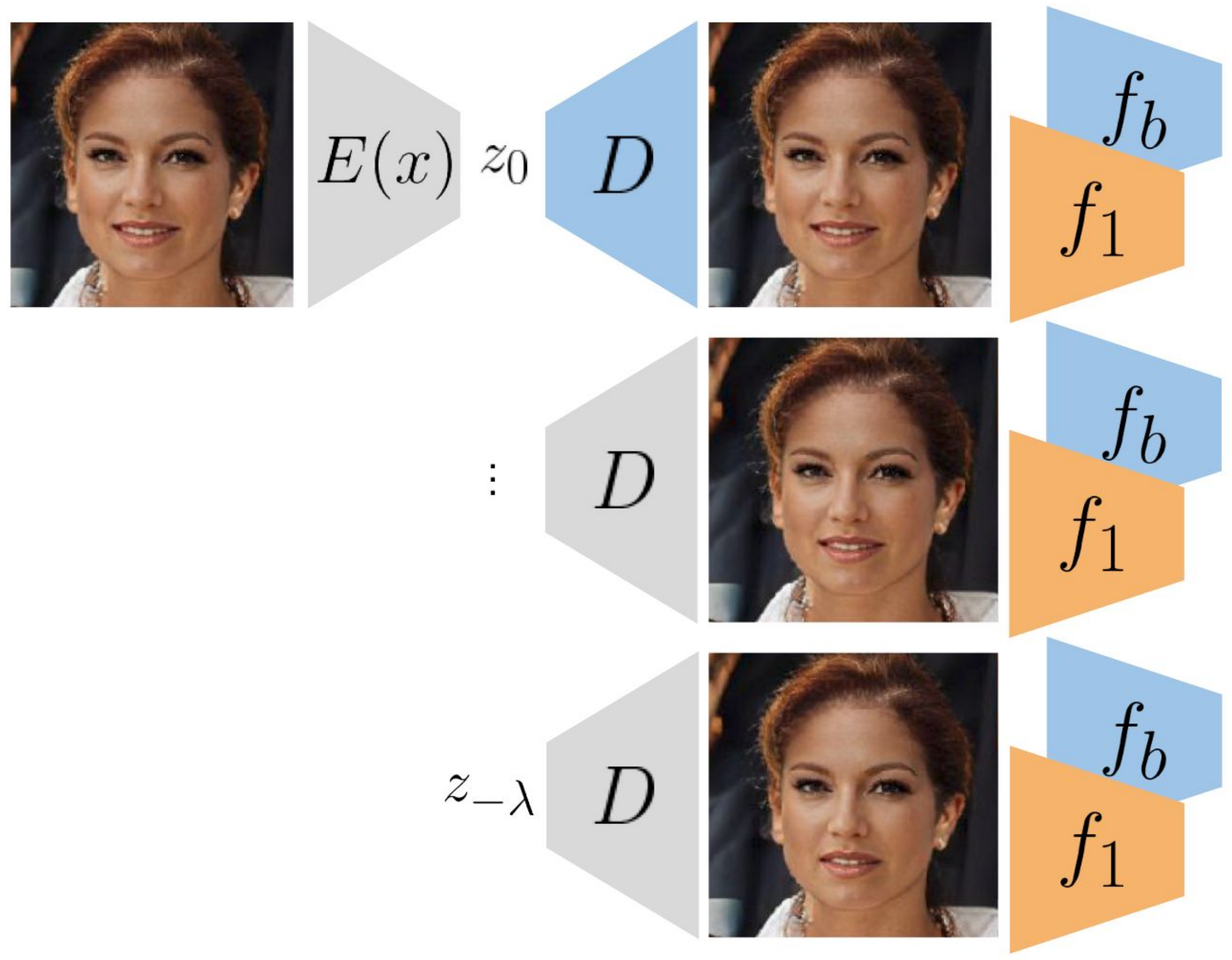
Identifying Spurious Correlations using Counterfactual Alignment

Joseph Paul Cohen, Louis Blankemeier, Akshay Chaudhari

Stanford

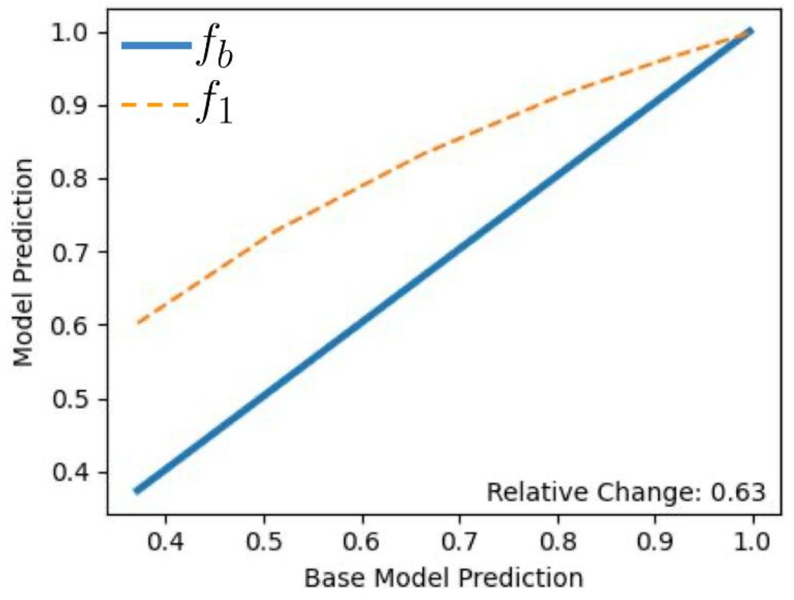
Counterfactual Alignment



Images generated at different lambdas w.r.t. the base classifier.

Images processed with multiple classifiers.

Plotting relative to base classifier.

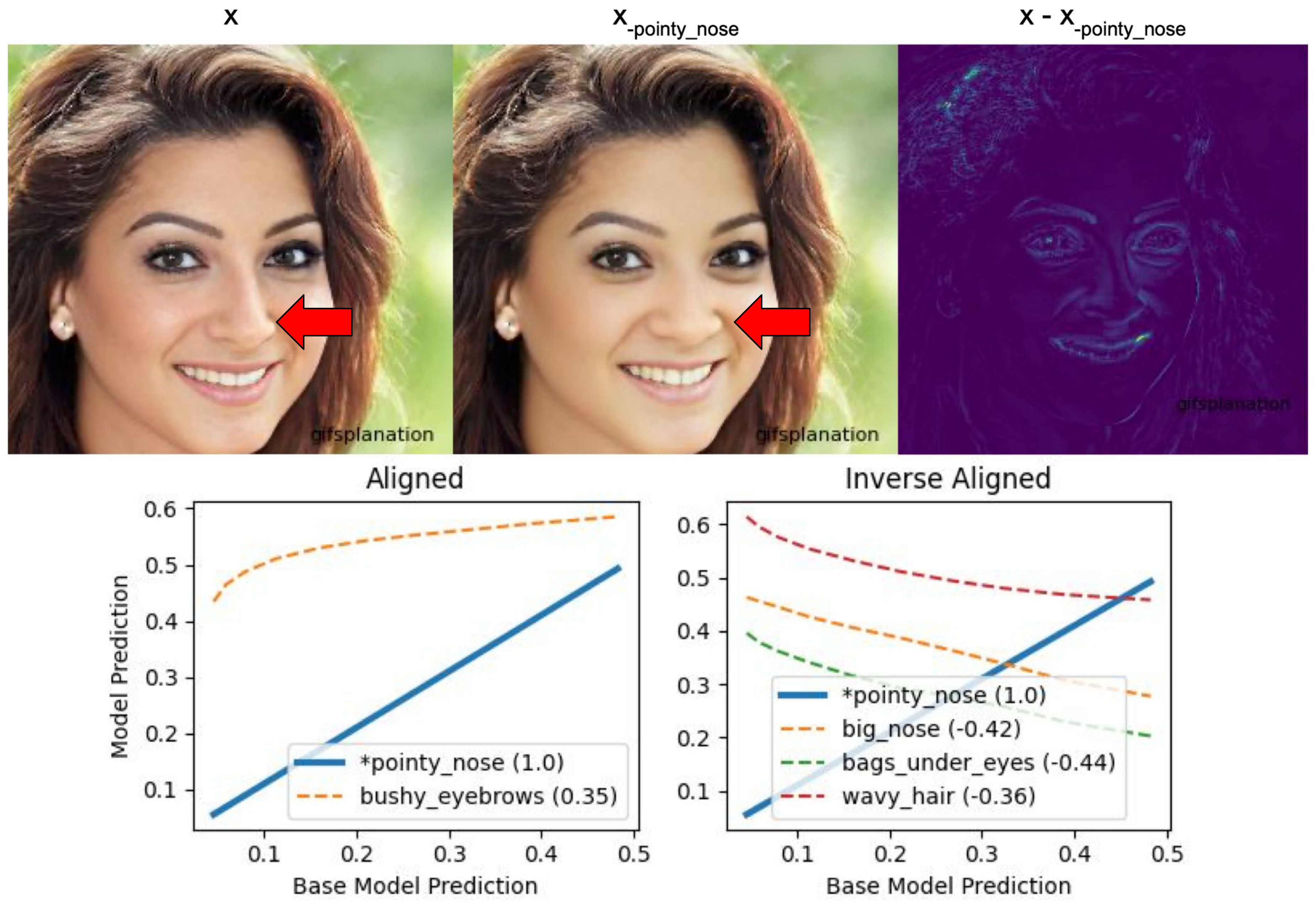


Compute relative change between the predictions:

$$\frac{f_1(D(z_\lambda)) - f_1(D(z_0))}{f_b(D(z_\lambda)) - f_b(D(z_0))}$$

- Counterfactuals are generated by subtracting the gradient of the classifier output w.r.t. the latent representation.
- The representation is reconstructed back into multiple images with different magnitudes of change (depending on the λ).
- Reconstructed images are processed with multiple classifiers.
- Base model predictions can be used as the x-axis to more easily compare it to the predictions of another classifier.
- Changes can be quantified and compared using relative change.

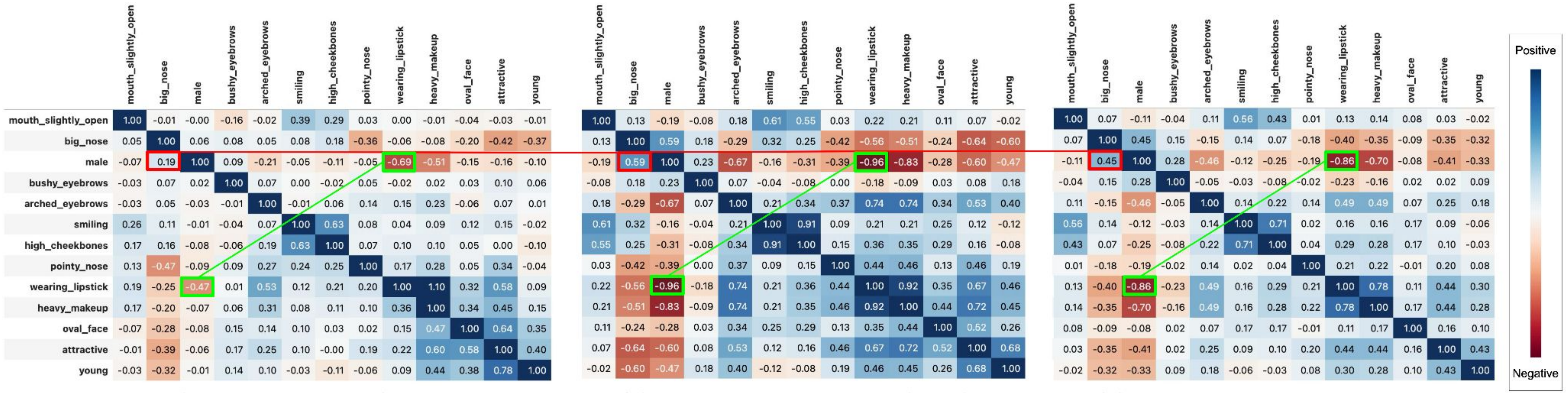
Studying Face Attribute Classifiers



CF alignment for pointy nose showing an inverse alignment with big nose and potential spurious relationships with eyebrows eyes and hair. The relative change is shown next to each classifier name.

Aggregate Spurious Correlations

Relationships between face attribute classifiers using different metrics.
Base classifiers are along the rows and downstream classifiers are along the columns.



a) Relative change using CF alignment.

	mouth_slightly_open	big_nose	male	bushy_eyebrows	arched_eyebrows	smiling	high_cheekbones	pointy_nose	wearing_lipstick	heavy_makeup	oval_face	attractive	young
mouth_slightly_open	1.00	-0.01	-0.00	-0.16	-0.02	0.39	0.29	0.03	0.00	-0.01	-0.04	-0.03	-0.01
big_nose	0.05	1.00	0.06	0.08	0.05	0.08	0.18	-0.36	-0.06	-0.08	-0.20	-0.42	-0.37
male	-0.07	0.19	1.00	0.09	-0.21	-0.05	-0.11	-0.05	-0.69	-0.51	-0.15	-0.16	-0.10
bushy_eyebrows	-0.03	0.07	0.02	1.00	0.07	0.00	-0.02	0.05	-0.02	0.02	0.03	0.10	0.06
arched_eyebrows	-0.03	0.05	-0.03	-0.01	1.00	-0.01	0.06	0.14	0.15	0.23	-0.06	0.07	0.01
smiling	0.26	0.11	-0.01	-0.04	0.07	1.00	0.63	0.08	0.04	0.09	0.12	0.15	-0.02
high_cheekbones	0.17	0.16	-0.08	-0.06	0.19	0.63	1.00	0.07	0.10	0.10	0.05	0.00	-0.10
pointy_nose	0.13	-0.47	-0.09	0.09	0.27	0.24	0.25	1.00	0.17	0.28	0.05	0.34	-0.04
wearing_lipstick	0.19	-0.25	-0.47	0.01	0.53	0.12	0.21	0.20	1.00	1.10	0.32	0.58	0.09
heavy_makeup	0.17	-0.20	-0.07	0.06	0.31	0.08	0.11	0.10	0.36	1.00	0.34	0.45	0.15
oval_face	-0.07	-0.28	-0.08	0.15	0.14	0.10	0.03	0.02	0.15	0.47	1.00	0.64	0.35
attractive	-0.01	-0.39	-0.06	0.17	0.25	0.10	-0.00	0.19	0.22	0.60	0.58	1.00	0.40
young	-0.03	-0.32	-0.01	0.14	0.10	-0.03	-0.11	-0.06	0.09	0.44	0.38	0.78	1.00

b) Correlation between predictions of classifier.

	mouth_slightly_open	big_nose	male	bushy_eyebrows	arched_eyebrows	smiling	high_cheekbones	pointy_nose	wearing_lipstick	heavy_makeup	oval_face	attractive	young
mouth_slightly_open	1.00	0.13	-0.19	-0.08	0.18	0.61	0.55	0.03	0.22	0.21	0.11	0.07	-0.02
big_nose	0.13	1.00	0.59	0.18	-0.29	0.32	0.25	-0.42	-0.56	-0.51	-0.24	-0.64	-0.60
male	-0.19	0.59	1.00	0.23	-0.67	-0.16	-0.31	-0.39	-0.96	-0.83	-0.28	-0.60	-0.47
bushy_eyebrows	-0.08	0.18	0.23	1.00	0.07	-0.04	-0.08	0.00	-0.18	-0.09	0.03	0.08	0.18
arched_eyebrows	0.18	-0.29	-0.67	0.07	1.00	0.21	0.34	0.37	0.74	0.74	0.34	0.53	0.40
smiling	0.61	0.32	-0.16	-0.04	0.21	1.00	0.91	0.09	0.21	0.21	0.25	0.12	-0.12
high_cheekbones	0.55	0.25	-0.31	-0.08	0.34	0.91	1.00	0.15	0.36	0.35	0.29	0.16	-0.08
pointy_nose	0.03	-0.42	-0.39	0.00	0.37	0.09	0.15	1.00	0.44	0.46	0.13	0.46	0.19
wearing_lipstick	0.22	-0.56	-0.96	-0.18	0.74	0.21	0.36	0.44	1.00	0.92	0.35	0.67	0.46
heavy_makeup	0.21	-0.51	-0.83	-0.09	0.74	0.21	0.35	0.46	0.92	1.00	0.44	0.72	0.45
oval_face	0.11	-0.24	-0.28	0.03	0.34	0.25	0.29	0.13	0.35	0.44	1.00	0.52	0.26
attractive	0.07	-0.64	-0.60	0.08	0.53	0.12	0.16	0.46	0.67	0.72	0.52	1.00	0.68
young	-0.02	-0.60	-0.47	0.18	0.40	-0.12	-0.08	0.19	0.46	0.45	0.26	0.68	1.00

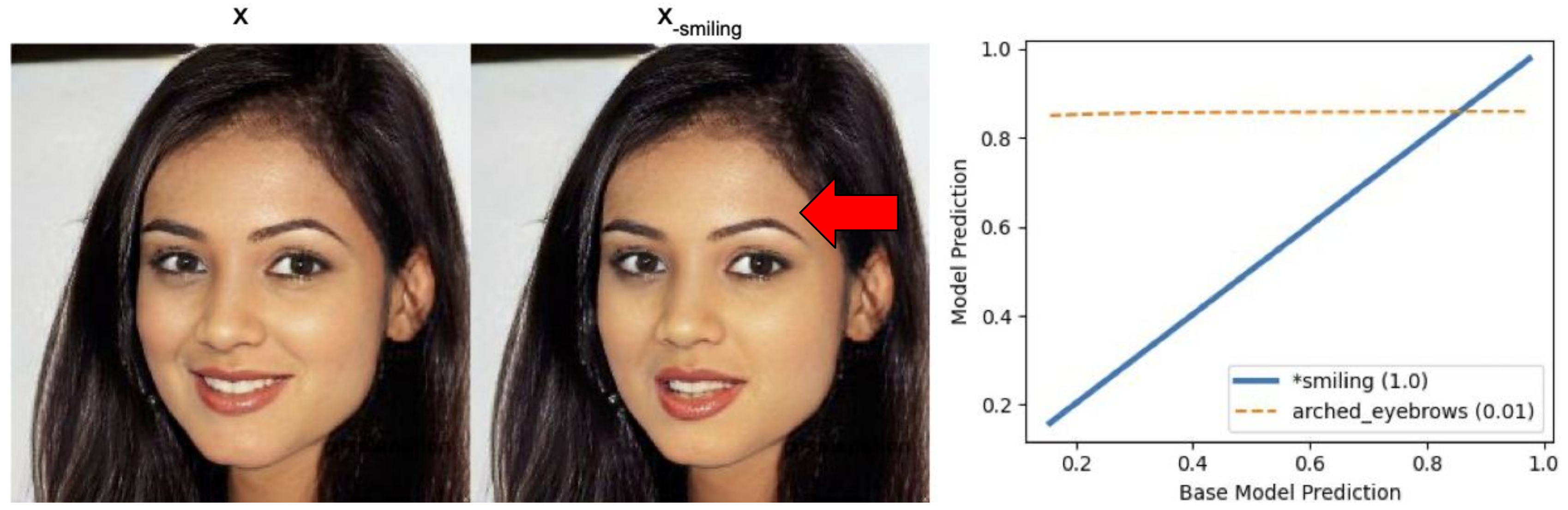
c) Correlation between labels in training data.

	mouth_slightly_open	big_nose	male	bushy_eyebrows	arched_eyebrows	smiling	high_cheekbones	pointy_nose	wearing_lipstick	heavy_makeup	oval_face	attractive	young
mouth_slightly_open	1.00	0.07	-0.11	-0.04	0.11	0.56	0.43	0.01	0.13	0.14	0.08	0.03	-0.02
big_nose	0.07	1.00	0.45	0.15	-0.15	0.14	0.07	-0.18	-0.40	-0.35	-0.09	-0.35	-0.32
male	-0.11	0.45	1.00	0.28	-0.46	-0.12	-0.25	-0.19	-0.86	-0.70	-0.08	-0.41	-0.33
bushy_eyebrows	-0.04	0.15	0.28	1.00	-0.05	-0.03	-0.08	-0.02	-0.23	-0.16	0.02	0.02	0.09
arched_eyebrows	0.11	-0.15	-0.46	-0.05	1.00	0.14	0.22	0.14	0.49	0.49	0.07	0.25	0.18
smiling	0.56	0.14	-0.12	-0.03	0.14	1.00	0.71	0.02	0.16	0.16	0.17	0.09	-0.06
high_cheekbones	0.43	0.07	-0.25	-0.08	0.22	0.71	1.00	0.04	0.29	0.28	0.17	0.10	-0.03
pointy_nose	0.01	-0.18	-0.19	-0.02	0.14	0.02	0.04	1.00	0.21	0.22	-0.01	0.20	0.08
wearing_lipstick	0.13	-0.40	-0.86	-0.23	0.49	0.16	0.29	0.21	1.00	0.78	0.11	0.44	0.30
heavy_makeup	0.14	-0.35	-0.70	-0.16	0.49	0.16	0.28	0.22	0.78	1.00	0.17	0.44	0.28
oval_face	0.08	-0.09	-0.08	0.02	0.07	0.17	0.17	-0.01	0.11	0.17	1.00	0.16	0.10
attractive	0.03	-0.35	-0.41	0.02	0.25	0.09	0.10	0.20	0.44	0.44	0.16	1.00	0.43
young	-0.02	-0.32	-0.33	0.09	0.18	-0.06	-0.03	0.08	0.30	0.28	0.10	0.43	1.00

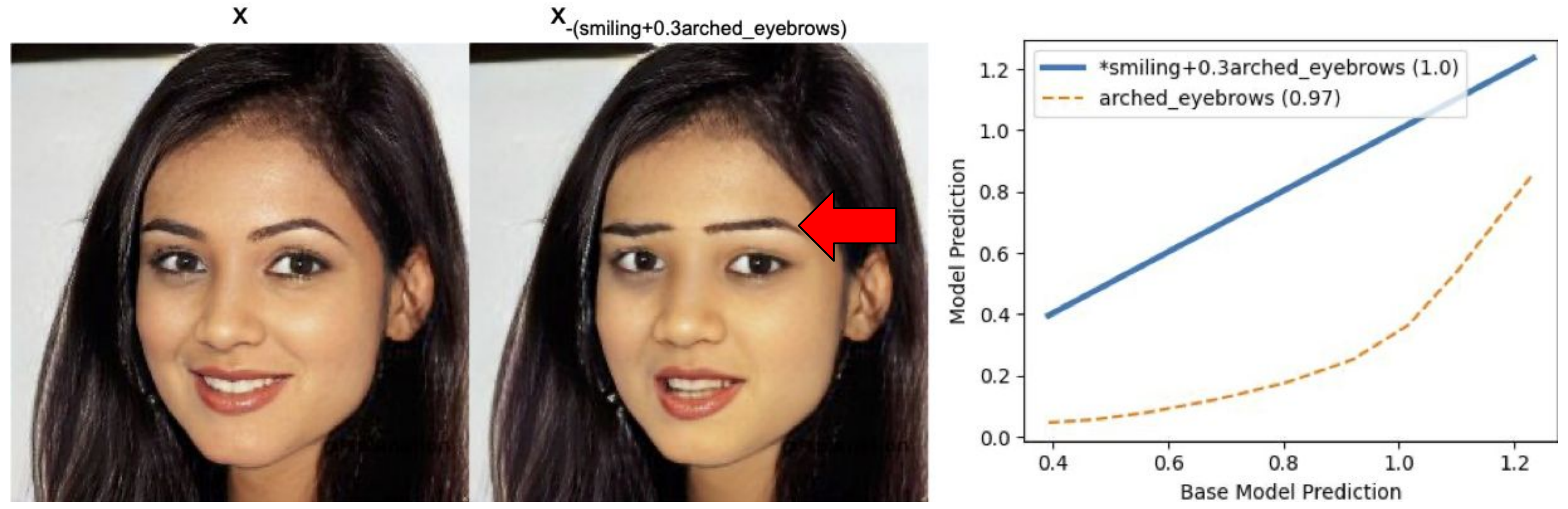
- Comparing (a) to (b) and (c), shows that many relationships reflected in the CF outputs are preserved from correlations in the training data.
- A non-symmetric relationship between male and wearing lipstick (highlighted in green) reflects an expected relationship.
- The relationship between male and big nose (highlighted in red) is strong in both the classifier predictions and ground truth labels but low in CF alignment, indicating that although correlated, these features aren't exploited by the classifier.

Detecting Induced Spurious Correlations

Example detecting a spurious correlation in a biased classifier. A classifier is biased with arched eyebrows and this is observed in the CF alignment plot as well as in the counterfactual image. We can observe that in order to decrease the prediction of the (now biased) classifier, the CF image also removes the arched eyebrows.



CF for smiling showing eyebrows are unchanged. The horizontal line indicates the prediction of arched eyebrows is not influenced by the features used for smiling in this image.



CF for the modified smiling classifier with an arched eyebrow spurious correlation showing that eyebrows are now changed.

Paper URL: https://arxiv.org/abs/2312.02186
Project Website: https://latentshift.github.io/

Code: https://github.com/ieee8023/latentshift
Contact: joseph@josephpcohen.com