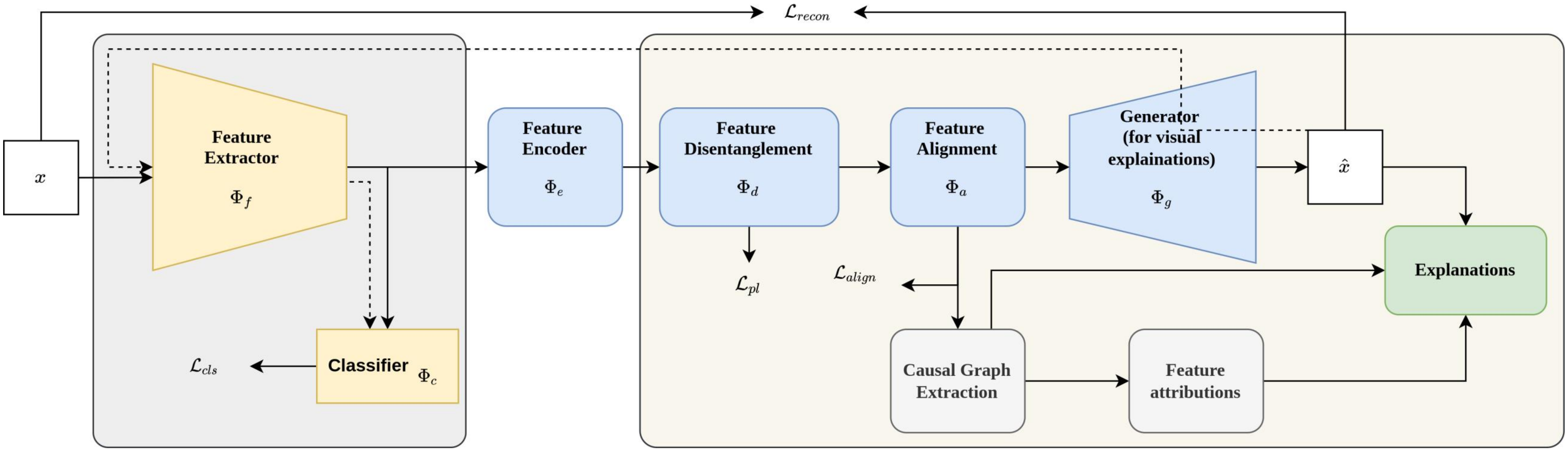
Algorithm 1 Causal discovery algorithm

Input Dataset $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$, Trained model $\Phi_q, \Phi_e, \Phi_d, \Phi_f$
Compute $z \sim \mathcal{N}(\Phi_f(x), 0.1), x \in \mathcal{X}$
for z_i **in** z **do**
    $ch_{z_i} = HardIntervention(z_i)$ # step 2 in 3.3
    $ch_{z_i} = RemoveSpurious(z_i, ch_{z_i})$ # step 3 in 3.3
end for

Controlled intervention: $do(Y=y_x)$
Measure change: $Z=z_y$
Check: $|z_y - z_x|$

Example: X -> Y -> Z

Framework


$\mathcal{L}_{recon}$
x
Feature Extractor
Φ_f
Classifier
Φ_c
$\mathcal{L}_{cls}$
Feature Encoder
Φ_e
Feature Disentanglement
Φ_d
$\mathcal{L}_{pl}$
Feature Alignment
Φ_a
$\mathcal{L}_{align}$
Generator (for visual explainations)
Φ_g
$\hat{x}$
Explanations
Causal Graph Extraction
Feature attributions

Explanations

- **Global Explanations:**
  - Perceived data generating process

- **Local Explanations:**
  - Latent feature attributions
  - Counterfactual explanations

Evaluations

Graph Evaluation

- Structural Hamming Distance:

  - Distance between estimated and ground truth graph

- Graph Correctness:

  - Stability: measures the variation in discoveries across multiple datasets
  - Consistency: measures the variation in discoveries across multiple runs

$$correctnessIndex \triangleq \frac{1}{PQ} \sum_P \sum_Q \frac{\#CorrectEdgesPredicted - \#AdditionalEdges}{\#TotalEdges}$$

Explanation Evaluation

- Faithfulness:

  - Measures the contribution of model in generating explanation

$$faithfulnessindex = \frac{\mathcal{I}(\mathcal{E}';\mathcal{E}xps)}{\sqrt{\mathcal{H}(\mathcal{E}xps)\mathcal{H}(\mathcal{E}')}}$$

- Stability:

  - Measures variance in the generated explanations for similar images

$$stability(\mathcal{E}xps) \triangleq -\frac{1}{P}\Sigma_P \mathbb{E}_{x\sim\mathcal{E}xps}((x - \mathbb{E}(x))^2)$$

Experiments

Experiments - Case Study 1: MorphoMNIST

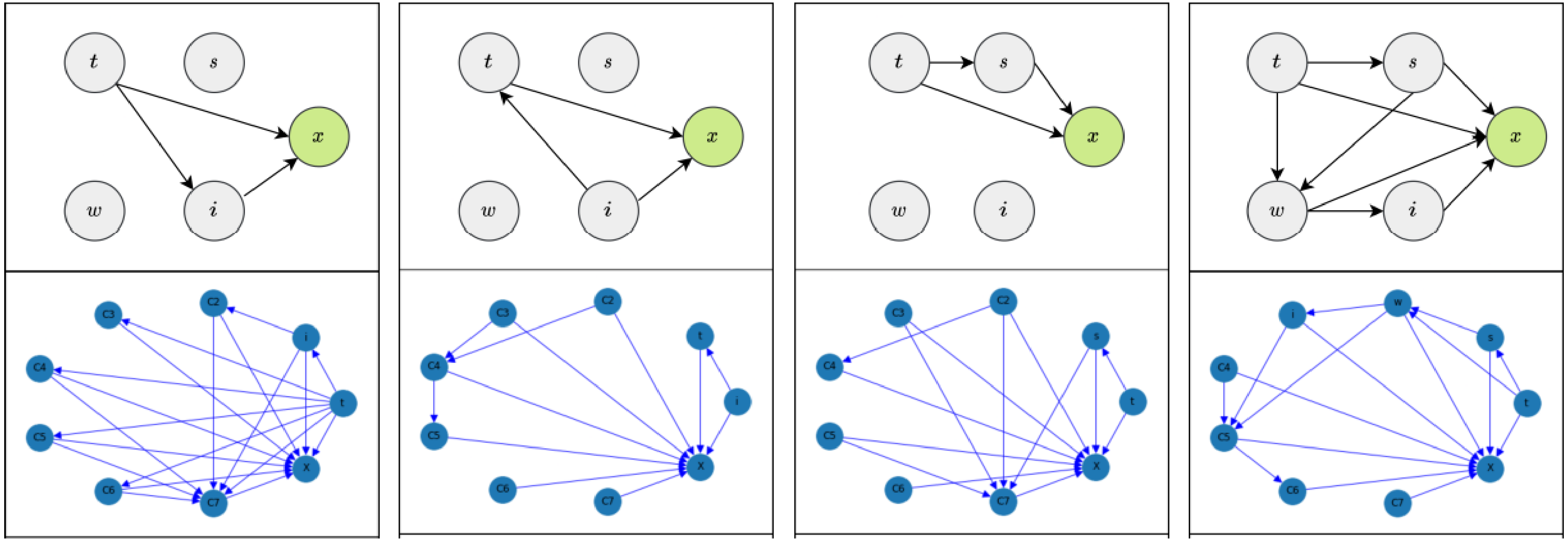
$$t := f_t \triangleq 0.5 + \epsilon_t \quad \epsilon_t \sim \Gamma(10, 5)$$
$$i := f_i \triangleq 64 + 191 * \sigma(2 * w + 5) + \epsilon_i \quad \epsilon_i \sim \mathbb{N}(0, 1)$$
$$x := f_x = SetIntensity(SetThickness(X; t); i)$$

$$i := f_i \triangleq \epsilon_i \quad \epsilon_i \sim \mathbb{U}(60, 255)$$
$$t := f_t \triangleq 3 + \sigma(i/255) + \epsilon_s \quad \epsilon_s \sim \mathbb{N}(0, 0.5)$$
$$x := f_x = SetThickness(SetIntensity(X; i); t)$$

$$t := f_t \triangleq \epsilon_t \quad \epsilon_t \sim \Gamma(0, 5)$$
$$s := f_s \triangleq 10 + 5 * \sigma(2 * t - 5) + \epsilon_s \quad \epsilon_s \sim \mathbb{N}(0, 0.5)$$
$$x := f_x = SetSlant(SetThickness(X; t); s)$$

$$t := f_t \triangleq \epsilon_t \quad \epsilon_t \sim \Gamma(0, 5)$$
$$s := f_s \triangleq 10 + 20 * t + \epsilon_s \quad \epsilon_s \sim \mathbb{N}(0, 5)$$
$$w := f_w \triangleq 10 + 15 * \sigma(0.5 * t) - 0.25 * s + \epsilon_w \quad \epsilon_w \sim \mathcal{N}(0, 1)$$
$$i := f_i \triangleq 64 + 191 * \sigma(w/25) + \epsilon_i \quad \epsilon_i \sim \mathbb{N}(0, 1)$$
$$x := f_x = SetIntensity(SetWidth(SetSlant(SetThickness(X; t); s); w); i)$$

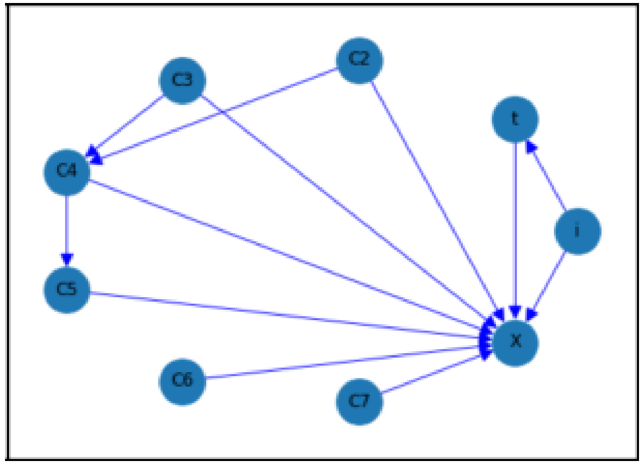
Experiments - Case Study 1: MorphoMNIST



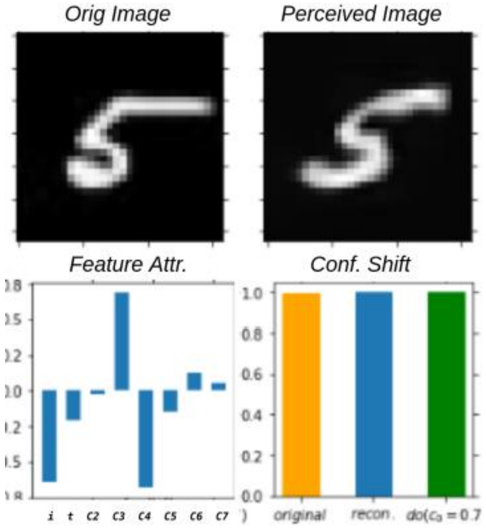
Datasets ↓ \ Methods →	LinGAM Based [36]	GES Based [37]	Ours
Morpho-MNIST (TI)	0.84	0.66	**1.0**
Morpho-MNIST (IT)	0.66	0.66	**1.0**
Morpho-MNIST (TS)	0.82	0.66	**0.98**
Morpho-MNIST (TSWI)	0.58	0.42	**0.94**

Experiments - Case Study 1: MorphoMNIST

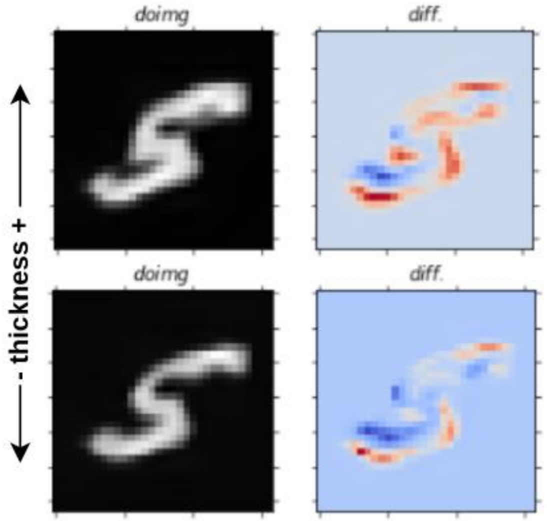
- Explanations



(a)



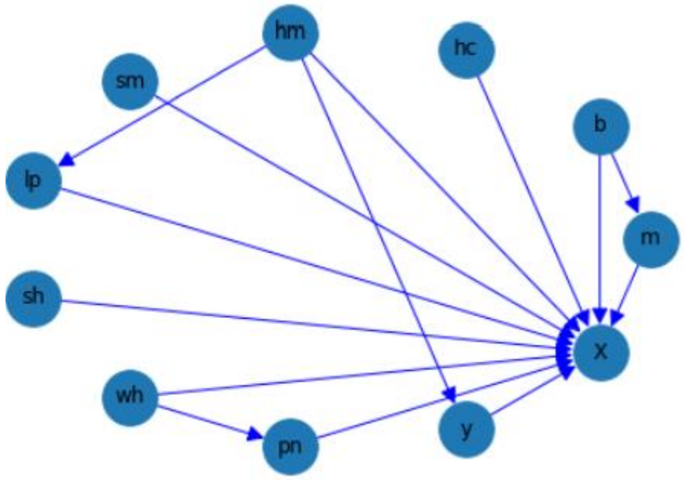
(b)

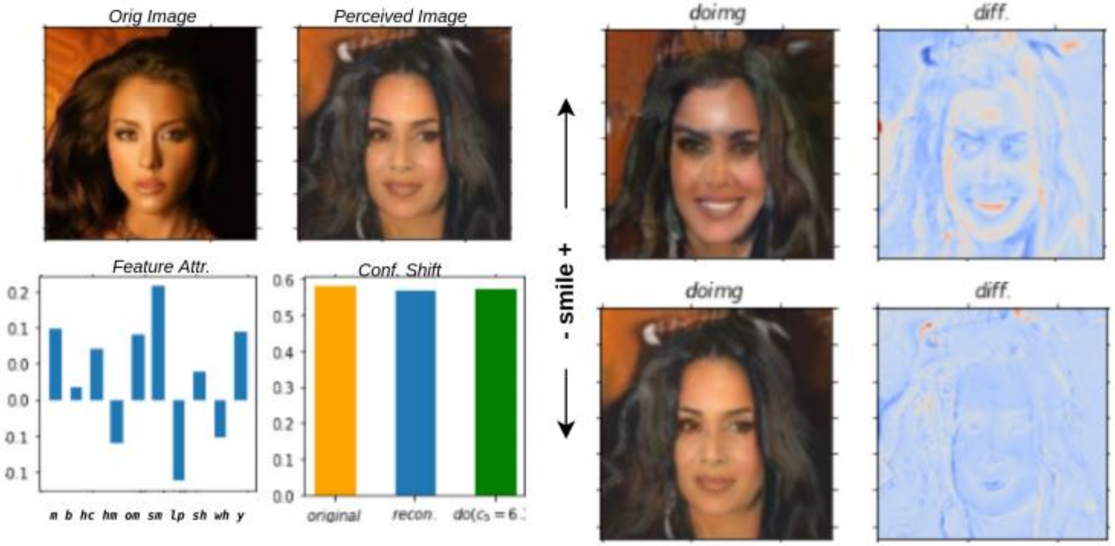


(c)

Experiments - Case Study 2: FFHQ

- FFHQ dataset





Conclusion

- Latent causal discoveries will help in generating better counterfactual explanations.
- Perceived data generating process can be retrieved from the latent classifier's knowledge.

- Future work: ways to expand this approach on datasets without meta information.