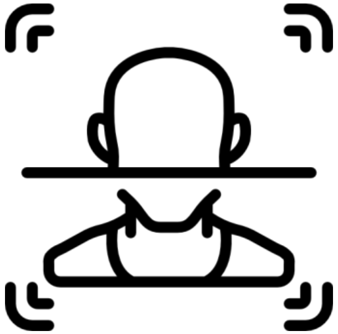
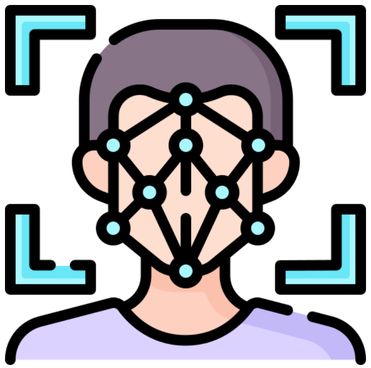
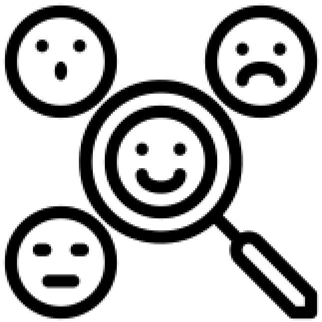
Working with faces can be challenging



Face recognition



Landmark detection



Emotion recognition

Collecting data is hard because of its **privacy sensitive** nature

⇕

You need a very **diverse dataset** to ensure fair and unbiased models

Image manipulation models provide a convenient way to enrichen your data

Existing dataset: FFHQ

Edit prompt: *"a face with grey hair"*

Generated data

Diverse

Privacy-friendly

But what if more is changed than you want?

Edit prompt: *"face with receding hairline"*

Expected changes

+ Receding hairline

Unexpected changes

+ Male features

– Wearing earrings

FACULTY OF ENGINEERING AND ARCHITECTURE

SynRDinBAS @ WACV 2025

TUCSON, ARIZONA • FEB 28 - MAR 4

Exploring Correlated Facial Attributes in Text-to-Image Models: Unintended Consequences in Synthetic Face Generation

Sander De Coninck, Sam Leroux, Pieter Simoens

GHENT UNIVERSITY

imec

IDLab INTERNET & DATA LAB

Flanders AI Research Program

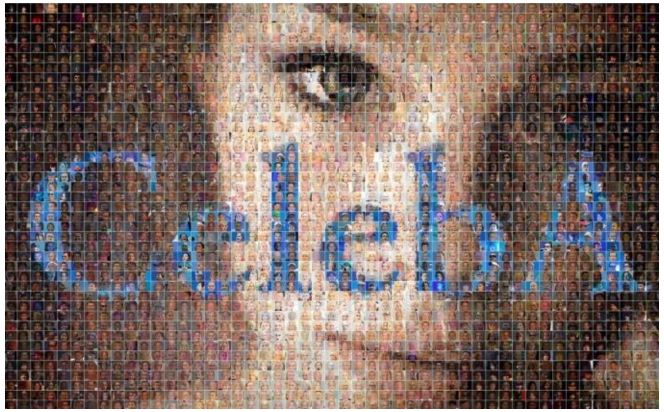
We think these unintended changes are related to inherent dataset correlation

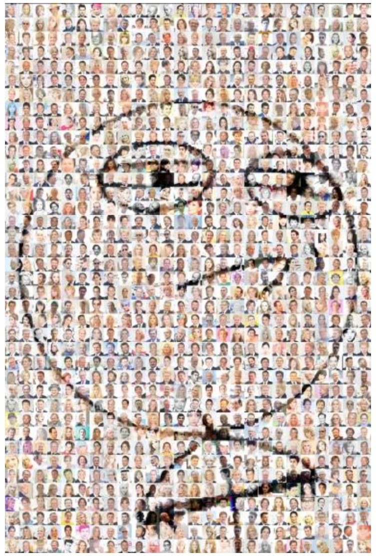
Identifying (spurious) correlations + Identifying unintended changes → Correlation?

We investigate correlations at the label level

Varied face attribute datasets

Dataset	# attributes	Type
CelebA	40	Binary
FFText-HQ	26	Binary & Multiclass
MAAD-Face	47	Binary





We investigate correlations at the label level

Varied face attribute datasets

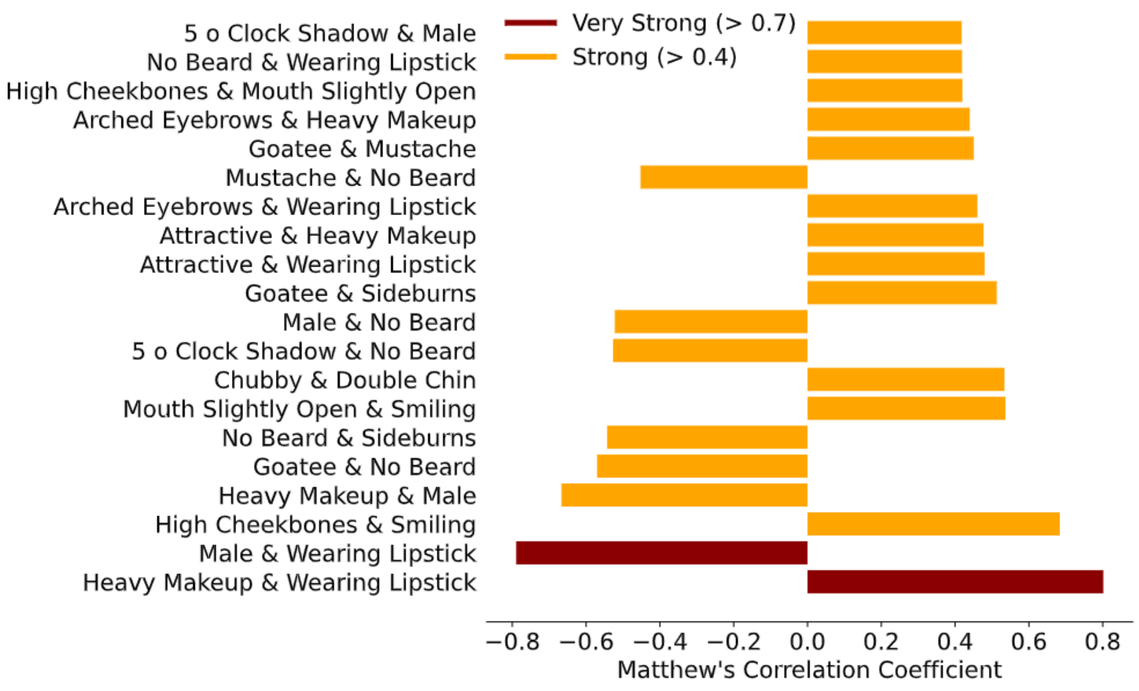
Dataset	# attributes	Type
CelebA	40	Binary
FFText-HQ	26	Binary & Multiclass
MAAD-Face	47	Binary

Categorical correlation metrics

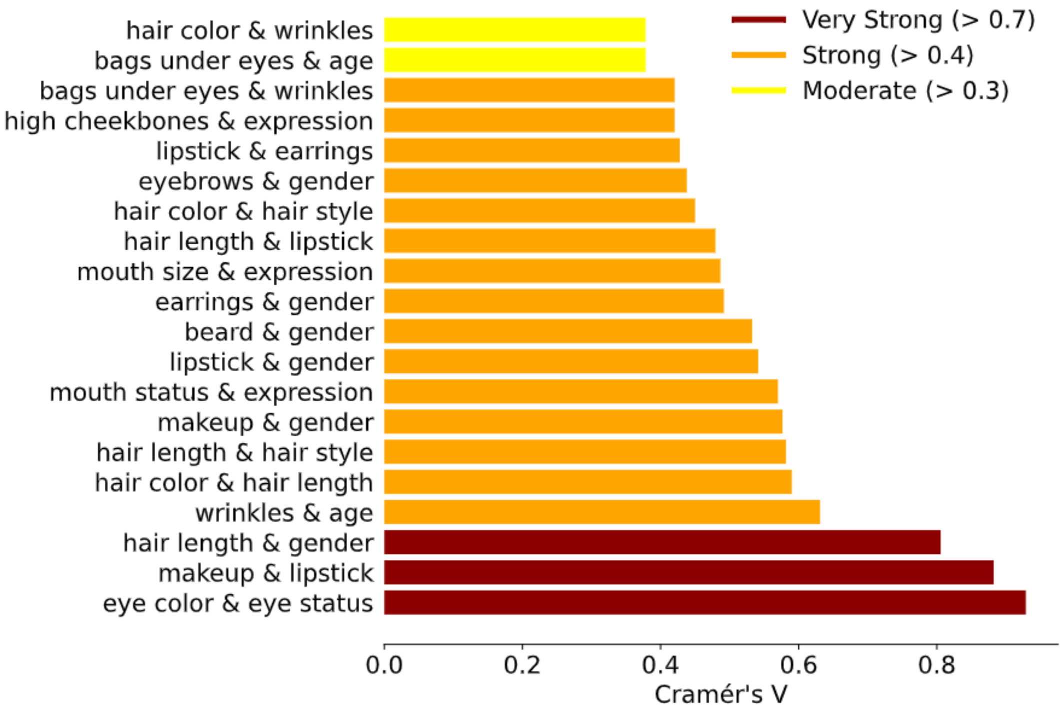
Correlation Metric	Data type
Matthew's correlation coefficient	Binary
Cramér's V	Multiclass
Uncertainty coefficient	Binary & Multiclass

→ Good for unbalanced data

Many strong correlations are present

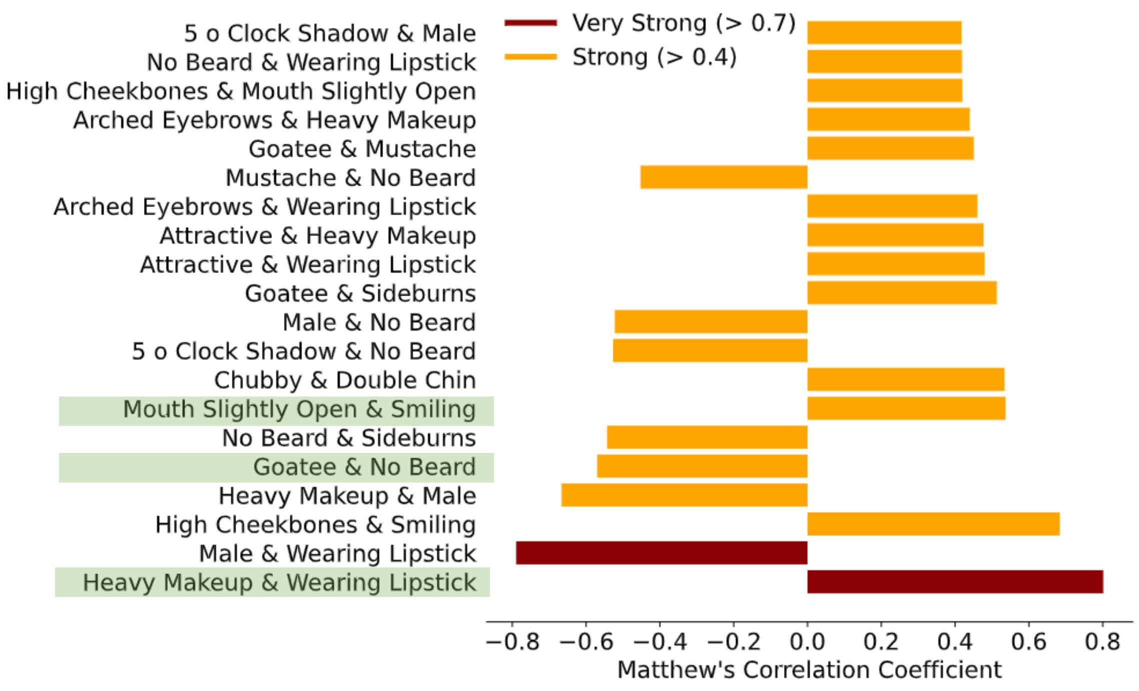


CelebA

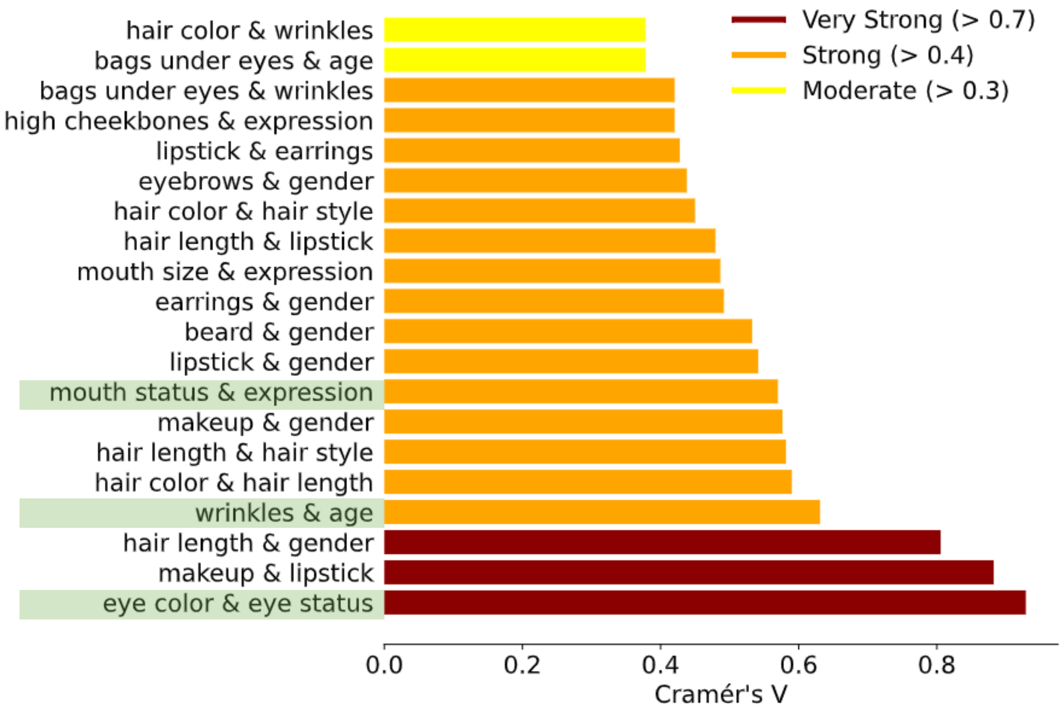


FFText-HQ

Some are quite logical

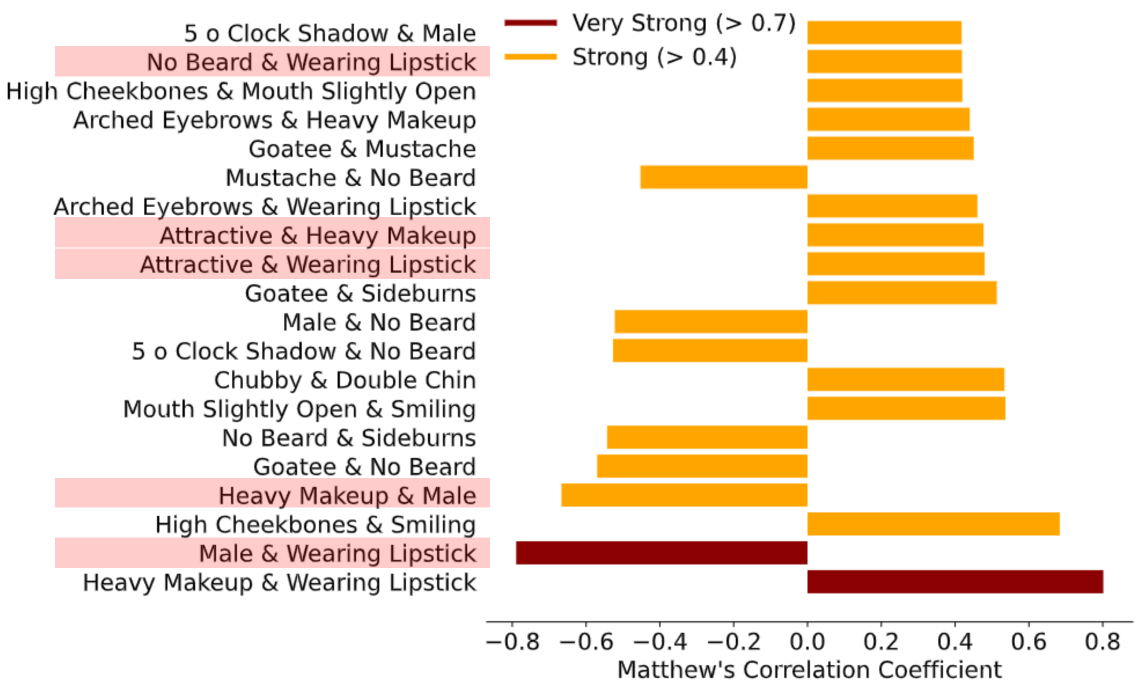


CelebA

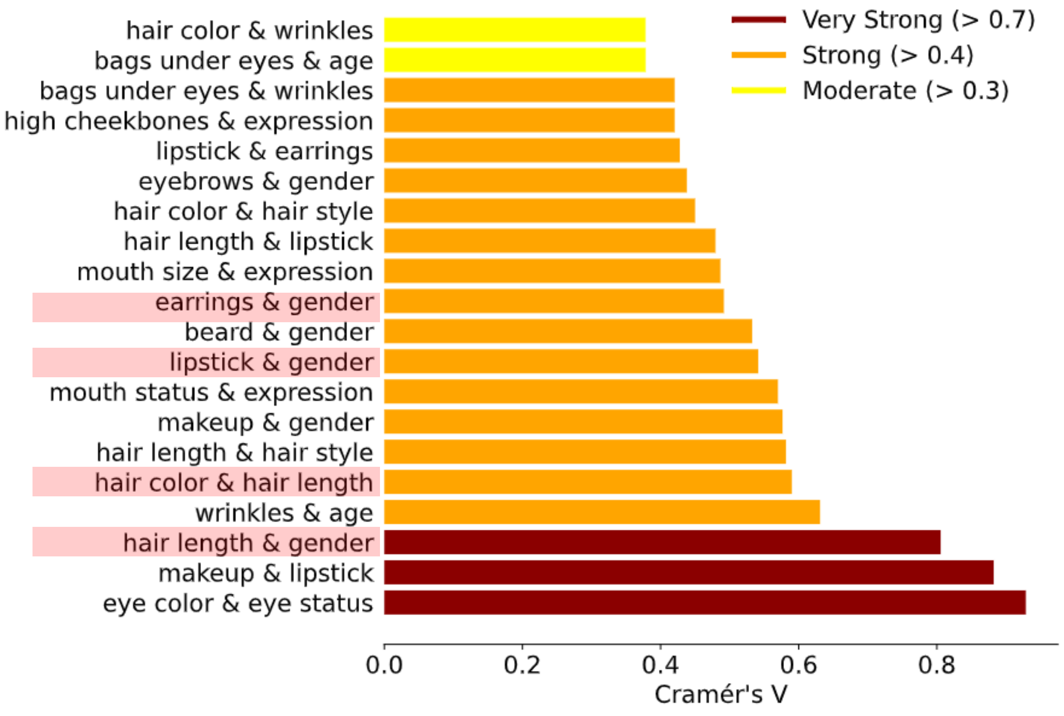


FFText-HQ

Others could propagate unwanted stereotypes



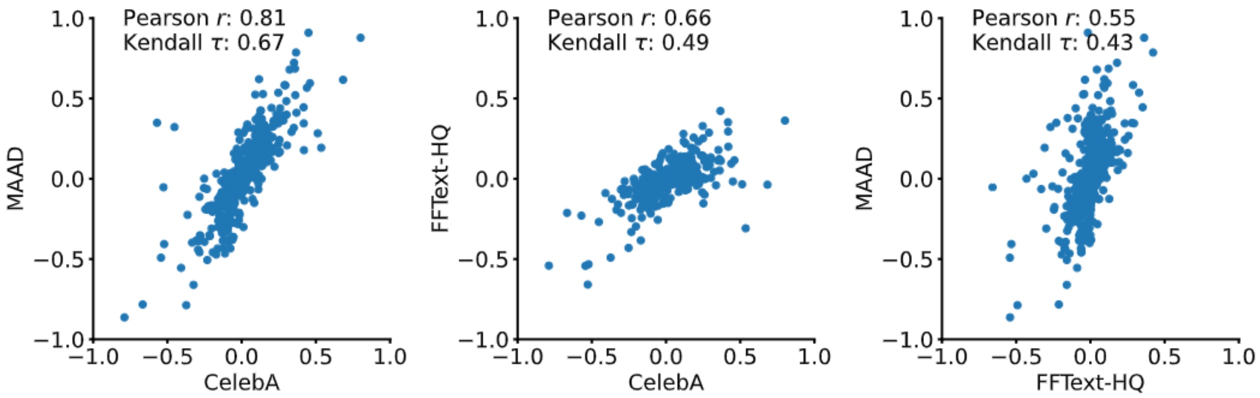
CelebA



FFText-HQ

Common aspects are found in the datasets

Strong relationship between correlations in the different datasets



With some noticeable differences

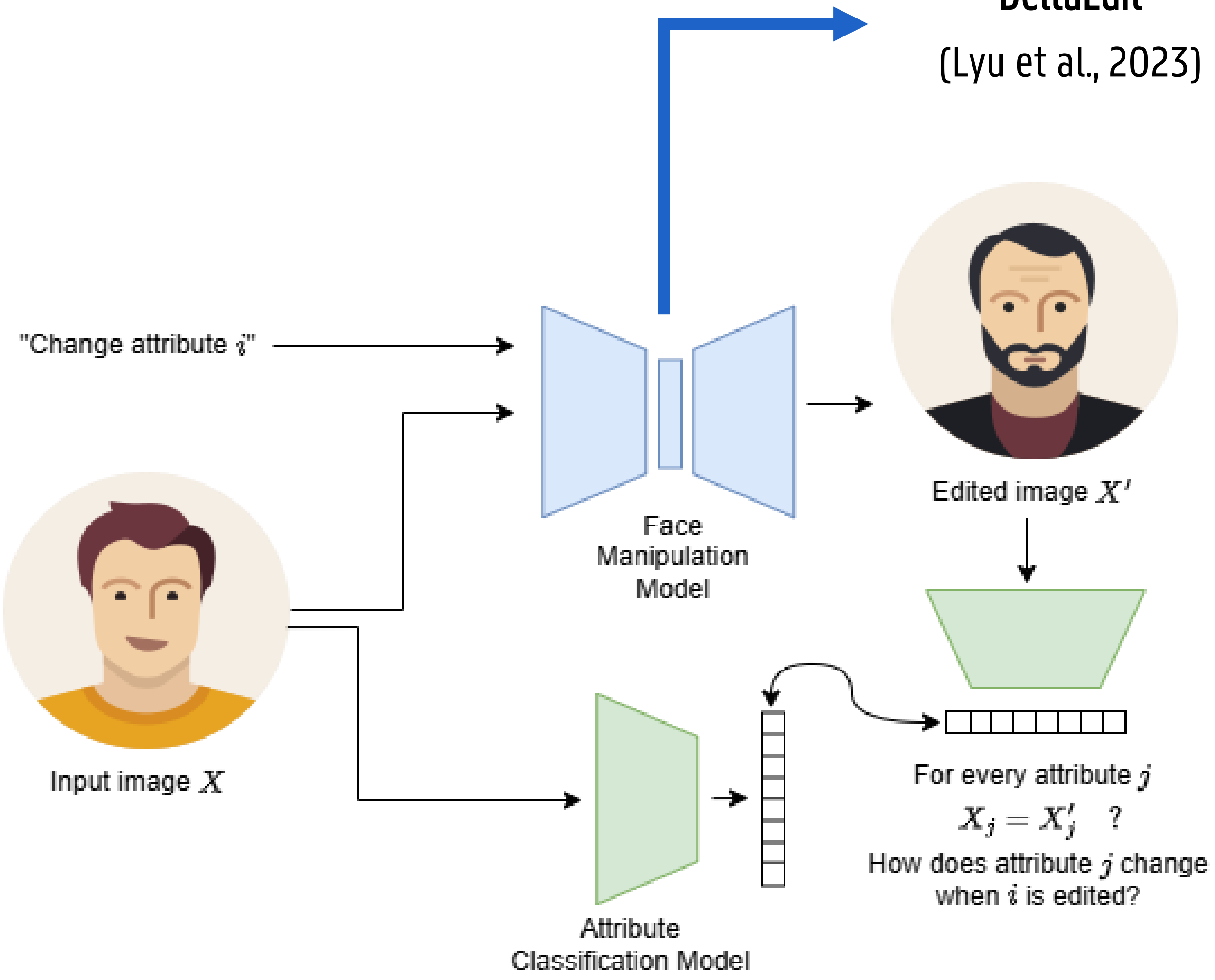
Combination	CelebA	MAAD	FFText-HQ
Heavy Makeup & Male	-0.67	-0.78	-0.21
Wavy Hair & Wearing Earrings	0.12	0.62	0.09

We test out 3 state of the art manipulation models

FFCLIP
(Zhu et al.,2022)

DeltaEdit
(Lyu et al., 2023)

ManiCLIP
(Wang et al., 2024)

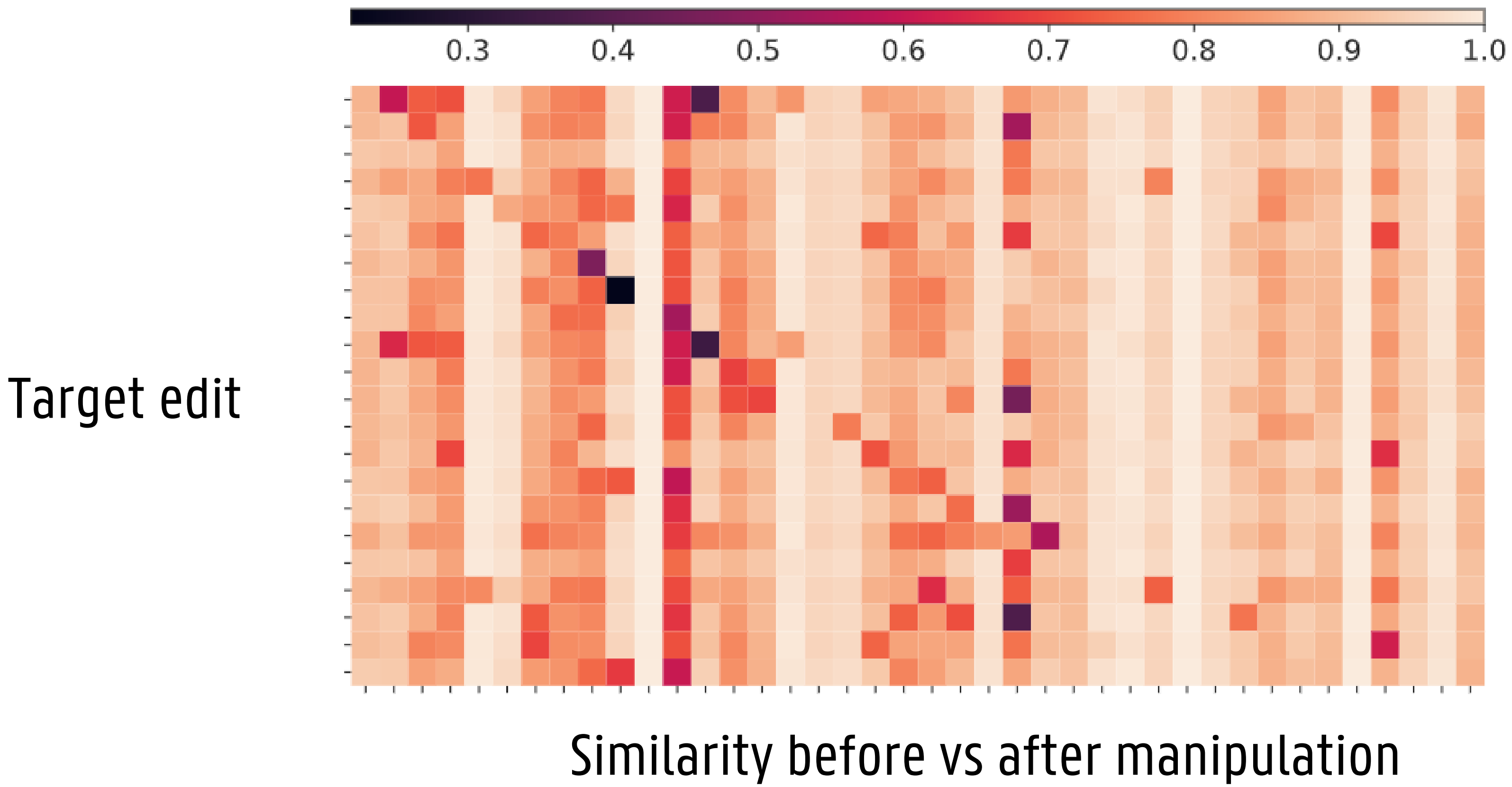


We notice two main errors arise when using manipulation methods

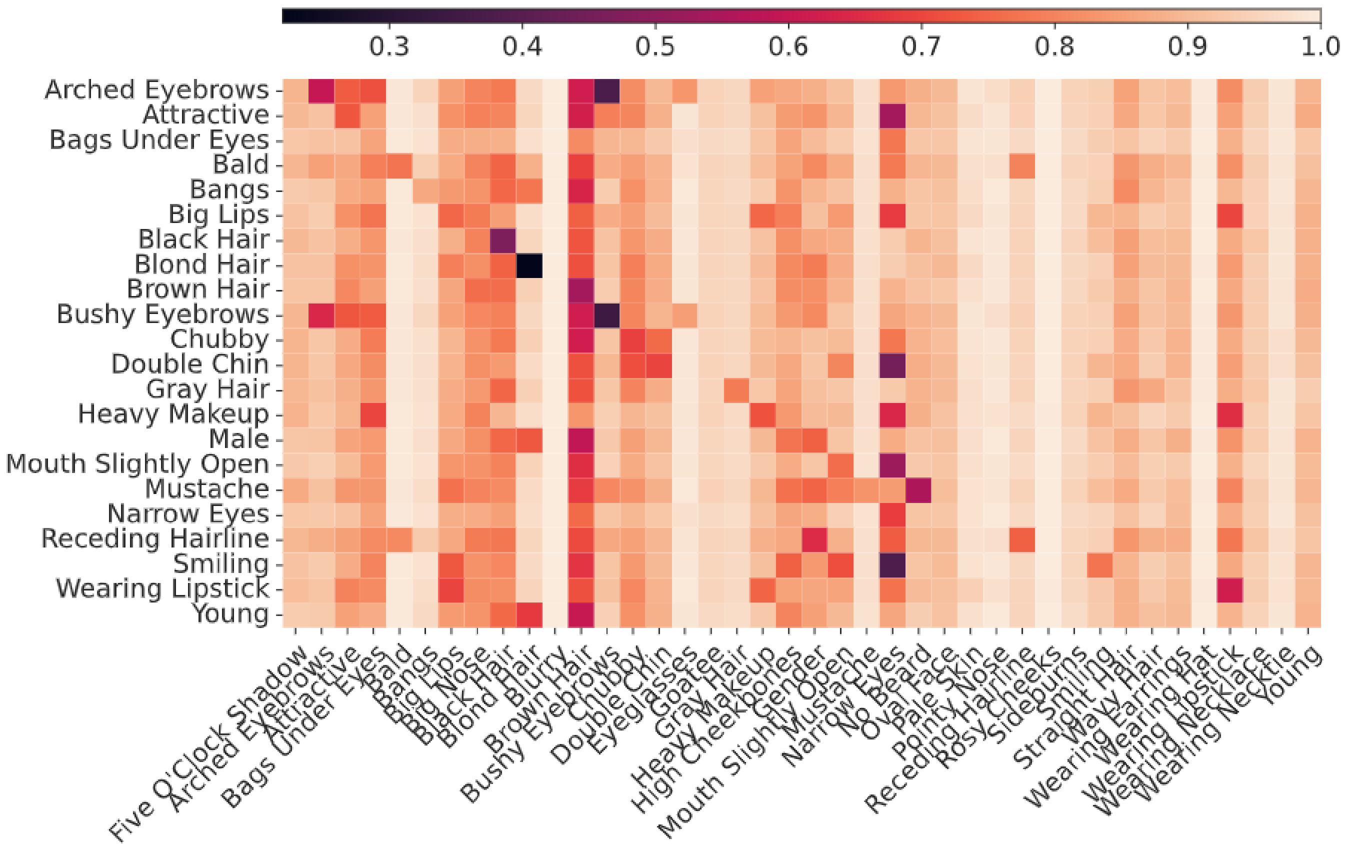
1. Manipulating one attribute brings on other, unexpected changes

2. Manipulating an attribute is less effective depending on the characteristics of the person

One change incurs other unwanted ones



One change incurs other unwanted ones

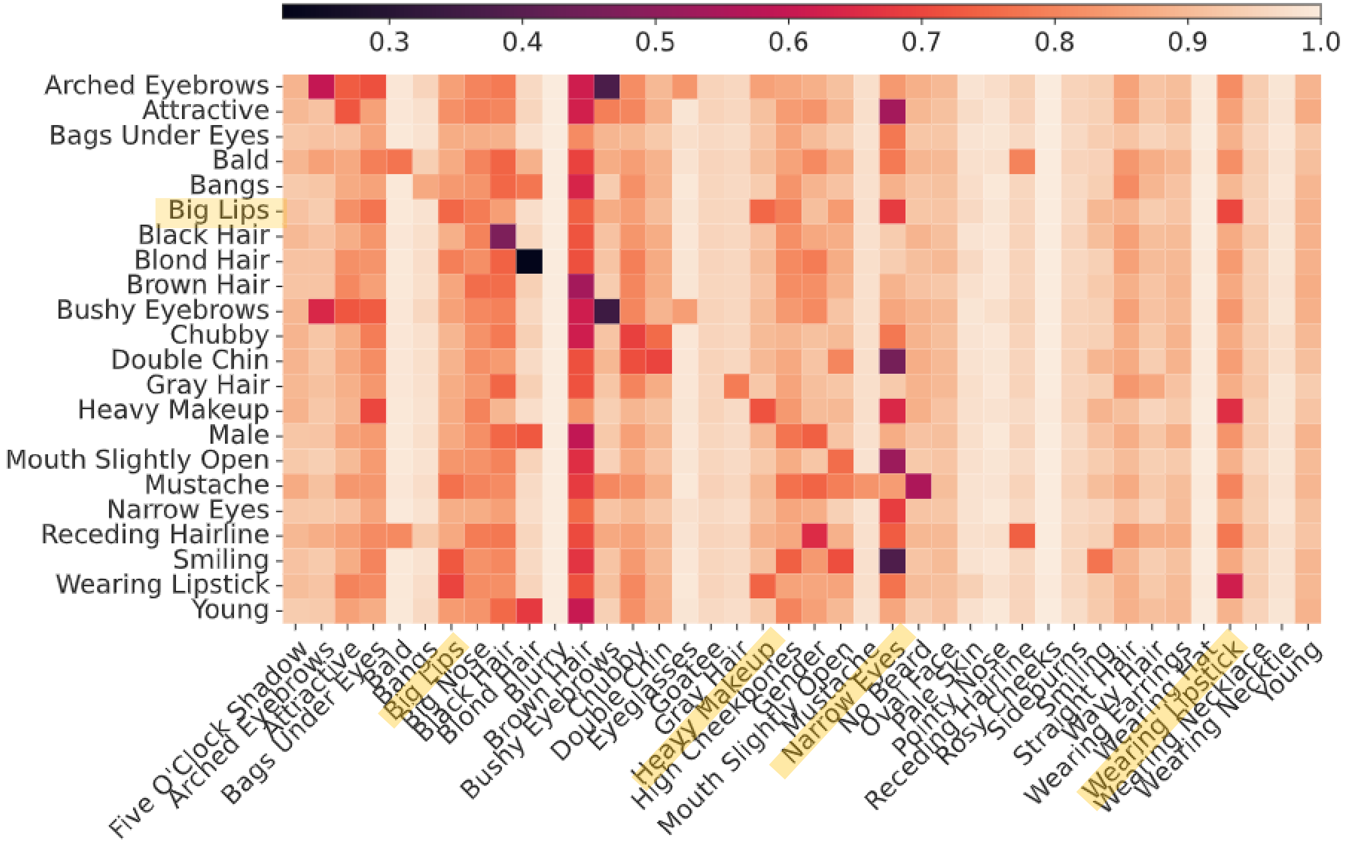


Target edit

Similarity before vs after manipulation

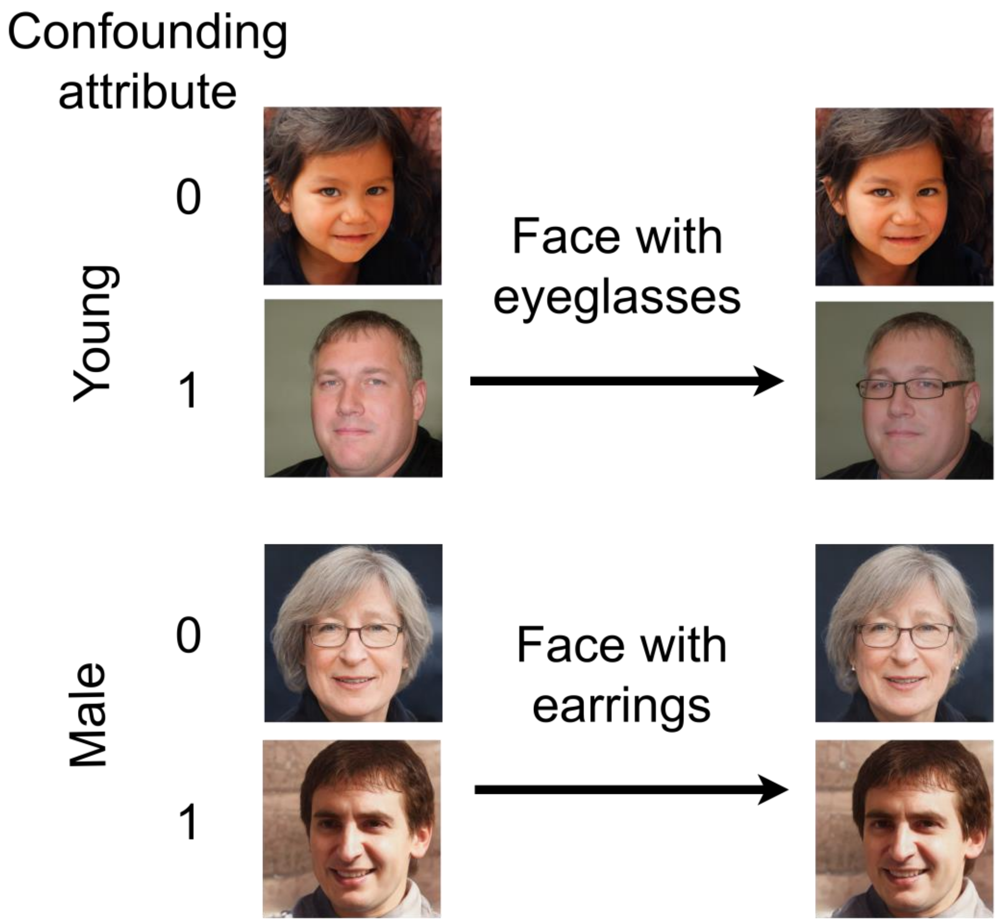
One change incurs other unwanted ones

Target edit



Similarity before vs after manipulation

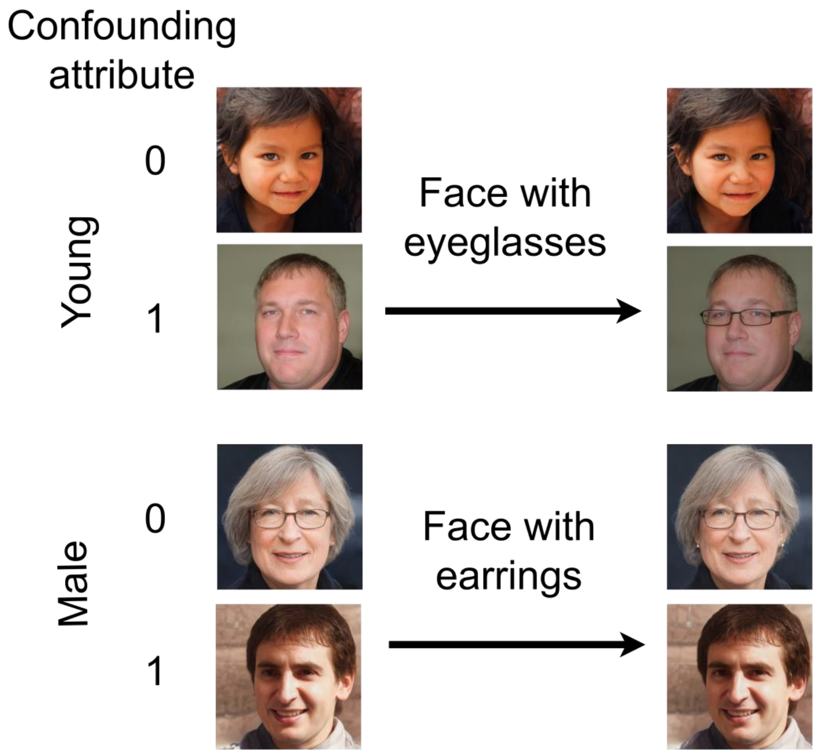
One change doesn't work as well based on a confounder



One change doesn't work as well based on a confounder

Percent changed

Confounder

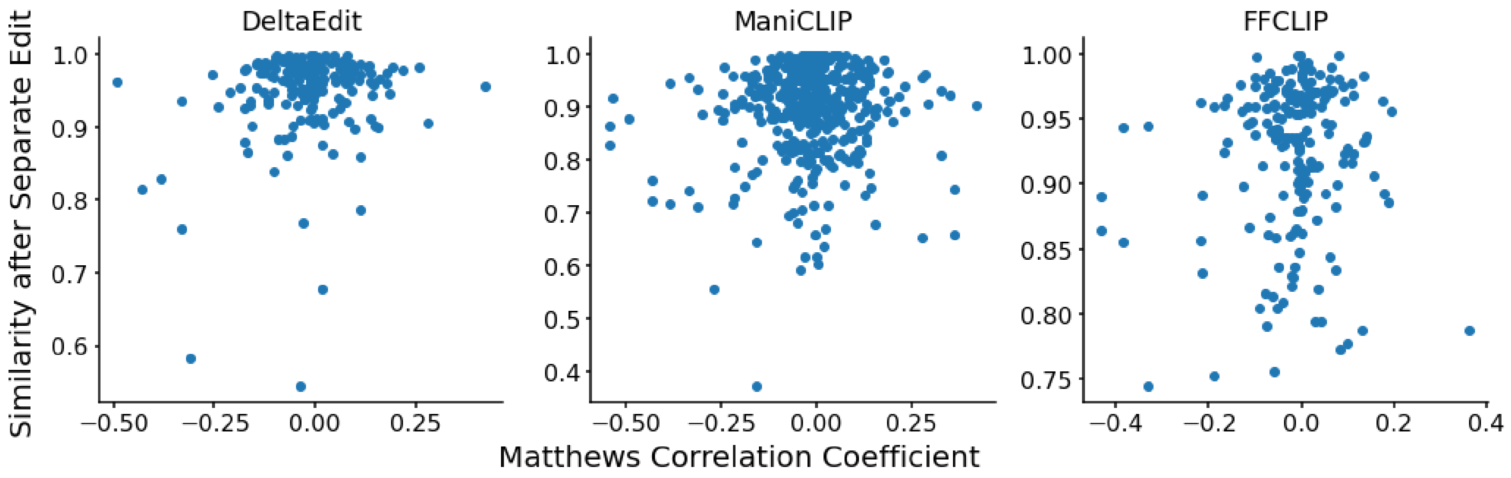


But the correlations are not related

Correlating MCC/UC with unintended change
after manipulating different attribute

Expectation: High correlation ⇨ more changes

Found: No relation



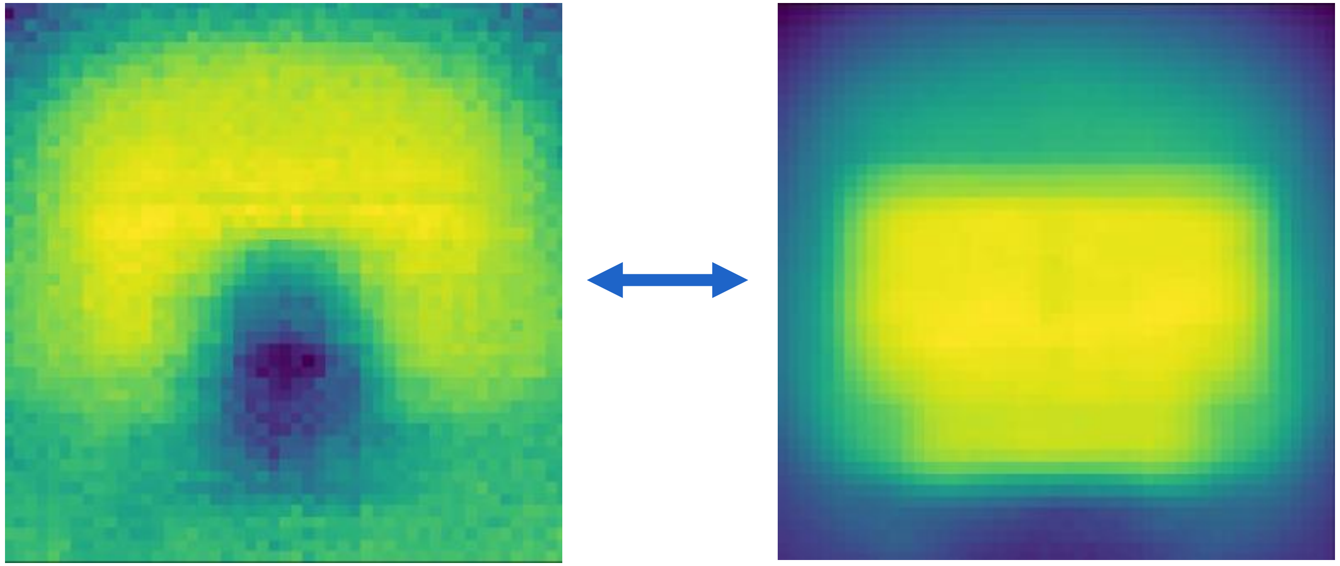
So, what else could it be?

Use explainability techniques to calculate attribute 'correlations'

XRAI

Region Detector

GRADCAM



Hair color

Perceived gender

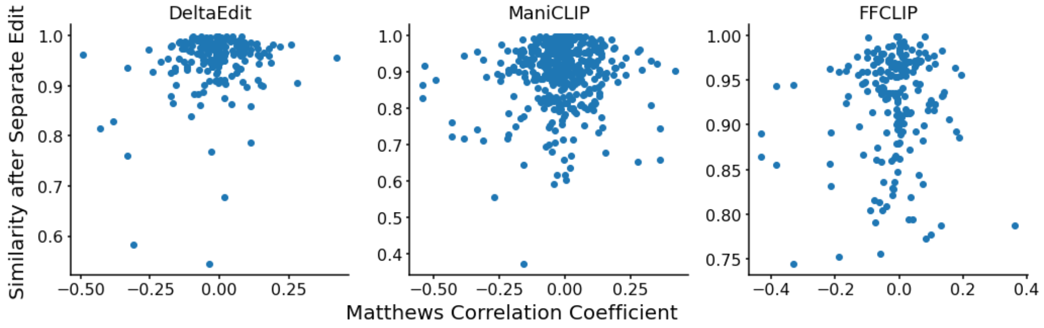
Conclusion

Unintended consequences were found

1. Unintended changes

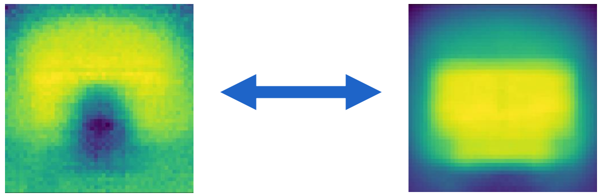
2. Confounder-dependent effectiveness

However, they weren't related to label correlations



So what else could it be?

Using explainability techniques to look at data attributions



Hair color

Perceived gender

Image manipulation models provide a convenient way to enrichen your data

Existing dataset: FFHQ

+

Edit prompt: *"a face with grey hair"*

Generated data

Diverse ✓

Privacy-friendly ✓