DIFFUSION MODELS FOR COUNTERFACTUAL EXPLANATIONS

GUILLAUME JEANNERET, LOIC SIMON AND FREDERIC JURIE

SCHEDULE

- Introduction
- Denoising Diffusion Probabilistic Models
- DiME: Diffusion Models for Counterfactual Explanations
- Results
- Conclusion

INTRODUCTION

INTRODUCTION

- Convolutional Neural Networks (CNN) have reached **unimaginable performances**.

INTRODUCTION

- The CNNs overparameterization makes these architectures **highly uninterpretable**.

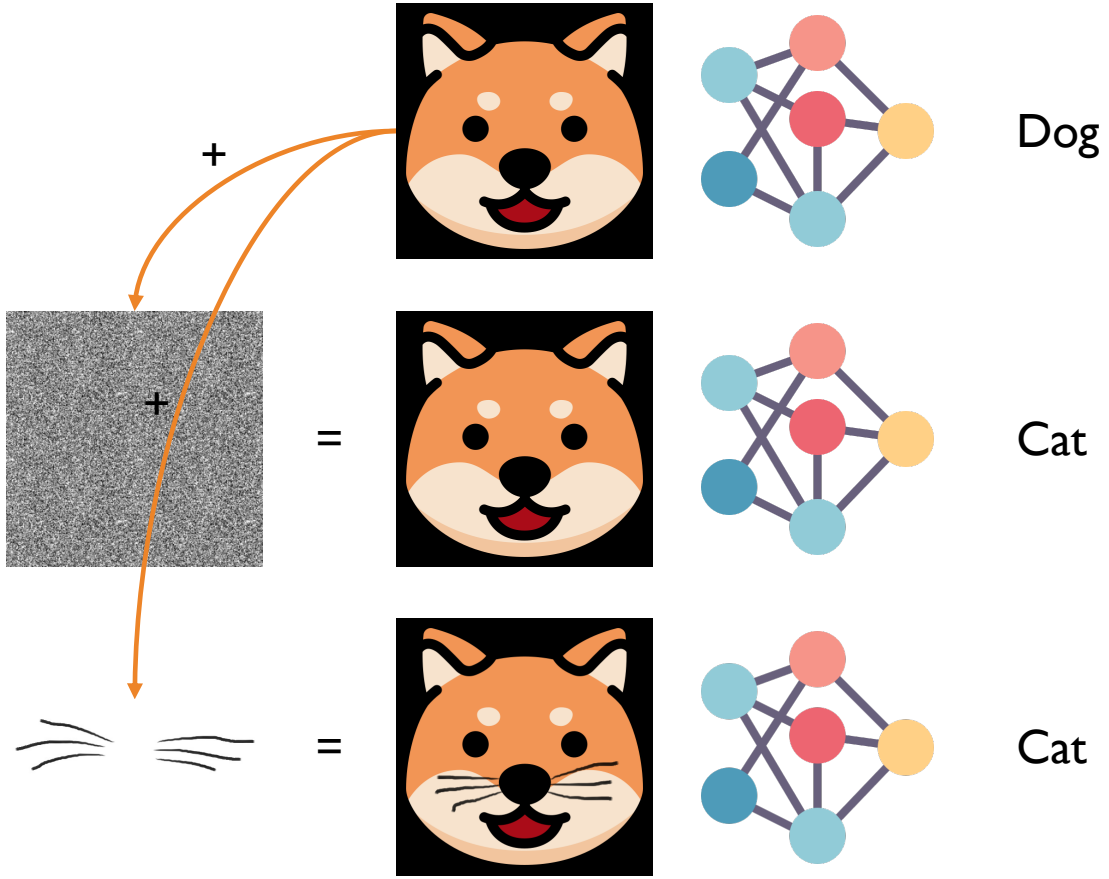
INTRODUCTION

- The research field of Explainable Artificial Intelligence searches to **uncover** the mysteries of within Machine Learning Models.

- In this presentation, we focus on ***Post-Hoc* explainability method**s to analyze black-box such as CNNs. In particular, we concentrate on the growing branch of **Counterfactual Explanations (CE)**.

WHAT ARE COUNTERFACTUAL EXPLANATIONS?

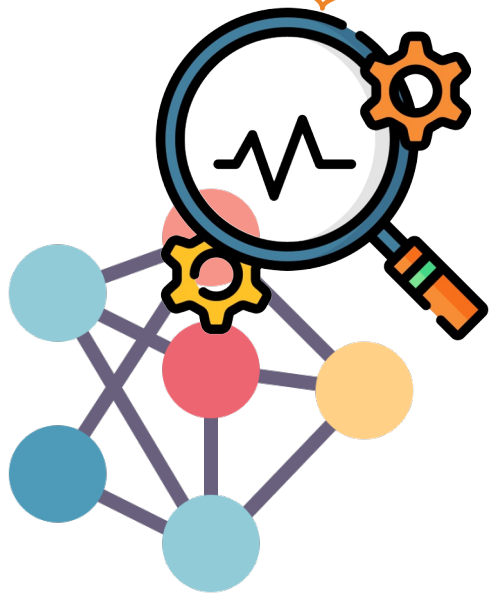
- Counterfactual Explanations perturb an input image to create **a minimal but meaningful perturbation** to change the original prediction.
  - Same objective as adversarial examples with different restrictions.
- Adversarial Examples: structured noise **imperceptible** to the human eye.
- Counterfactual Explanations: ***Plausible* and understandable** modifications.



COUNTERFACTUAL EXPLANATIONS' OBJECTIVES

1. Find variables used for the prediction.

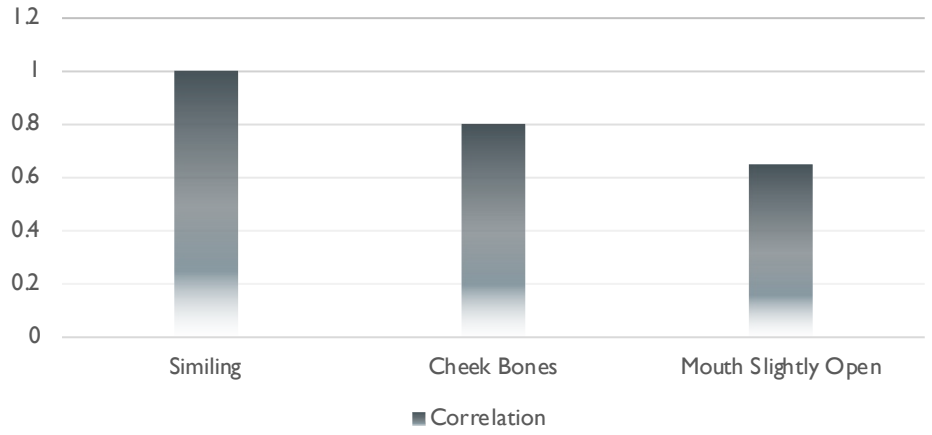
Cats:
- ✓ Whiskers
- ✓ Smaller
- ✓ Pointy Ears
- ✓ Prefer small talks



Dogs:
- ✓ Big mouth
- ✓ Larger
- ✓ Longer face
- ✓ Extrovertive

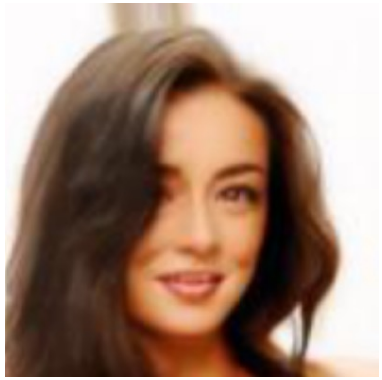
COUNTERFACTUAL EXPLANATIONS' OBJECTIVES

1. Find variables used for the prediction.
2. Uncover **hidden spurious correlations**.



Not Smiling

Smiling



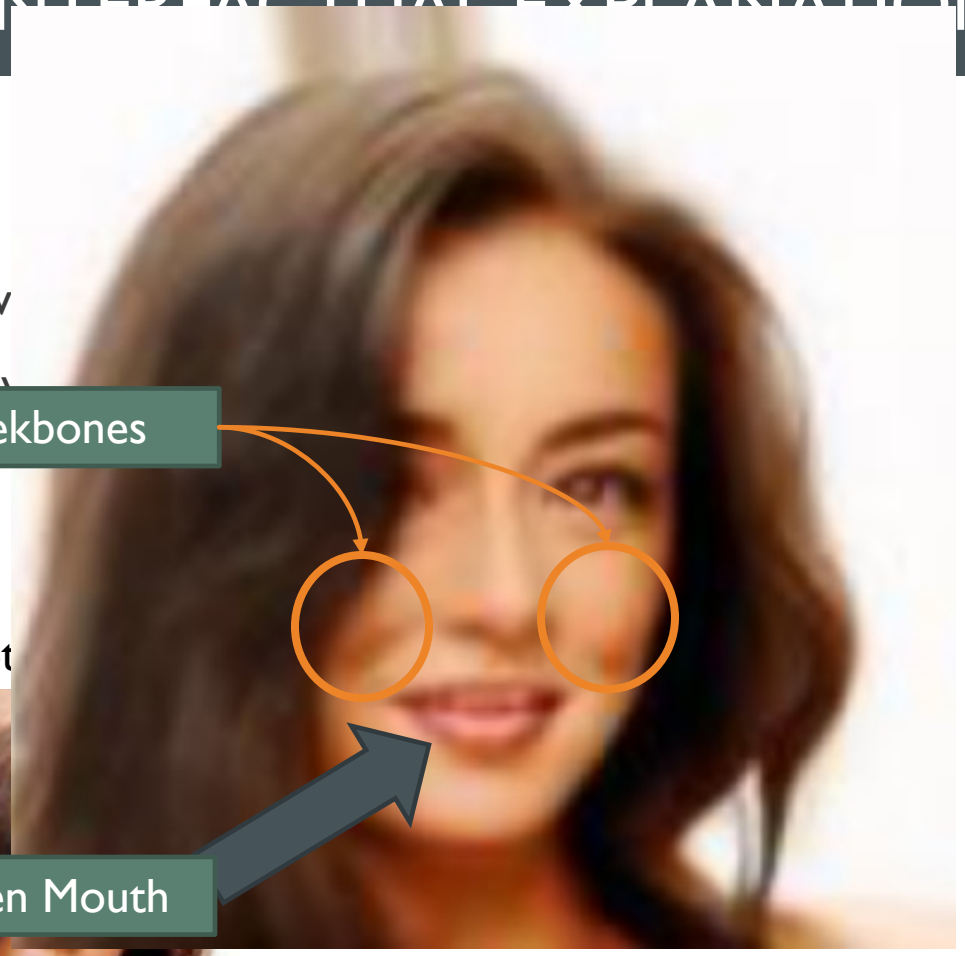
Smiling

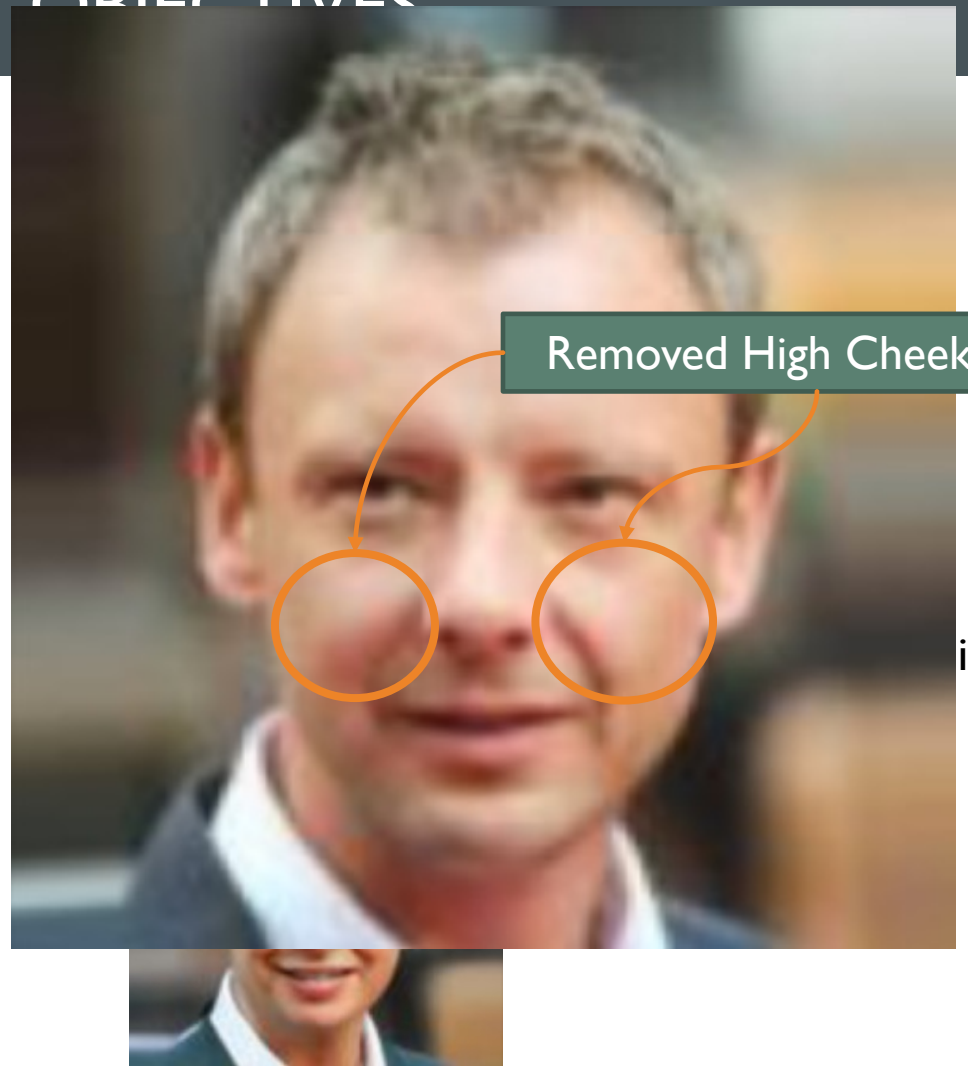
Not Smiling

COUNTERFACTUAL EXPLANATIONS' OBJECTIVES

1. Find v
2. Uncov

Not iling

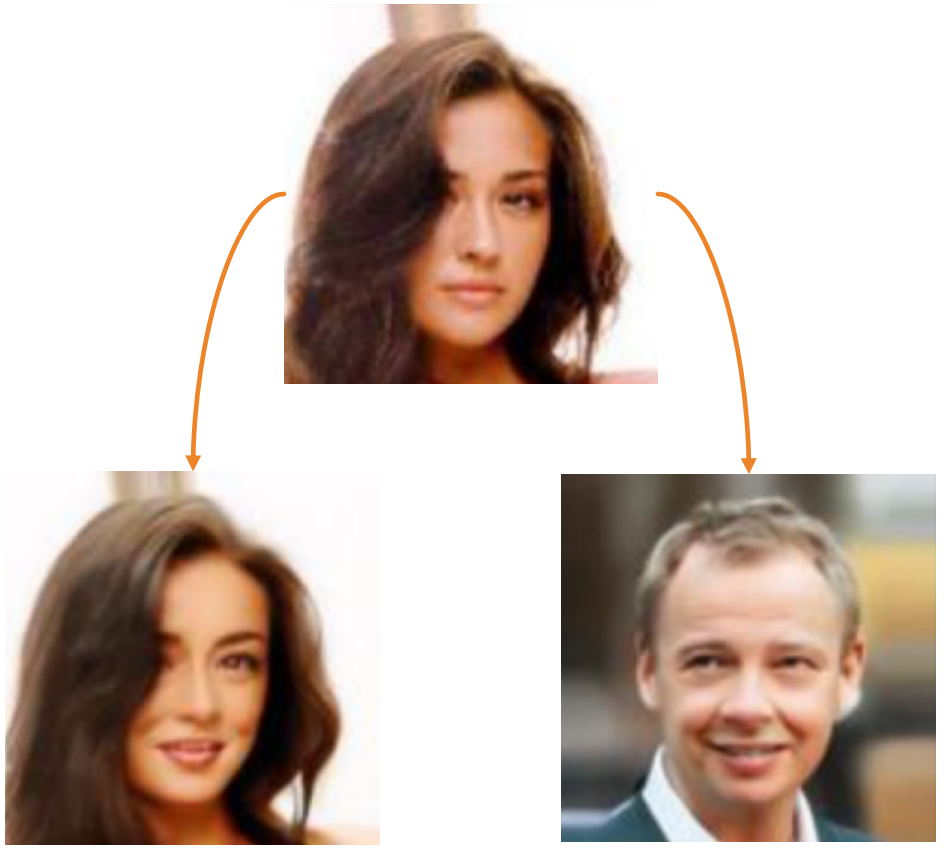




COUNTERFACTUAL EXPLANATIONS' OBJECTIVES

1. Find variables used for the prediction.
2. Uncover **hidden spurious correlations**.

- Desired Properties:
    1. **Proximal** explanations with **sparse** modification.

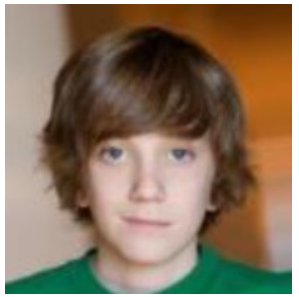


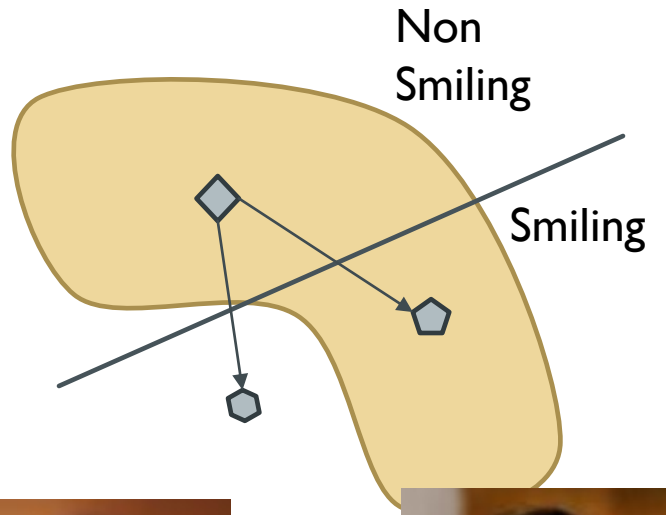
True Counterfactual Explanation

Trivial Counterfactual Explanation

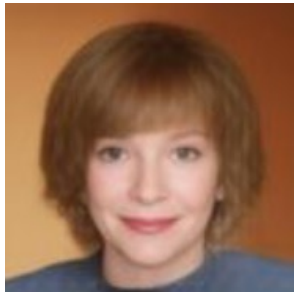
COUNTERFACTUAL EXPLANATIONS' OBJECTIVES

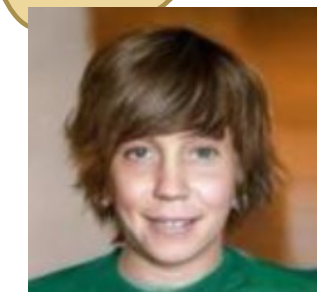
1. Find variables used for the prediction.
2. Uncover **hidden spurious correlations**.





- Desired Properties:
  1. **Proximal** explanations with **sparse** modification.
  2. **Realistic** and understandable by human.



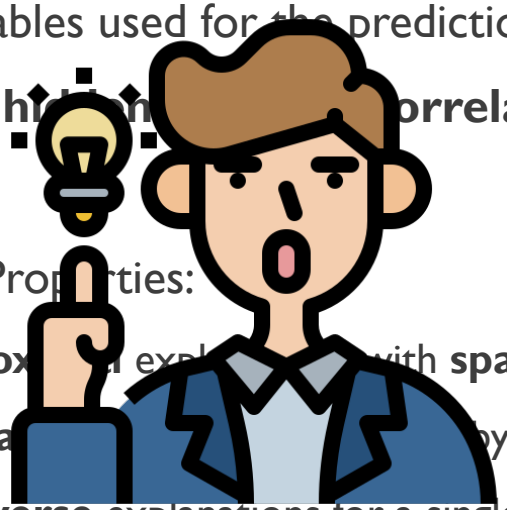


Particularly important in high risk areas!

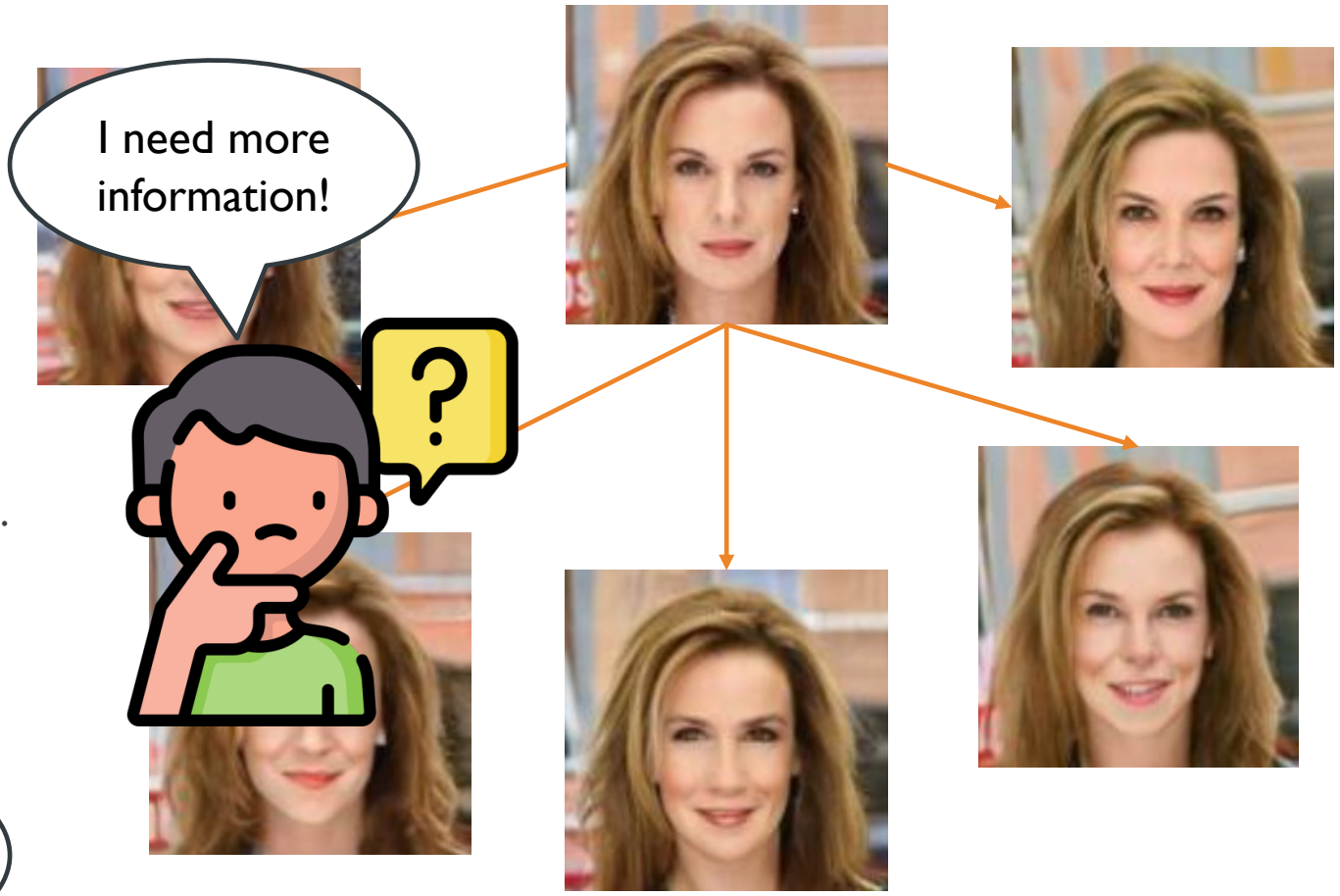
COUNTERFACTUAL EXPLANATIONS' OBJECTIVES

1. Find variables used for the prediction.
2. Uncover **hidden correlations**.

- Desired Properties:
  1. **Proximity** explanations with **sparse** modification.
  2. **Real** explanations by human.
  3. **Diverse** explanations for a single input.



Now I understand



CONTRIBUTIONS

- Three main contributions:

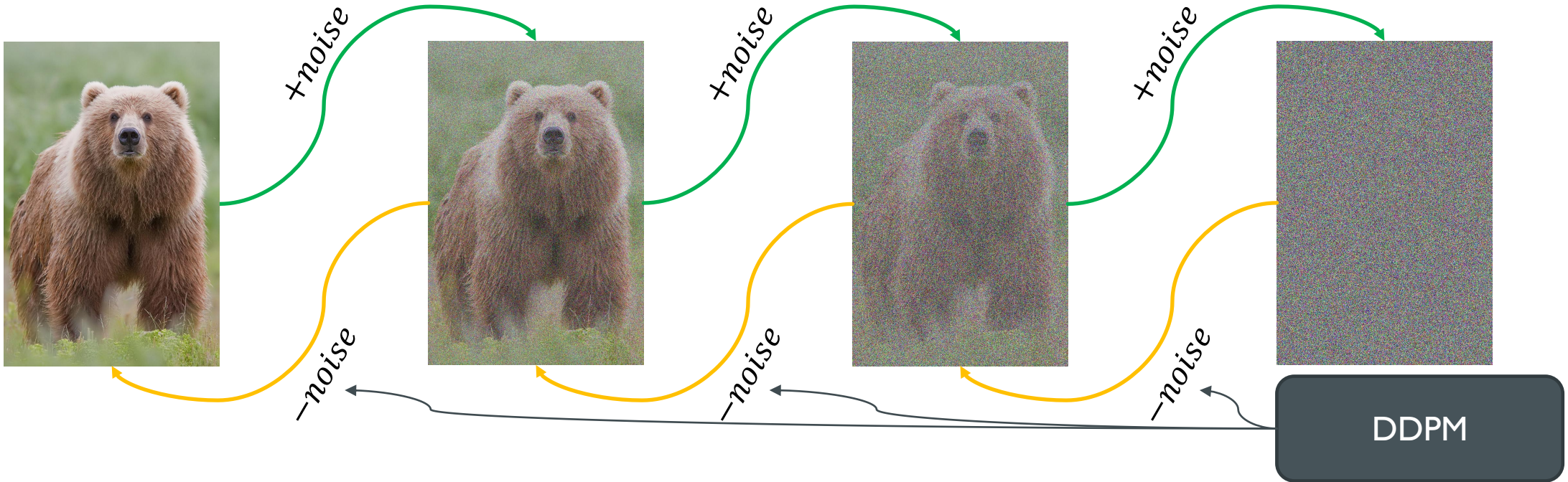
1. DiME uses the **recent diffusion models to generate counterfactual examples**.
2. We set a **new State-of-the-Art** result on a standard benchmark.
3. We **introduce** a new metric to evaluate subtle spurious correlations detection.

DENOISING DIFFUSION PROBABILISTIC MODELS

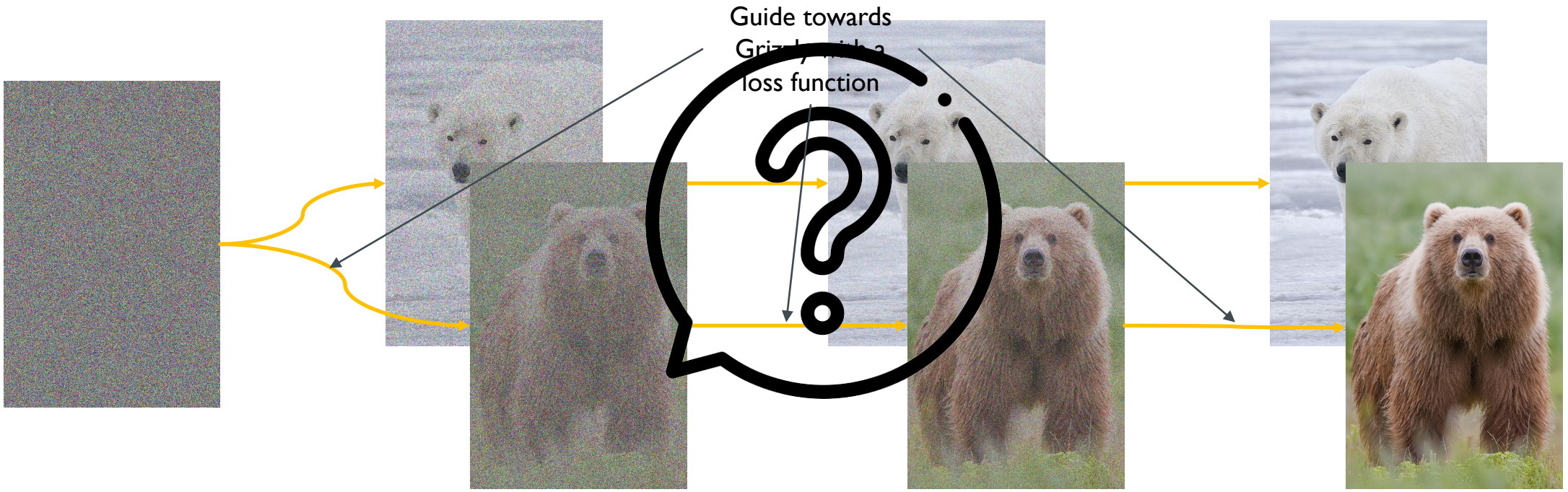
DDPM OR DIFFUSION MODELS

DIFFUSION MODELS

- Objective: **generate** an image from random noise.
- DDPMs rely on **two inverse processes**: Forward and Backward.
- The DDPM is **trained** so the backward process matches the forward.



GUIDED DIFFUSION



- There are **multiple way** to generate an image. How may we know **which** image it will produce?
- The work of Dhariwal and Nichol [1] proposes the **Guided Diffusion:** Each denoising step can be seen as an **optimization step guided with a classifier** trained on noisy images.

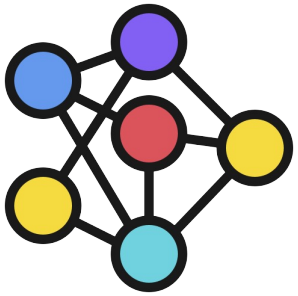
[1] Dhariwal, P. and Nichol, A. Diffusion Models Beat GANs on Image Synthesis. NeurIPS 2021.

DIFFUSION MODELS FOR COUNTERFACTUAL EXPLANATIONS

DIME

DIME: DIFFUSION MODELS FOR COUNTERFACTUAL EXPLANATIONS

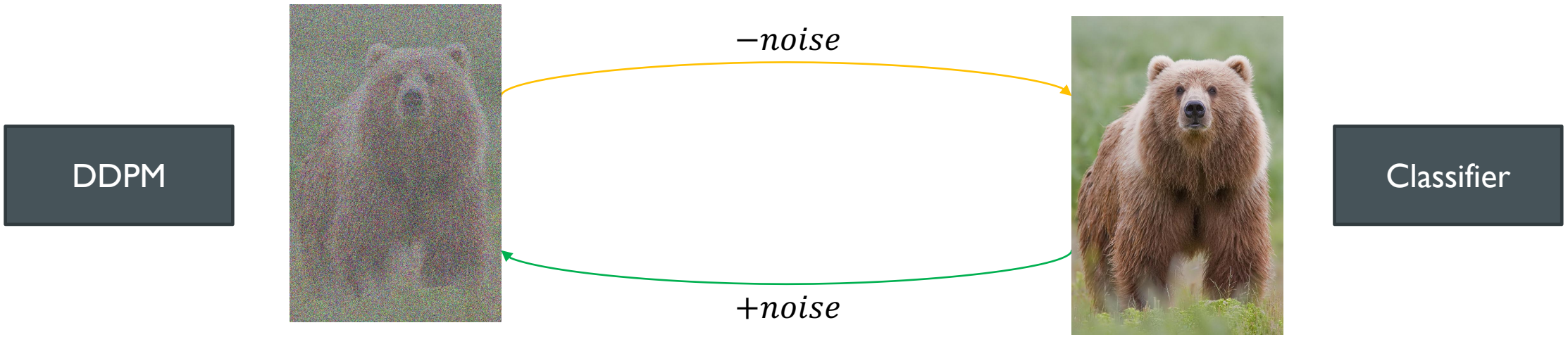
- Guided Diffusion requires **particular** classifier to operate: a classifier **trained on noisy instances**.
  - **Substantial limitation** in the context of CEs.
  - Traditional classifiers do **not perform well** on noisy images.



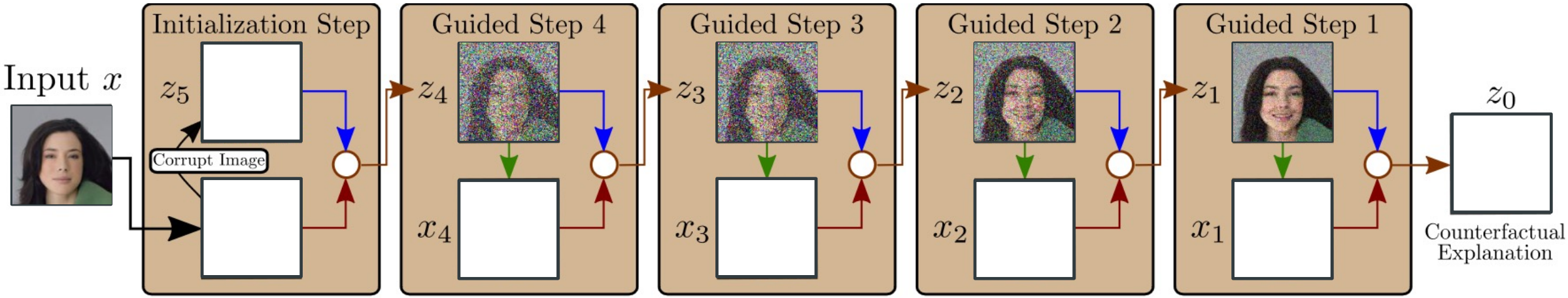
Particular	Traditional
Grizzly	Grizzly
Grizzly	Giraffe
Grizzly	Piano

DIME: DIFFUSION MODELS FOR COUNTERFACTUAL EXPLANATIONS

- **How** may we adapt it for the classifier under observation?
- We transfer the information from the clean instance given that the Forward and Backward processes are approximately the same.

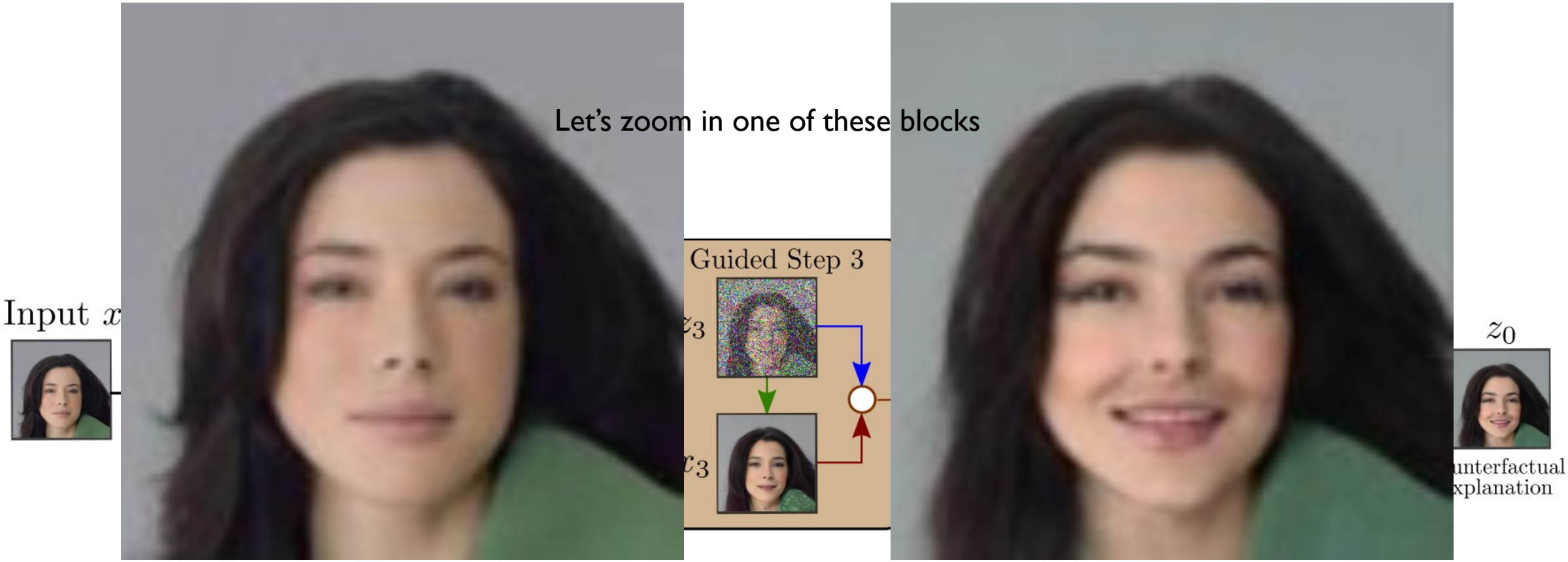


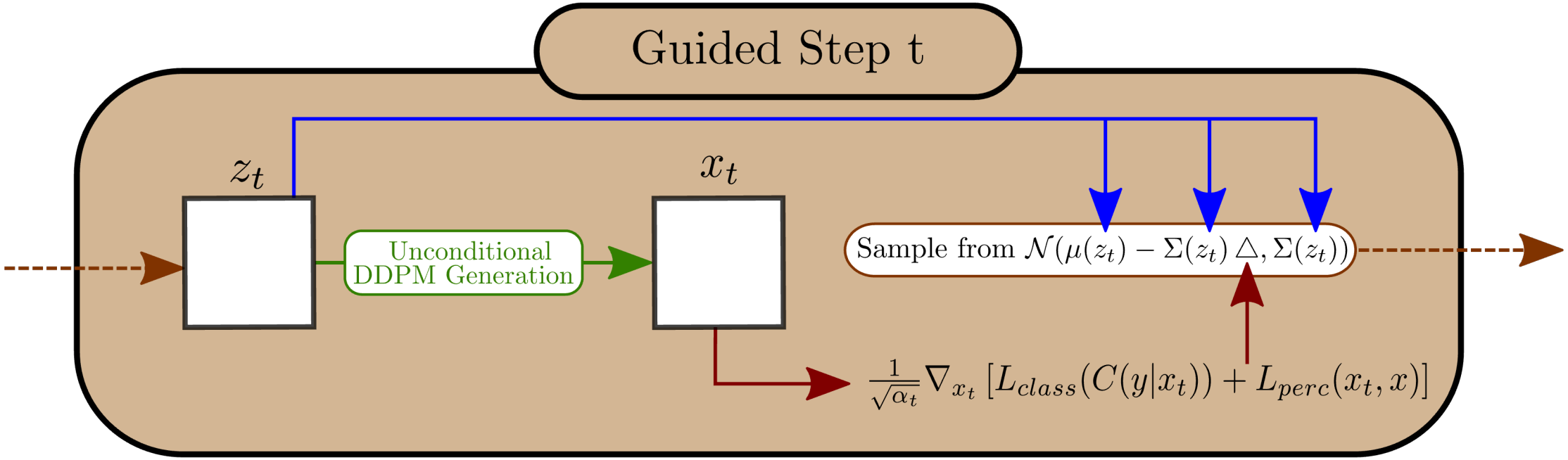
DIME: DIFFUSION MODELS FOR COUNTERFACTUAL EXPLANATIONS



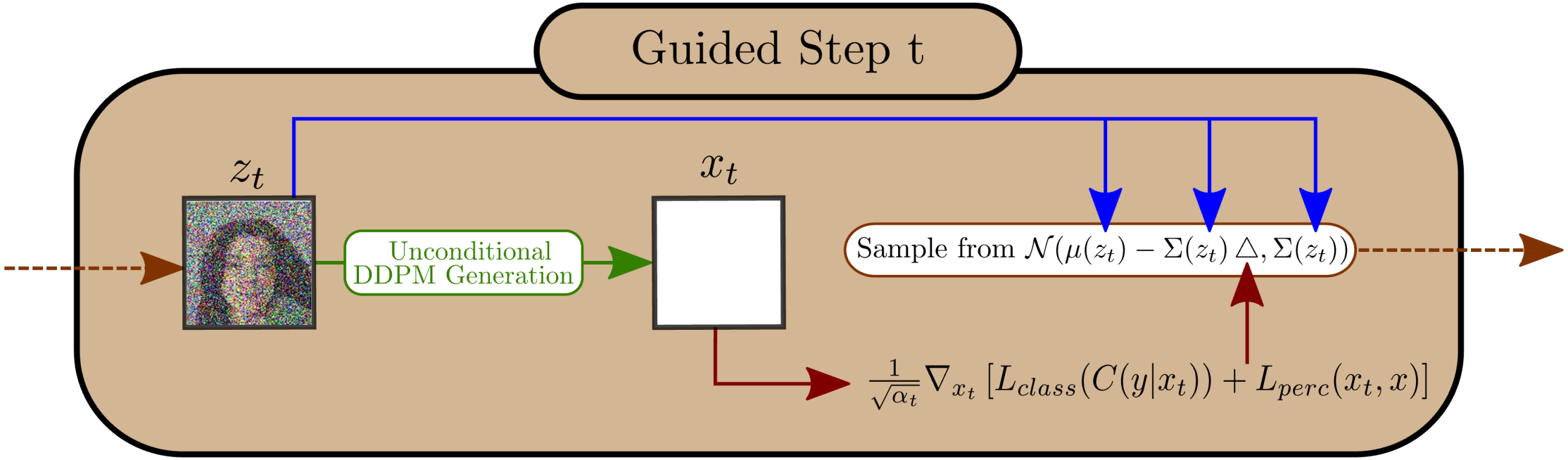
DIME: DIFFUSION MODELS FOR COUNTERFACTUAL EXPLANATIONS

Let's zoom in one of these blocks

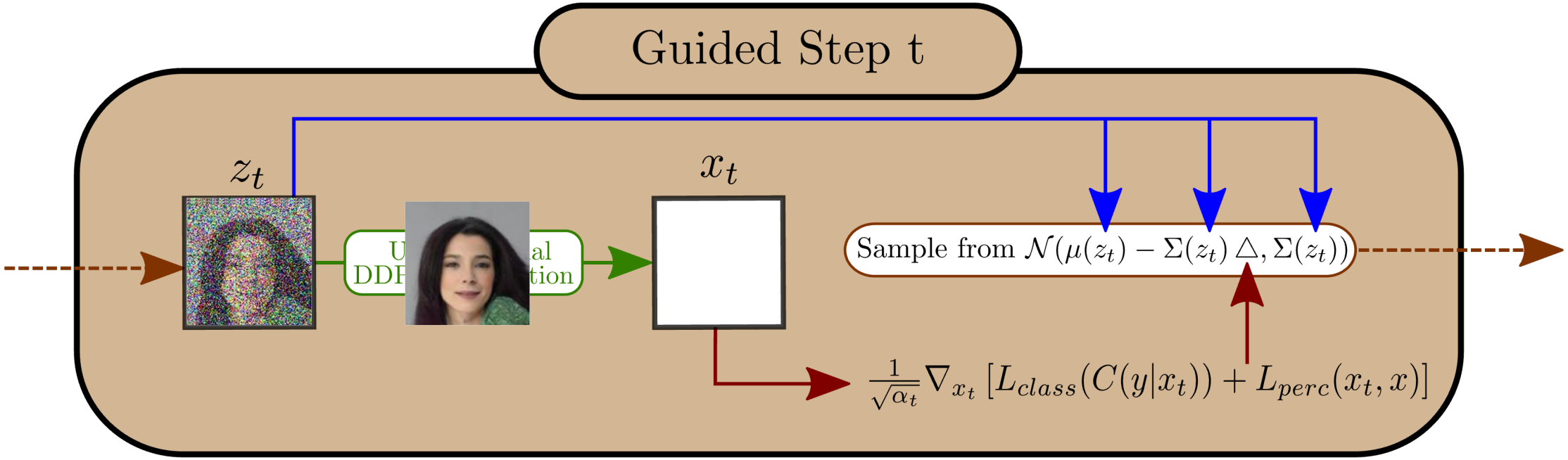




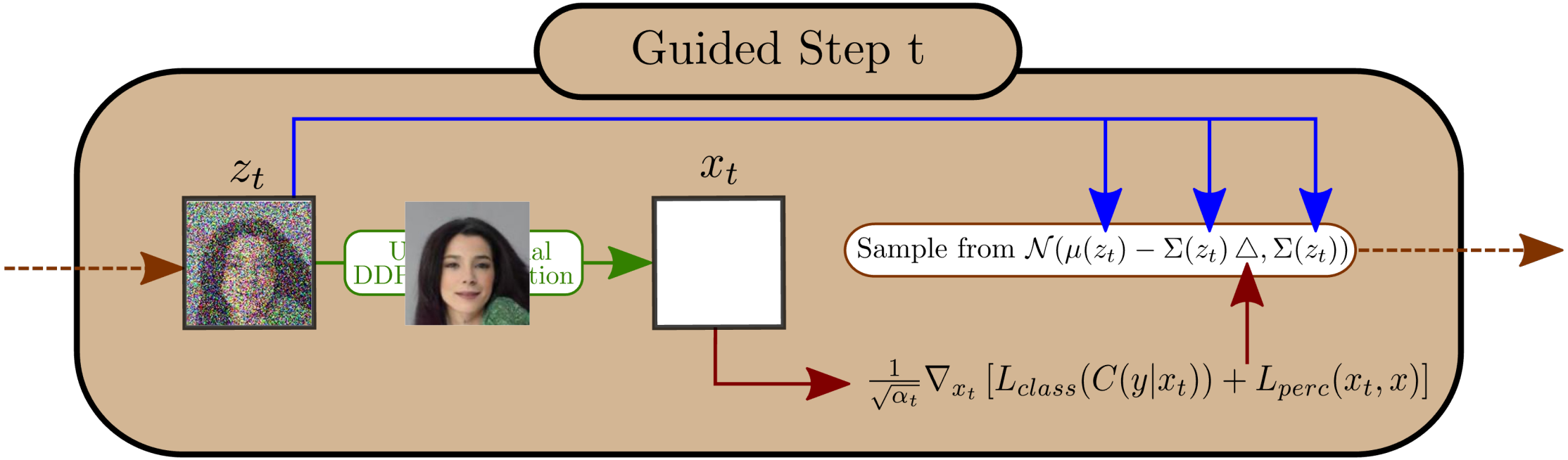
We begin with a noisy image



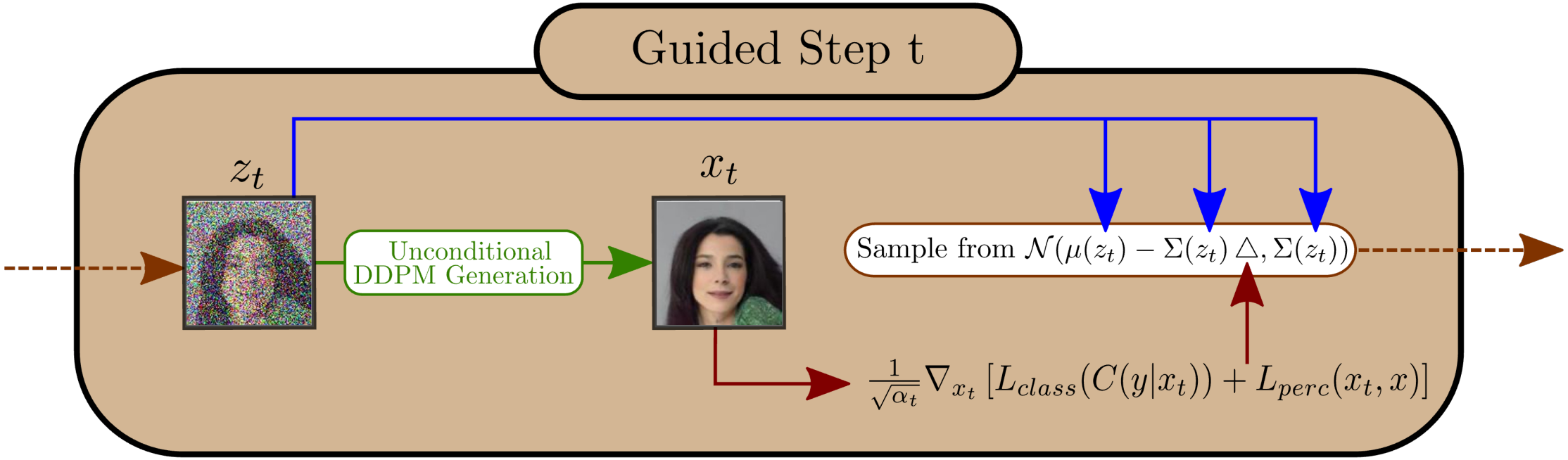
We begin with a noisy image



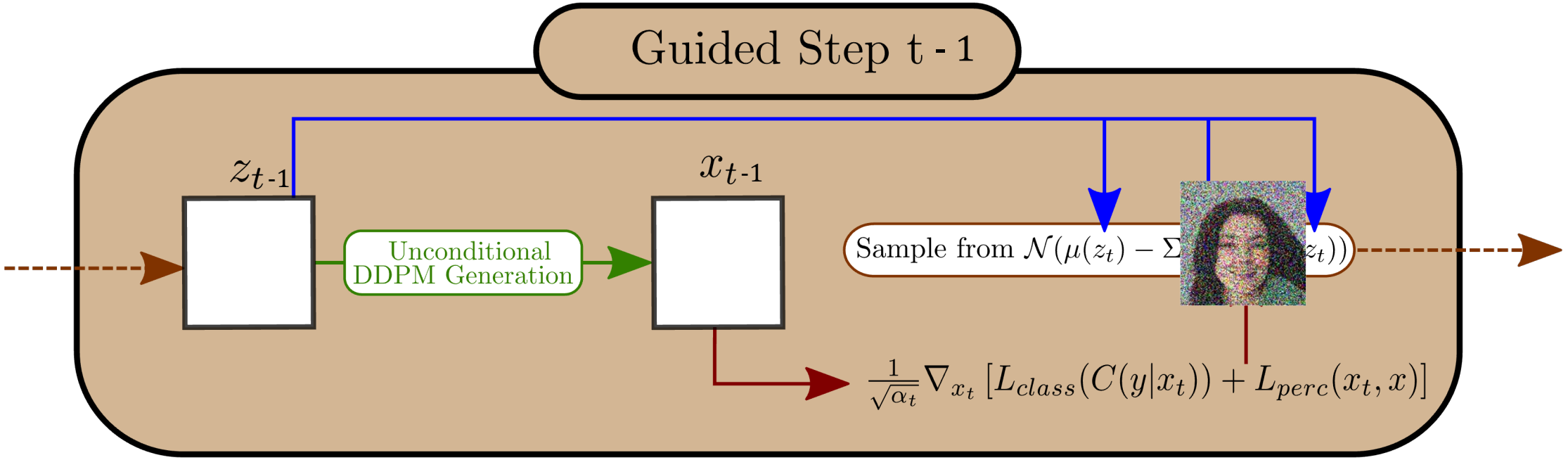
DiME generates a clean image using the iterative DDPM algorithm of the diffusion model using the noisy image as input



DiME generates a clean image using the iterative DDPM algorithm of the diffusion model using the noisy image as input



We sample the next guided noisy image via the **guided diffusion**, and the mean and variance produced by the DDPM



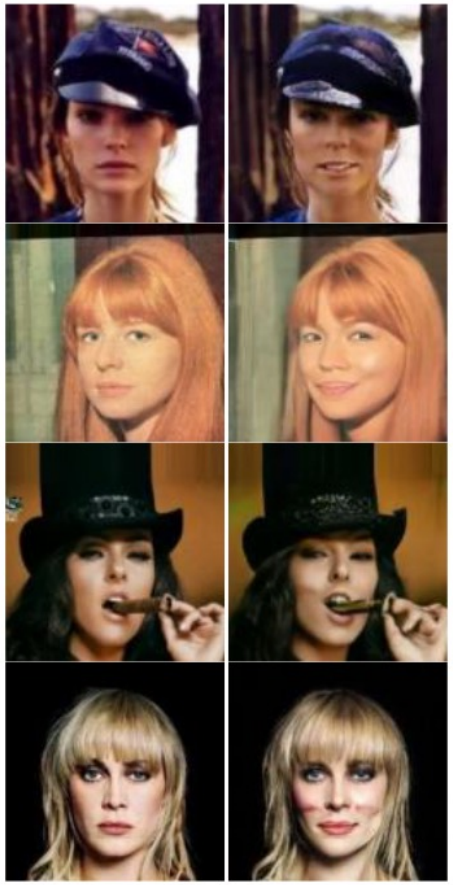
Then, we iterate until $t = 0$

RESULTS

RESULTS

Smile → NS

NS → Smile



Young → Old

Old → Young

Female → Male

Query

↓

DiME

DiVE

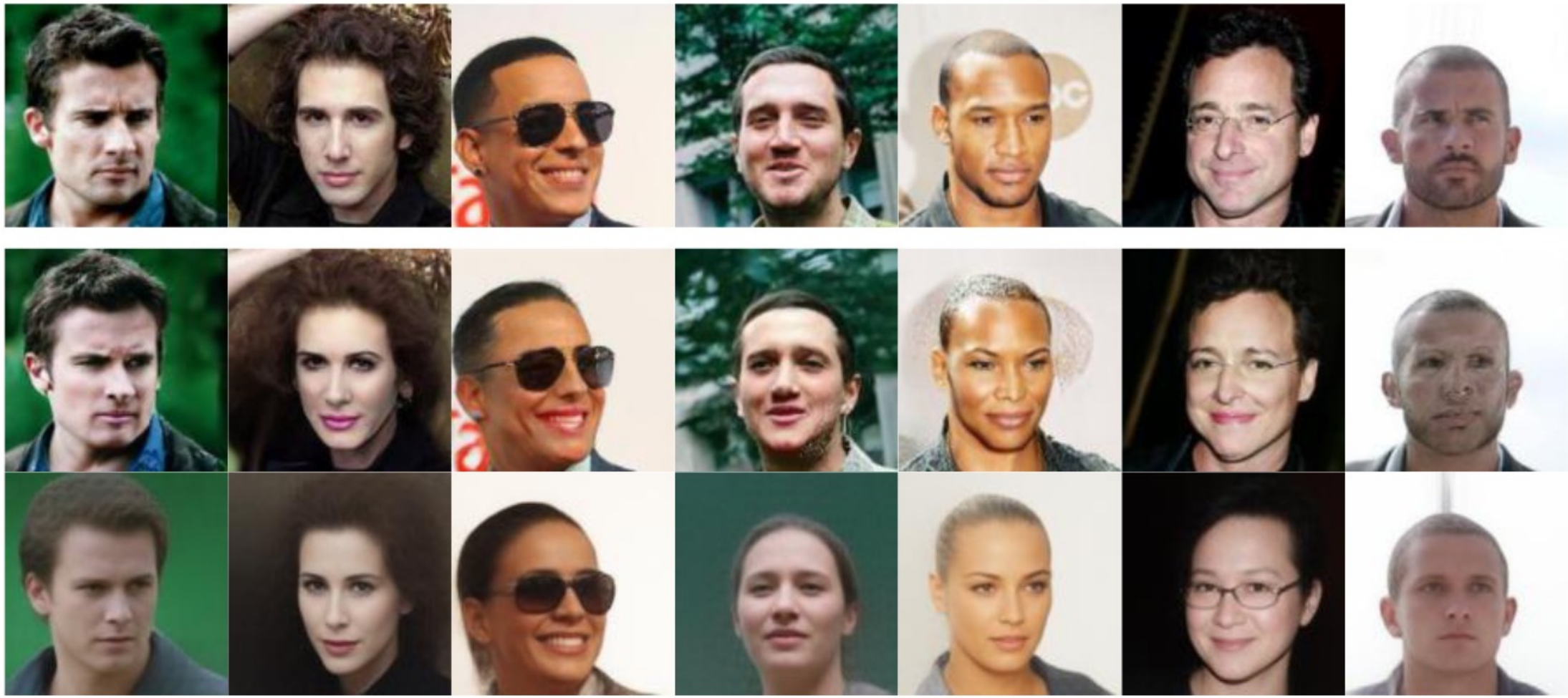
Male → Female

Query

↓

DiME

DiVE



Clear → Blurry

Query

↓

DiME

DiVE

Blurry → Clear

Query

↓

DiME

DiVE

No Eye Bags → Eye Bags

Eye Bags → No Eye Bags

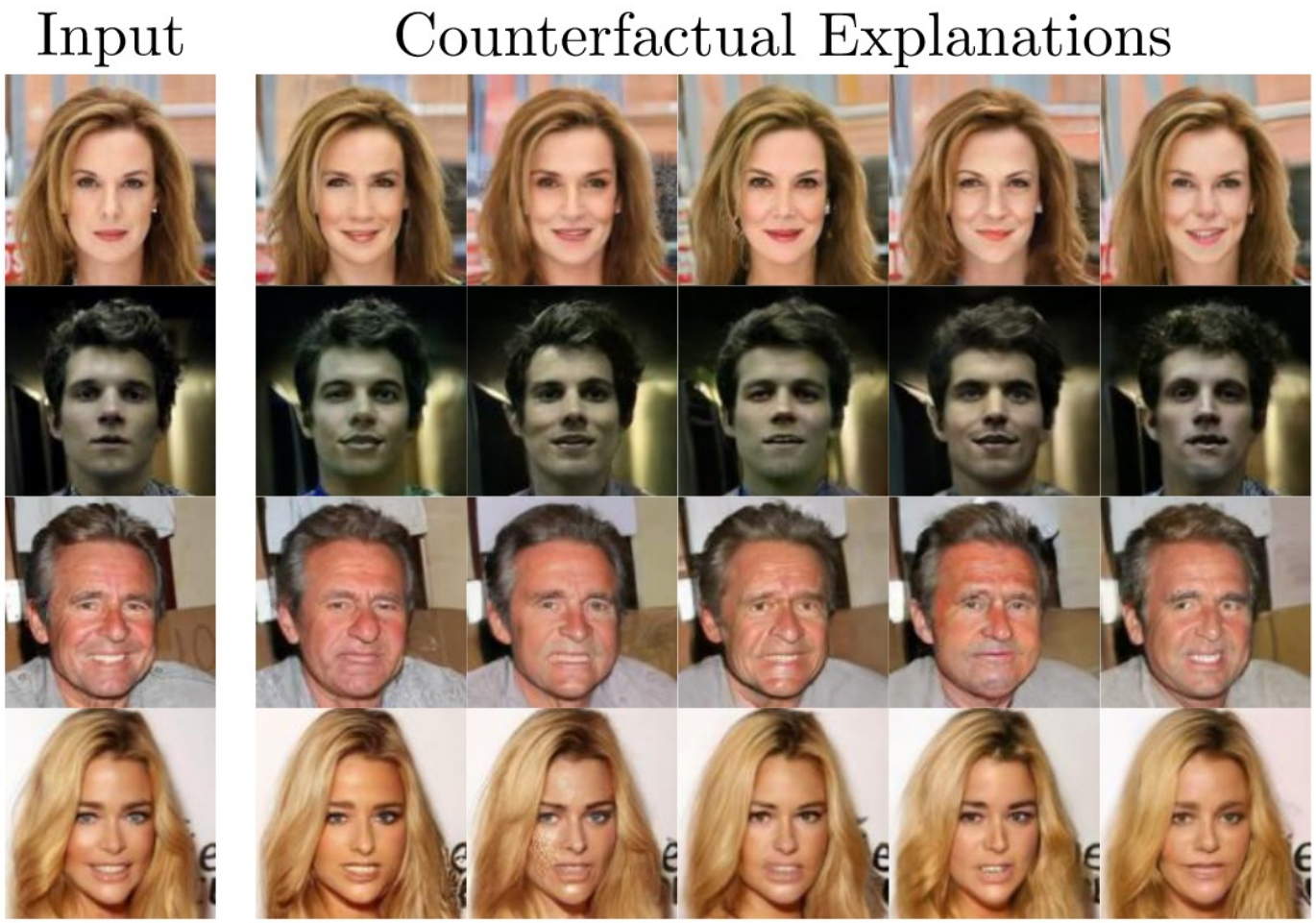
Query

↓

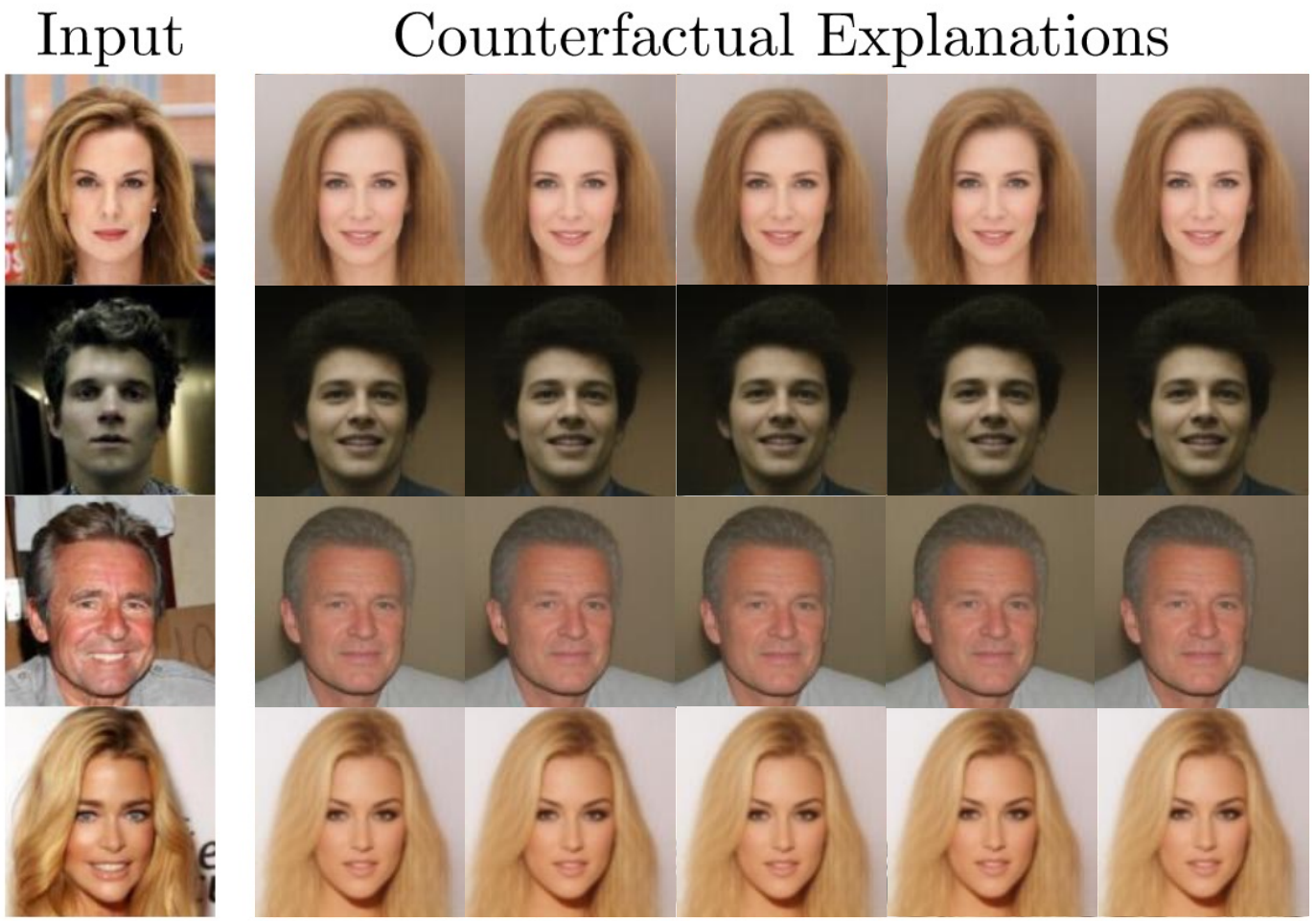
DiME

DiVE

DIVERSITY

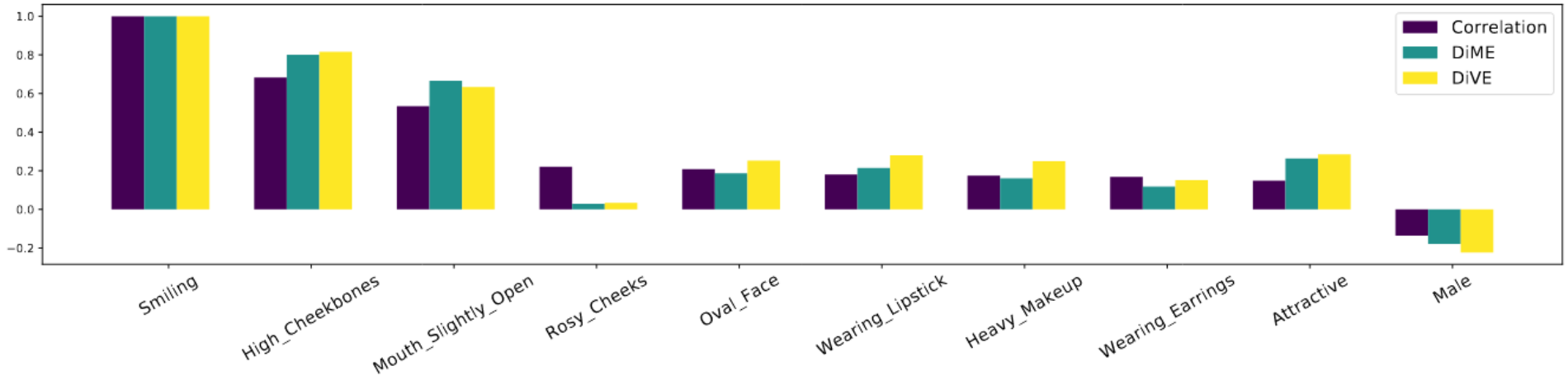


DIVERSITY



A NEW METRIC TO ASSESS SPURIOUS CORRELATION DETECTION

- **False sense of achievement** produced by standard metrics.
- Our model beats previous State-of-the-Art, but by **small margin**.



CONCLUSION

CONCLUSION

- We explore the new Diffusion Models for CEs.
- DiME is able to create diverse, realistic with sparse modifications; features desired for counterfactual explanations.
- We analyze current ways to assess spurious correlations detection and propose a new metric to correctly measure the detection of correlations.

QUESTIONS?

THANK YOU VERY MUCH FOR YOUR ATTENTION!