UNIVERSITY OF WARSAW

UNIVERSITY OF WARSAW
FACULTY OF MATHEMATICS, INFORMATICS AND MECHANICS

MI

Faculty of Mathematics and Information Sciences
WARSAW UNIVERSITY OF TECHNOLOGY

POLITECHNIKA WARSZAWSKA
WARSAW UNIVERSITY OF TECHNOLOGY

Global Counterfactual Directions

Bartlomiej Sobieski, Przemyslaw Biecek

ECCV
EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

ML in PL
CONFERENCE 2024

Outline

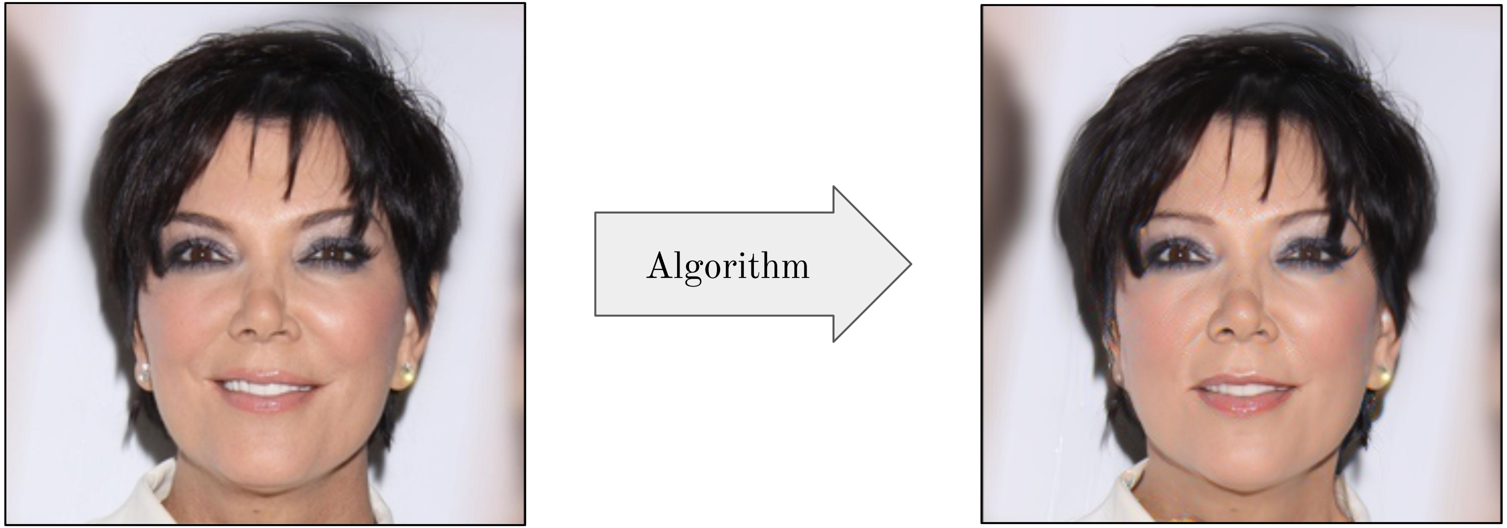
1. Visual Counterfactual Explanations
2. Diffusion Autoencoders
3. Global Counterfactual Directions for Black-Box Models
4. Conclusions

Visual Counterfactual Explanations

What are Visual Counterfactual Explanations (VCEs)?

For a given **classifier**, what is the **minimal semantic modification** of the **image** that **flips** the model's **decision**?

$f(\text{smile} \mid \mathbf{x}) = 0.97$ $\qquad$ $f(\text{smile} \mid \mathbf{x}') = 0.12$

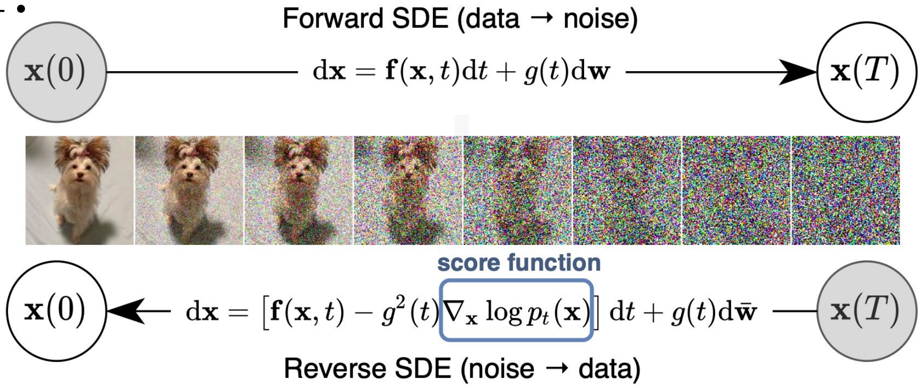


Previous works

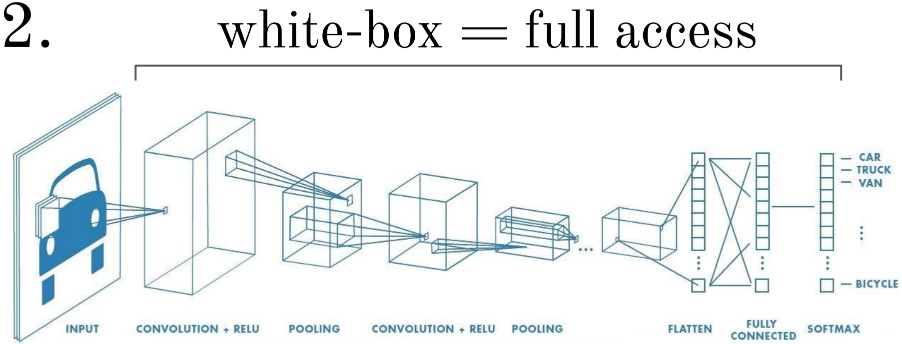
1. utilize *diffusion models* as generative priors

2. assume *white-box* access to the explained classifier

1.



2.

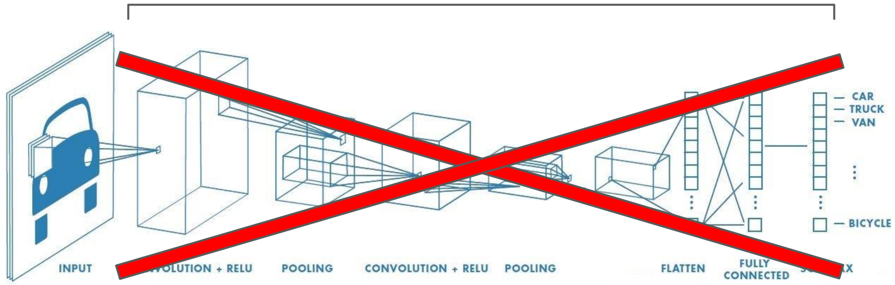


Limitations

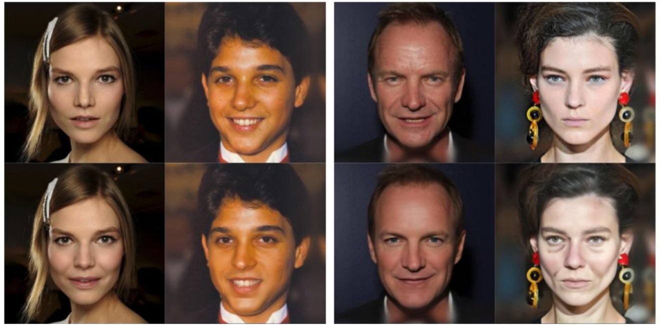
1. white-box access is often not realistic, think ChatGPT

2. VCEs are considered independently for each image

1. black-box = no access



2. Optimized independently



Central question

Can we simultaneously

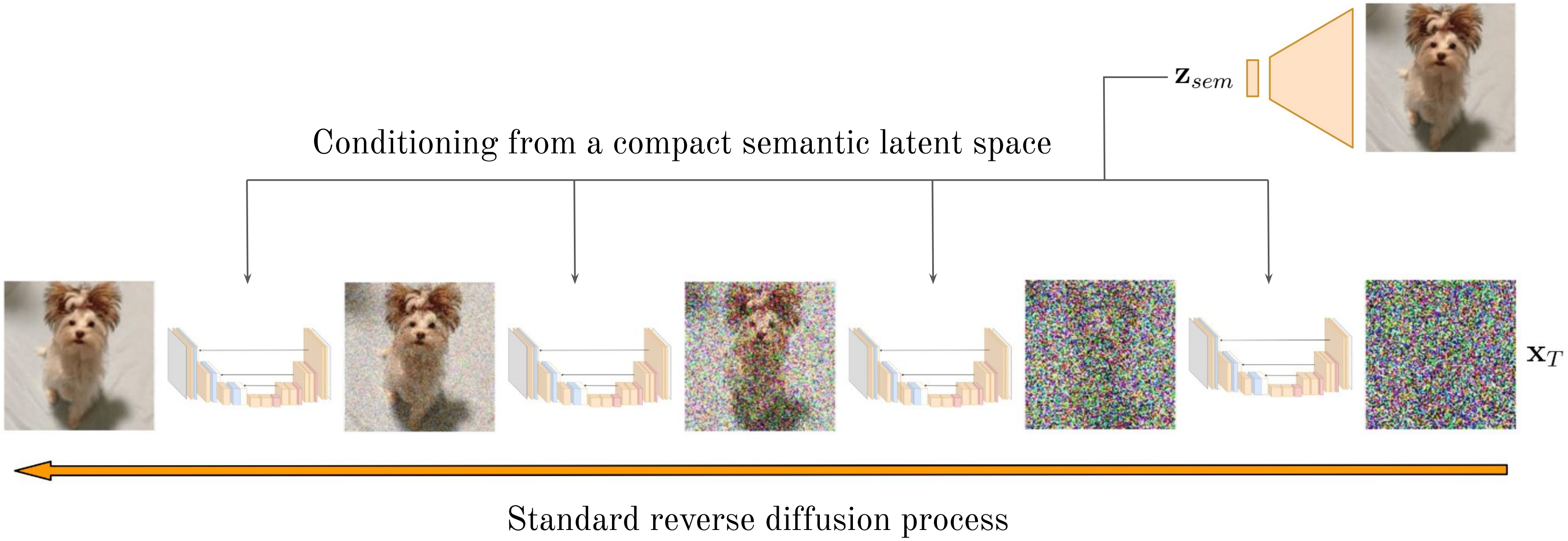
1. synthesize VCEs in a black-box setting using diffusion models

   and

2. find any possible links between the explanations?

Diffusion Autoencoders (DiffAE)

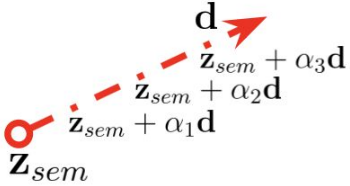
Diffusion Autoencoders (Preechakul et al., CVPR 2022)

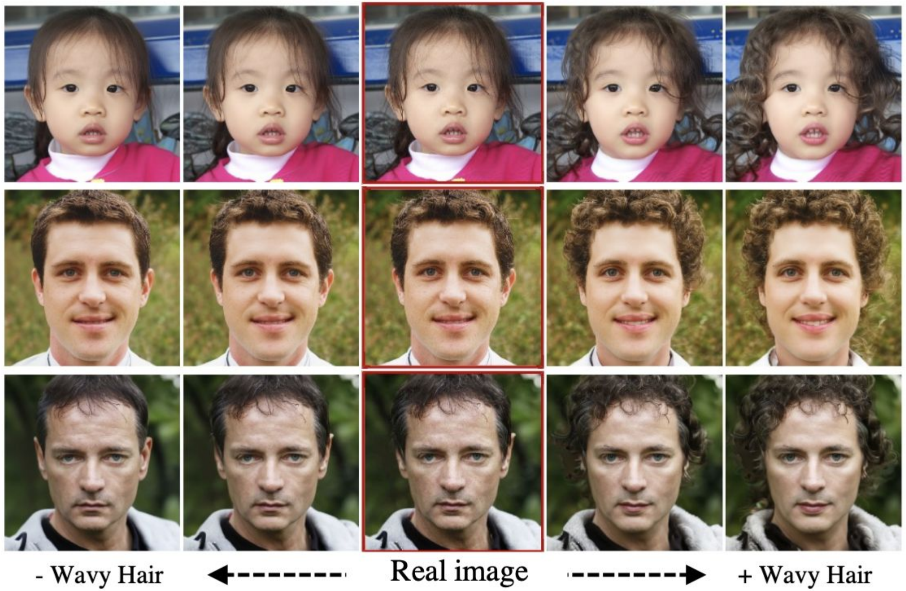


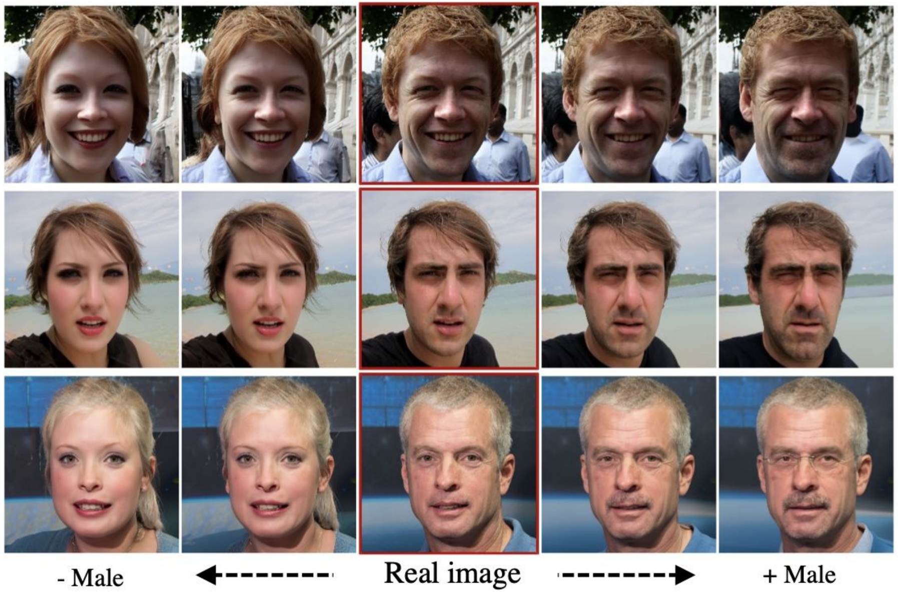
Disentangled properties

Input Reconstruction $(\mathbf{z}_{\text{sem}}, \mathbf{x}_T)$ Varying stochastic subcode $(\mathbf{z}_{\text{sem}}, \mathbf{x}_T^i)$

Global semantic directions

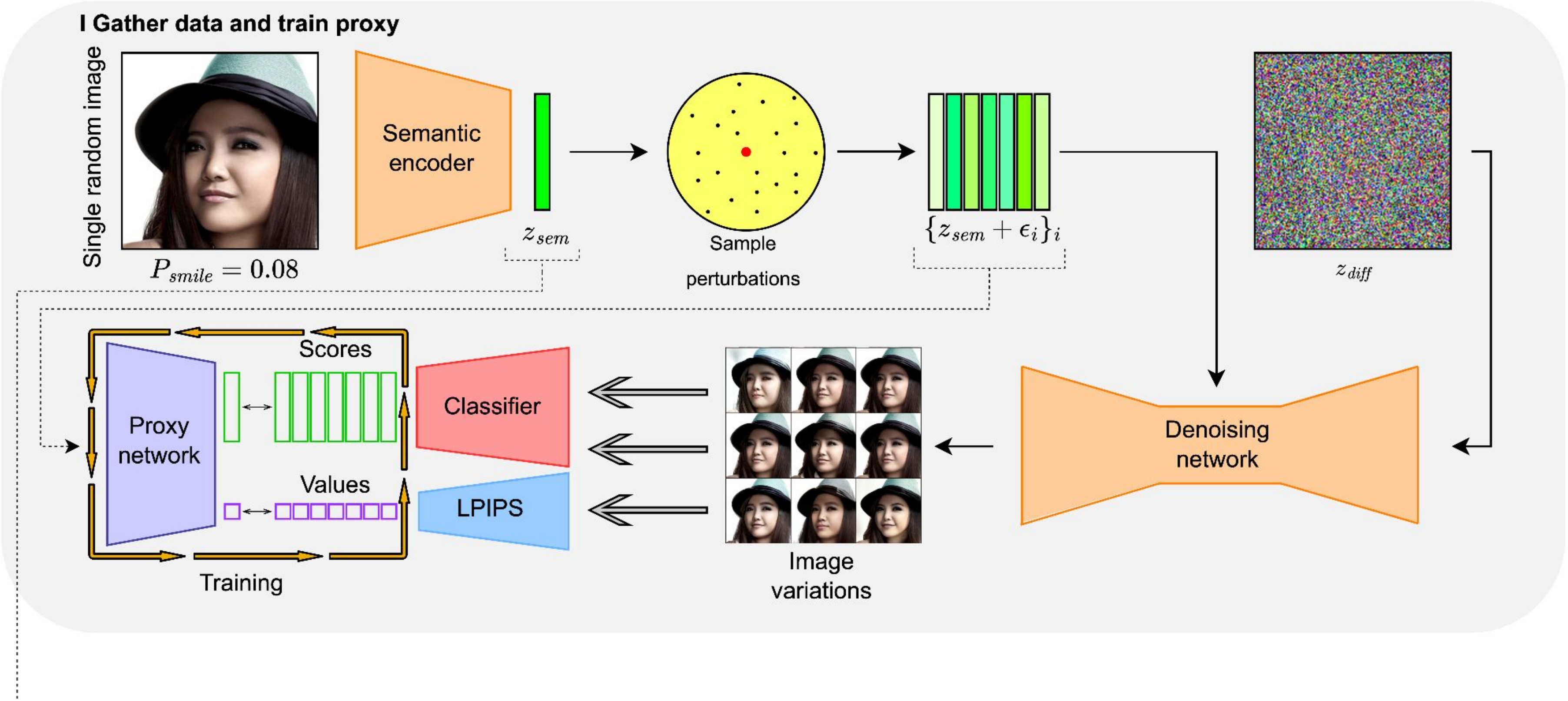




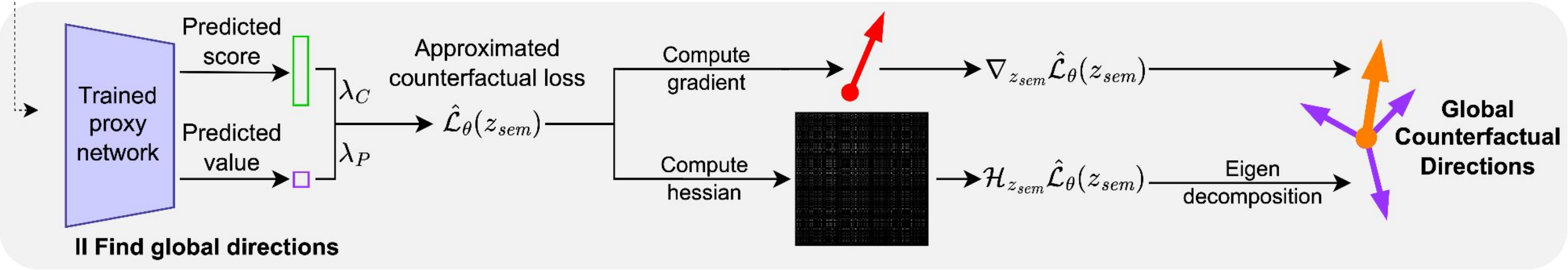


Global Counterfactual Directions for Black-Box Models

Local approximation of the classifier



Extracting influential directions



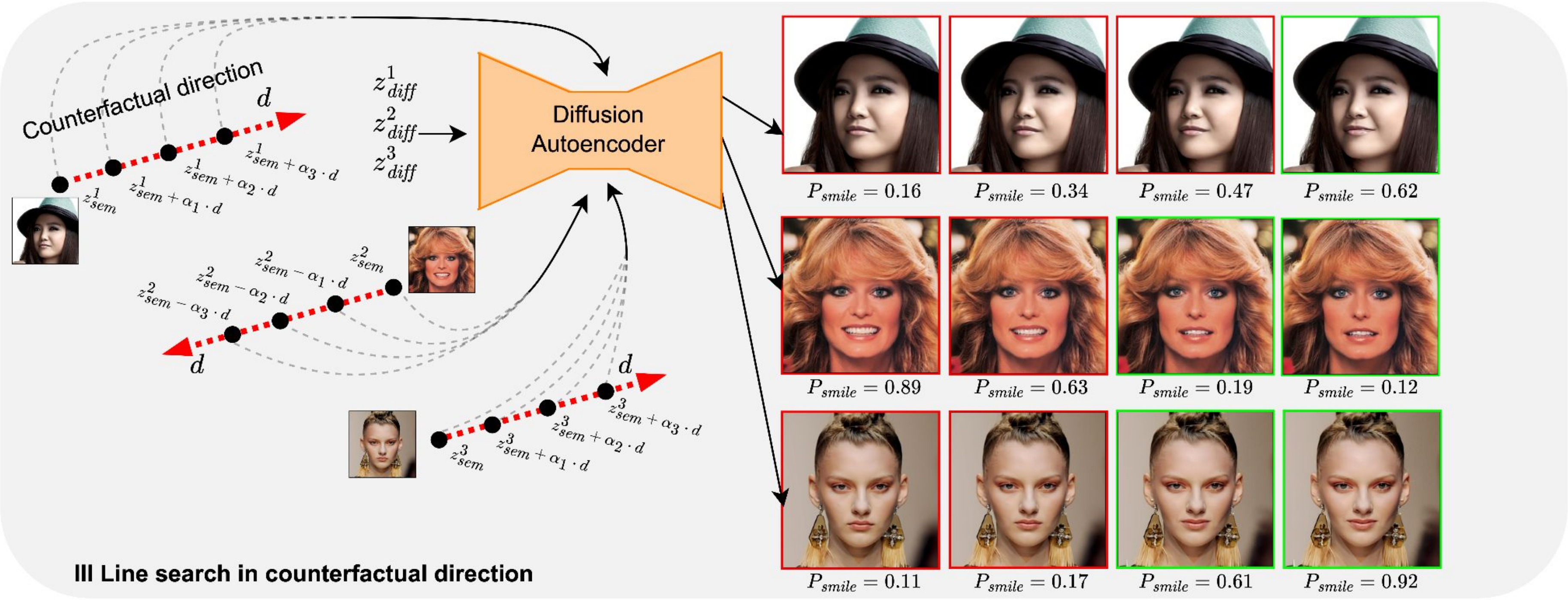
g-direction

$$\mathbf{d}_g = \nabla_{\mathbf{z}_{sem}}(p^f_{\boldsymbol{\theta}}(\mathbf{z}_{sem}) + \lambda p^s_{\boldsymbol{\theta}}(\mathbf{z}_{sem}))$$

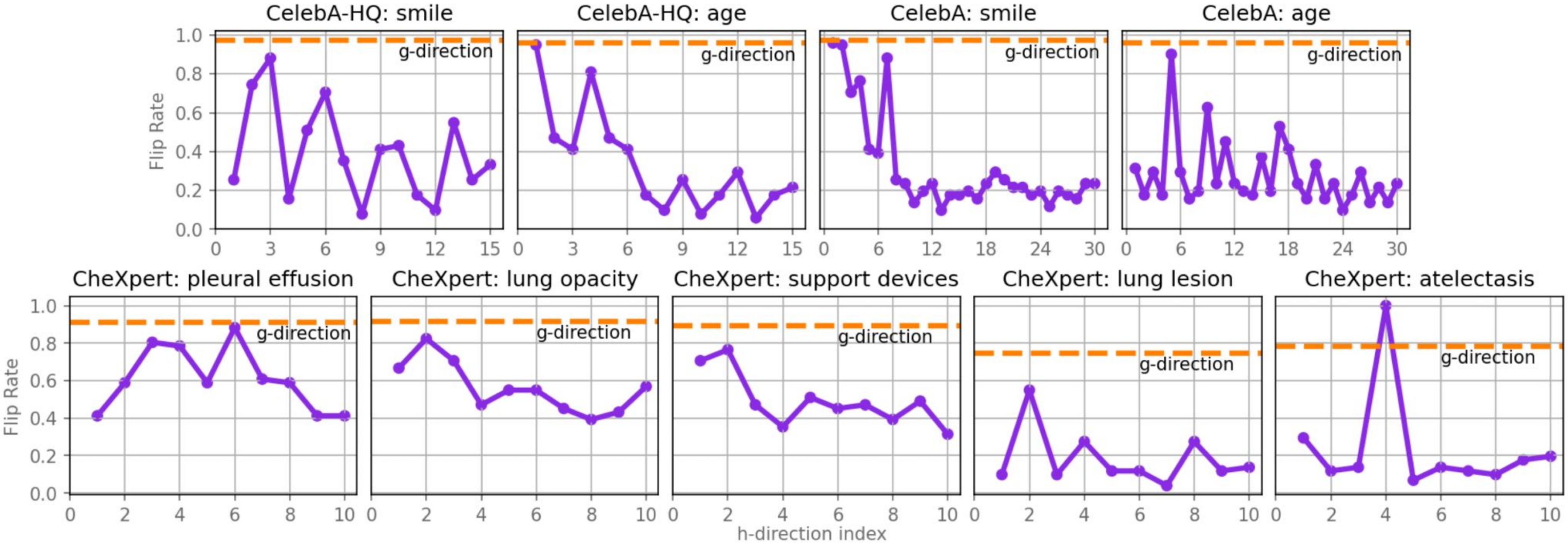
h-directions

$$\mathbf{H} = \nabla^2_{\mathbf{z}_{sem}}(p^f_{\boldsymbol{\theta}}(\mathbf{z}_{sem}) + \lambda p^s_{\boldsymbol{\theta}}(\mathbf{z}_{sem})) \to \{\mathbf{d}^i_h\}_i$$

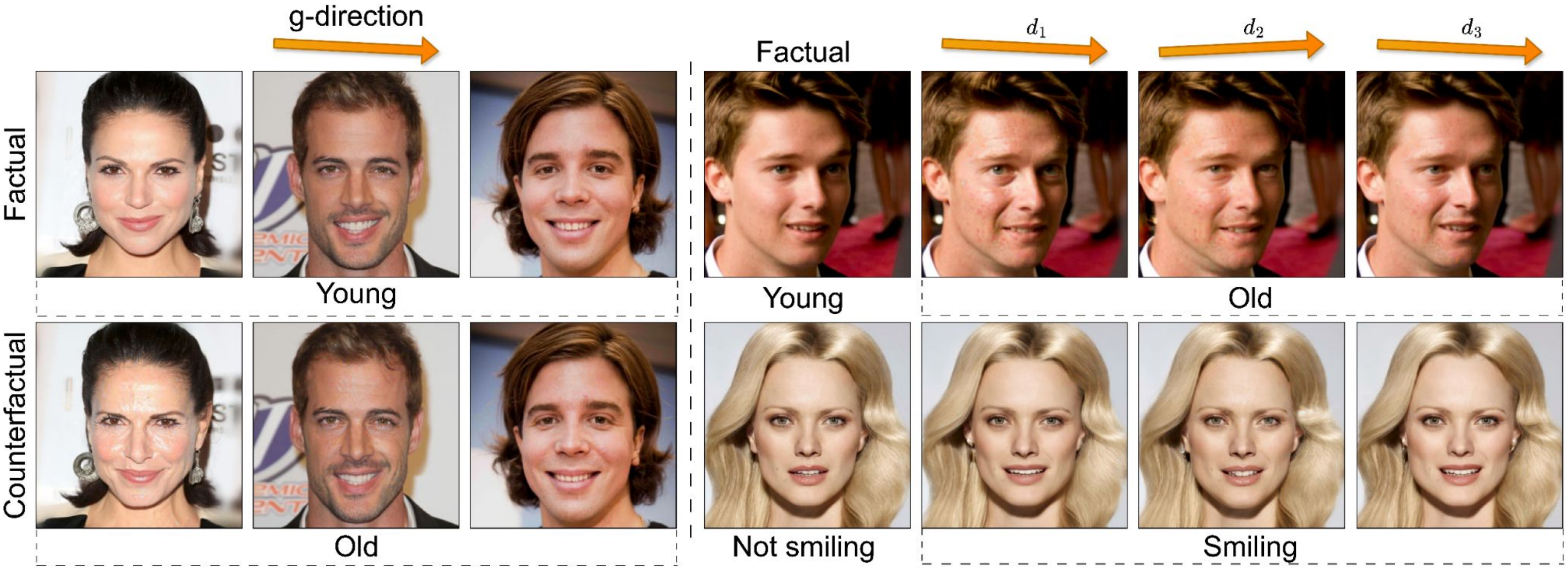
Directions from a single image are global!



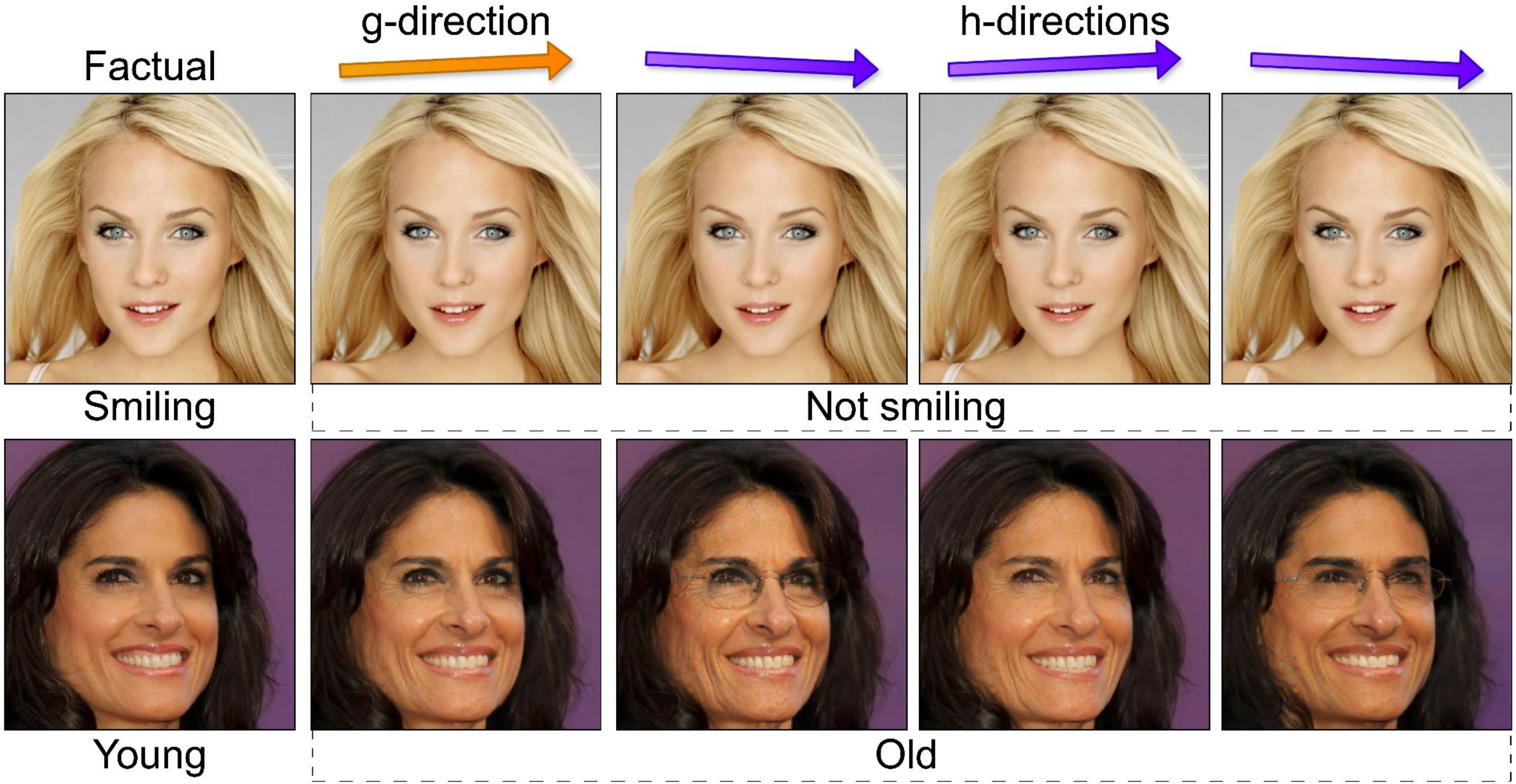
Quantitative assessment of globality



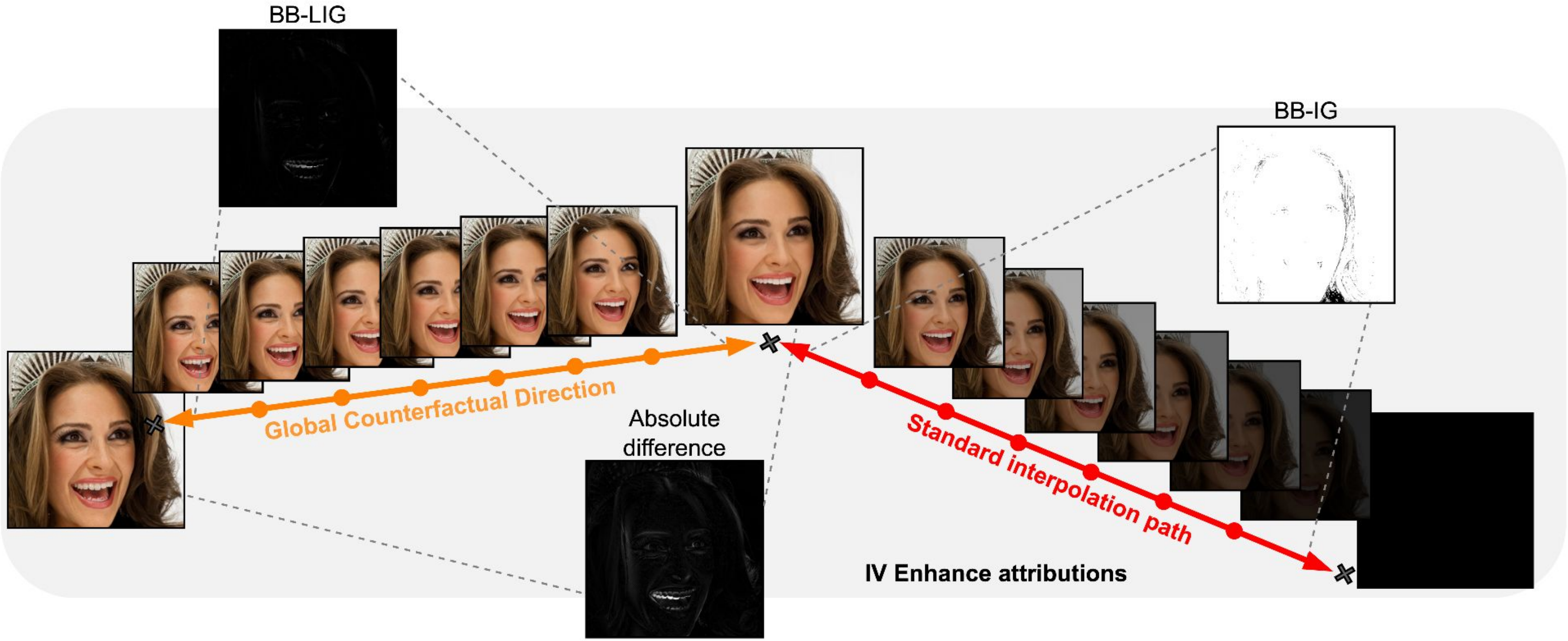
g-directions from different images differ



h-directions increase diversity



Black-box Latent Integrated Gradients



Enhanced explanation evaluation with GCD

$$BB\text{-}LIG_i(\mathbf{x}) = \frac{1}{m-1}(\mathbf{x}_i - \mathbf{x}'_i)\sum_{k=1}^{m-1}\frac{f(y \mid \tilde{\mathbf{x}}^{k+1}) - f(y \mid \tilde{\mathbf{x}}^{k})}{\tilde{\mathbf{x}}_i^{k+1} - \tilde{\mathbf{x}}_i^{k}}$$

Conclusions

Takeaways

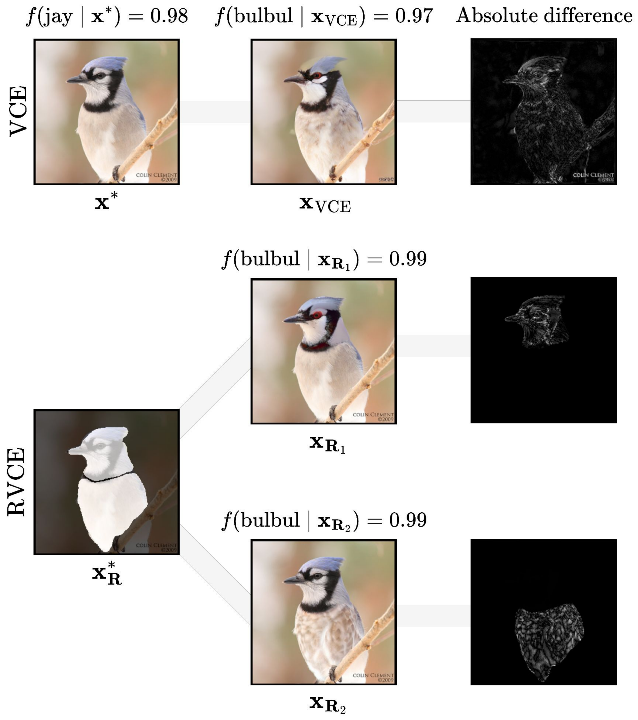
1. Black-box VCEs are possible but can they scale to very large datasets like ImageNet?

2. Extensions of standard generative frameworks offer highly non-trivial applications to XAI!

Follow-up

Rethinking Visual Counterfactual Explanations Through Region Constraint

arXiv



Want more?

GCD on arXiv

My LinkedIn

Our research group

Let's grab a coffee/beer and catch for a chat about (X)AI!

Thank you for attention

References

1. Song et al., *Score-based Generative Modeling Through Stochastic Differential Equations*, ICLR 2021,
2. Liu et al., *I2SB: Image-to-Image Schrödinger Bridge*, ICML 2023,
3. Sobieski and Biecek, *Global Counterfactual Directions*, ECCV 2024,
4. Sobieski et al., *Rethinking Visual Counterfactual Explanations Through Region Constraint*, arXiv 2024,
5. Saha, S., *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way*, Medium 2018,
6. Preechakul et al., *Diffusion Autoencoders: Toward a Meaningful and Decodable Representation*, CVPR 2022
7. Jeanneret et al., *Adversarial Visual Counterfactual Explanations*, CVPR 2023