Rethinking Visual Counterfactual Explanations Through Region Constraint

Bartlomiej Sobieski

Authors

Rethinking Visual Counterfactual Explanations Through Region Constraint

Bartlomiej Sobieski *
University of Warsaw
b.sobieski@uw.edu.pl

Jakub Grzywaczewski
Warsaw University of Technology
jakub.grzywaczewski2.stud@pw.edu.pl

Bartlomiej Sadlej
University of Warsaw
b.sadlej@student.uw.edu.pl

Matthew Tivnan
Harvard Medical School
mtivnan@mgh.harvard.edu

Przemyslaw Biecek
University of Warsaw, Warsaw University of Technology
przemyslaw.biecek@gmail.com

Bartłomiej Sobieski

Jakub Grzywaczewski

Bartłomiej Sadlej

Matthew Tivnan

Przemysław Biecek

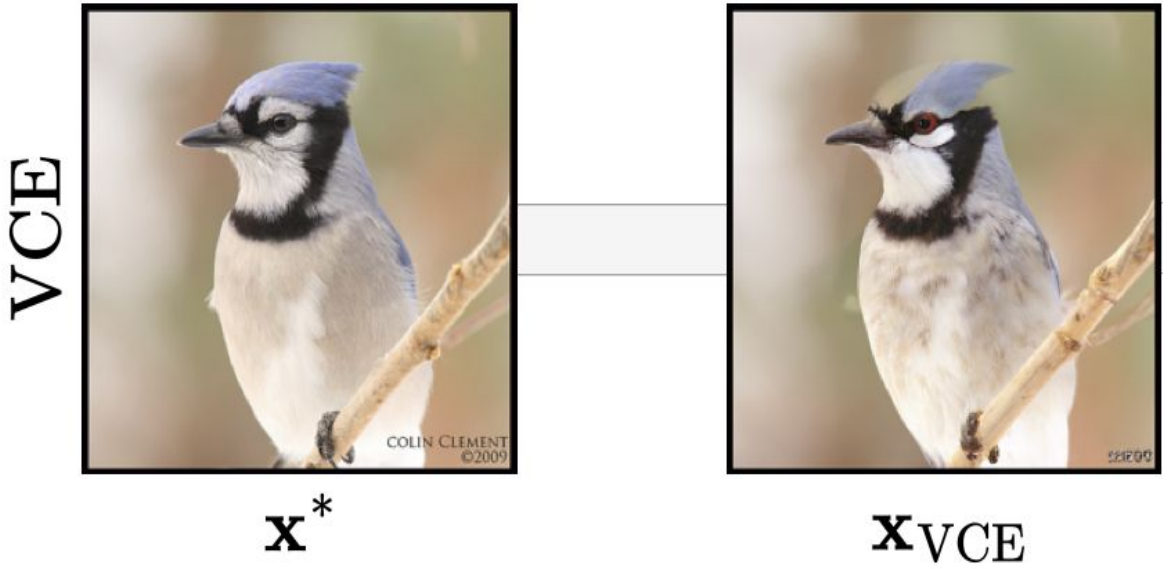
Outline

1. What is wrong with Visual Counterfactual Explanations?
2. Score-based Generative Models and Schrödinger Bridges
3. Region-constrained Counterfactual Schrödinger Bridge

What is wrong with Visual Counterfactual Explanations?

Visual Counterfactual Explanations (VCEs)

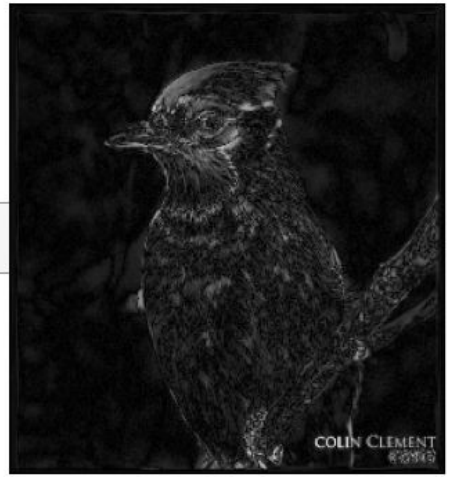
$$f(\text{jay} \mid \mathbf{x}^*) = 0.98 \qquad f(\text{bulbul} \mid \mathbf{x}_{\text{VCE}}) = 0.97$$



What is wrong?

$f(\text{jay} \mid \mathbf{x}^*) = 0.98$ $f(\text{bulbul} \mid \mathbf{x}_{\text{VCE}}) = 0.97$ Absolute difference

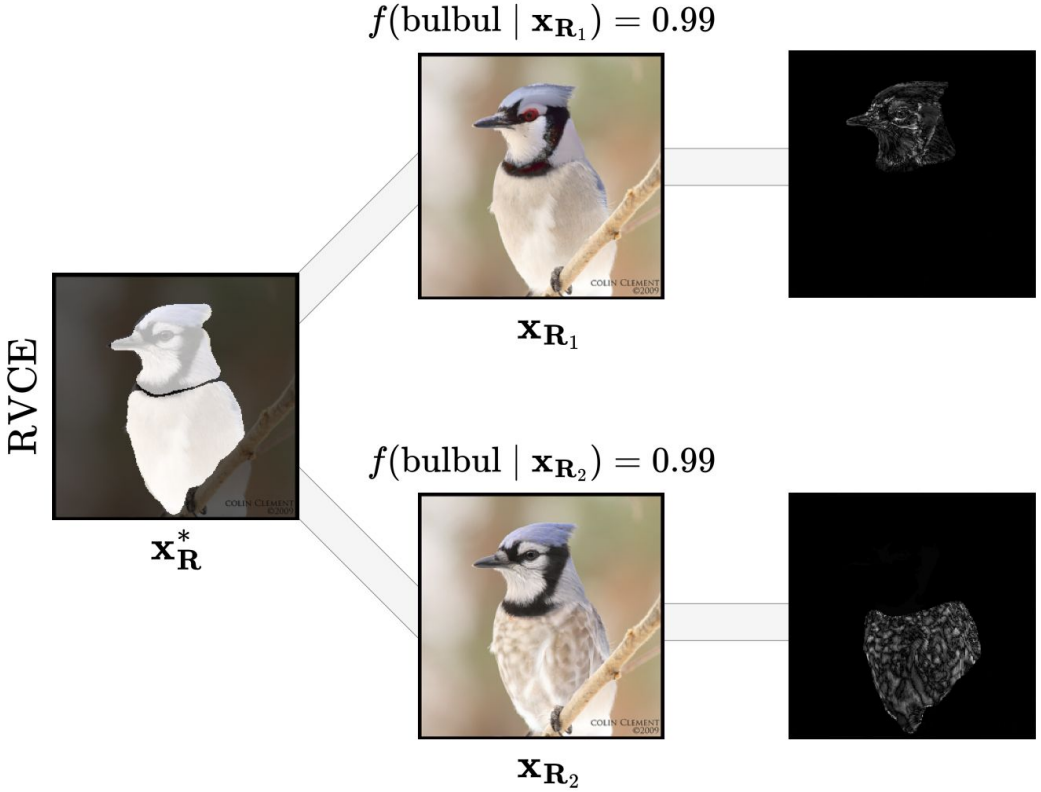
VCE



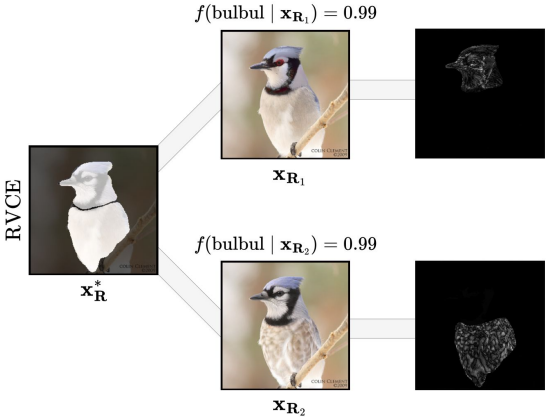
$\mathbf{x}^*$ $\mathbf{x}_{\text{VCE}}$

Introducing the region constraint

Region-constrained Visual Counterfactual Explanations



What are RVCEs?



Sample different explanations

Keep the region's complement unchanged

$$p(\mathbf{x} \mid \text{argmax}_{y'} f(y' \mid \mathbf{x}) = y, (\mathbf{1} - \mathbf{R}) \odot \mathbf{x} = (\mathbf{1} - \mathbf{R}) \odot \mathbf{x}^*)$$

Maximize classifier's probability for this class

Original image

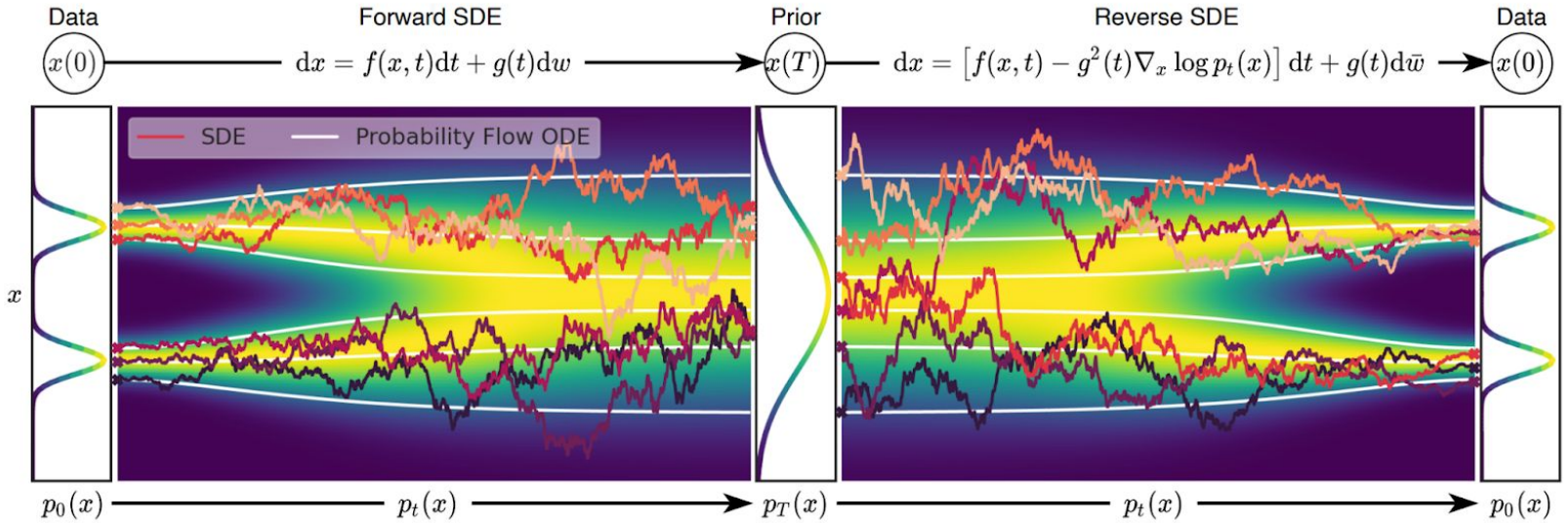
How to sample meaningful RVCEs?

We need

1. a strong prior for generating images from the data manifold
2. a way to condition the generation on the classifier

Score-based Generative Models and Schrödinger Bridges

SOTA image synthesis with score-based generative models



SGM

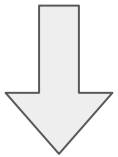
Adapting SGMs to conditional generation

Forward equation

$$\mathrm{d}\mathbf{x}_t = \mathbf{F}_t(\mathbf{x}_t)\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\mathbf{w}$$

Reverse equation

$$\mathrm{d}\mathbf{x}_t = (\mathbf{F}_t(\mathbf{x}_t) - \beta_t \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t))\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\bar{\mathbf{w}}$$

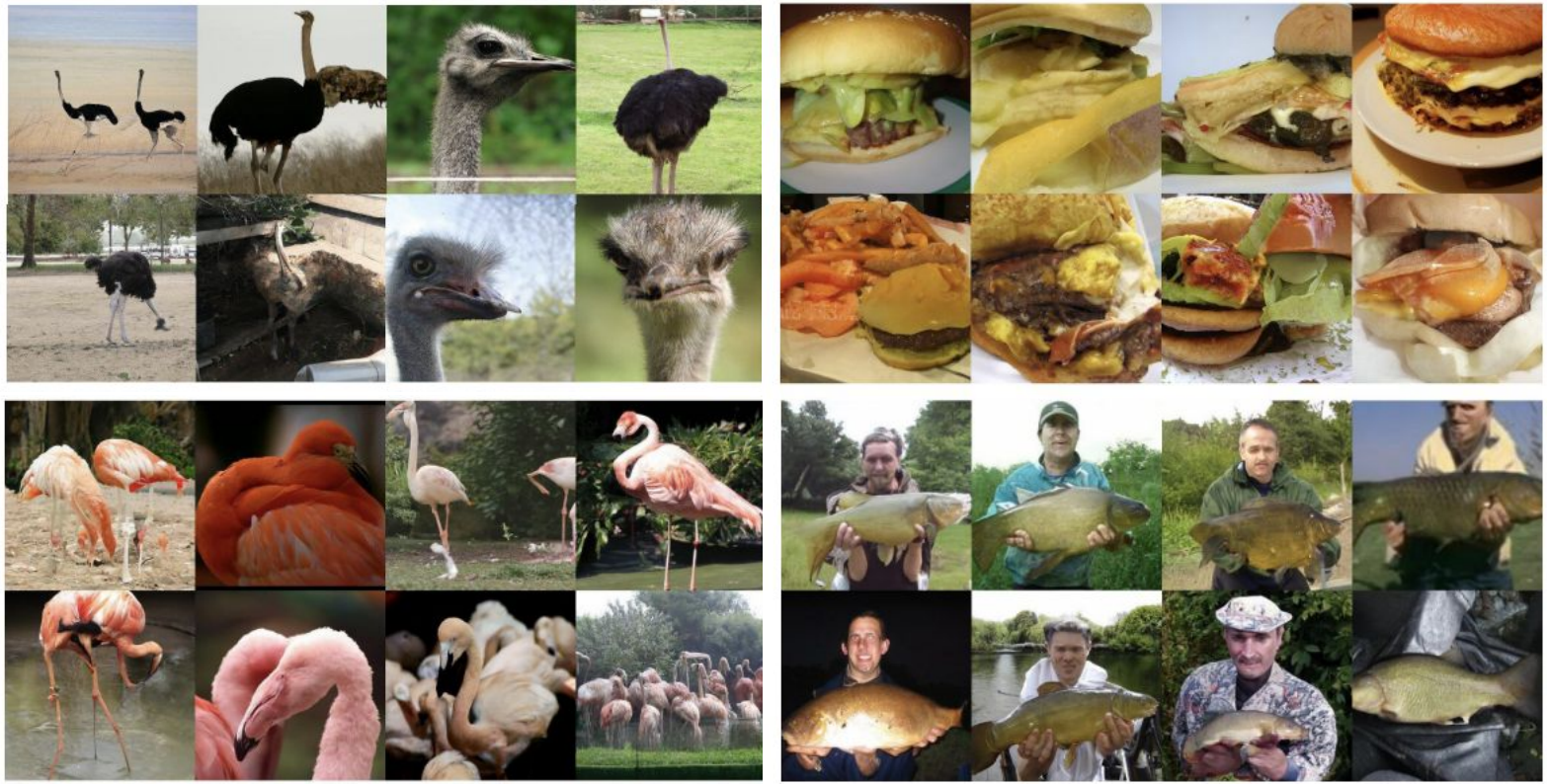


Bayes' Theorem

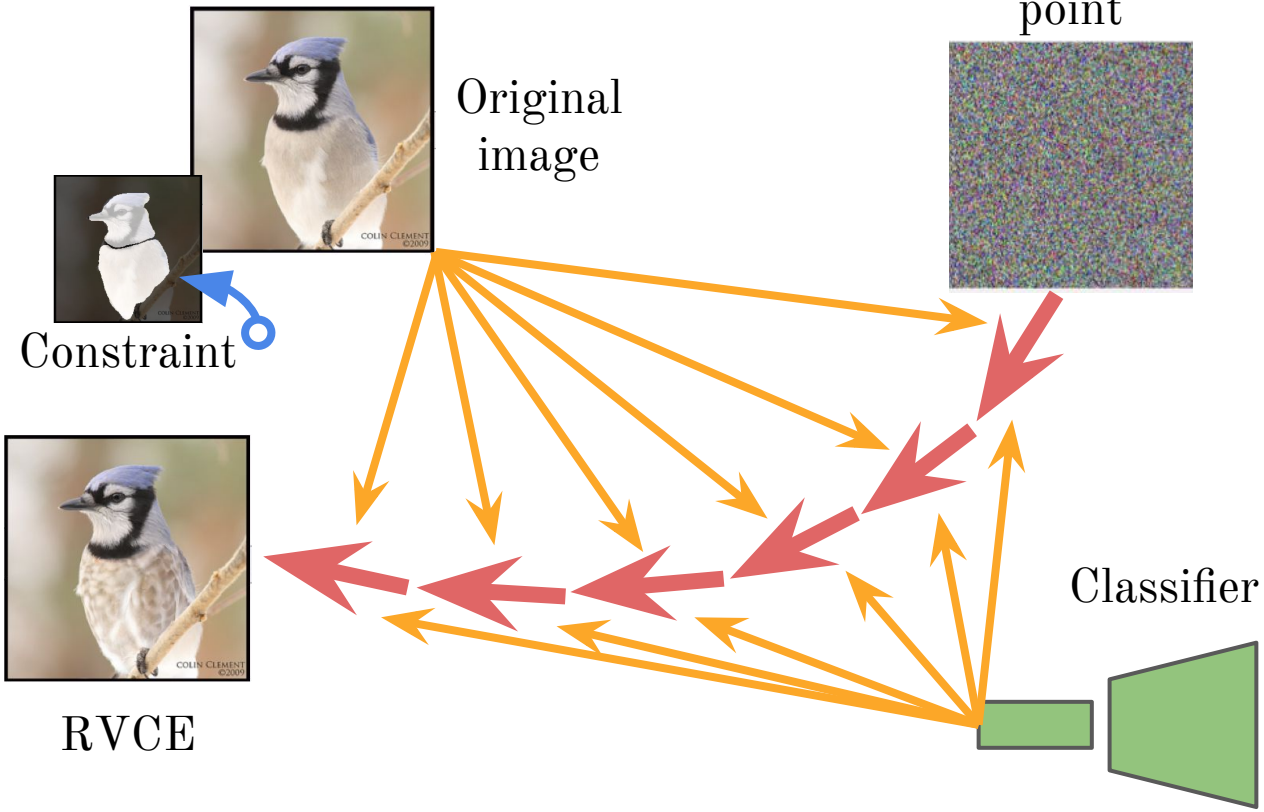
Conditional reverse equation

$$\mathrm{d}\mathbf{x}_t = (\mathbf{F}_t(\mathbf{x}_t) - \beta_t[\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t, t)])\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\bar{\mathbf{w}}$$

Sampling from conditional distributions



Suboptimal strategy

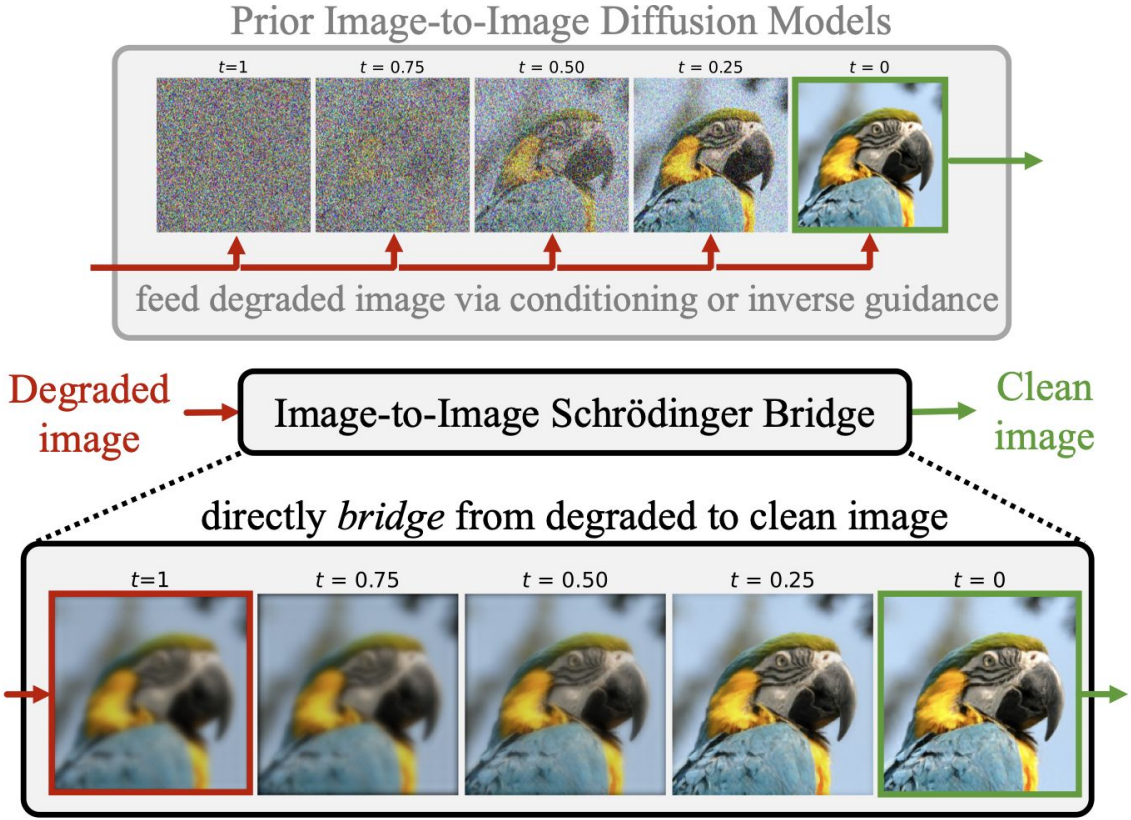


Why not?

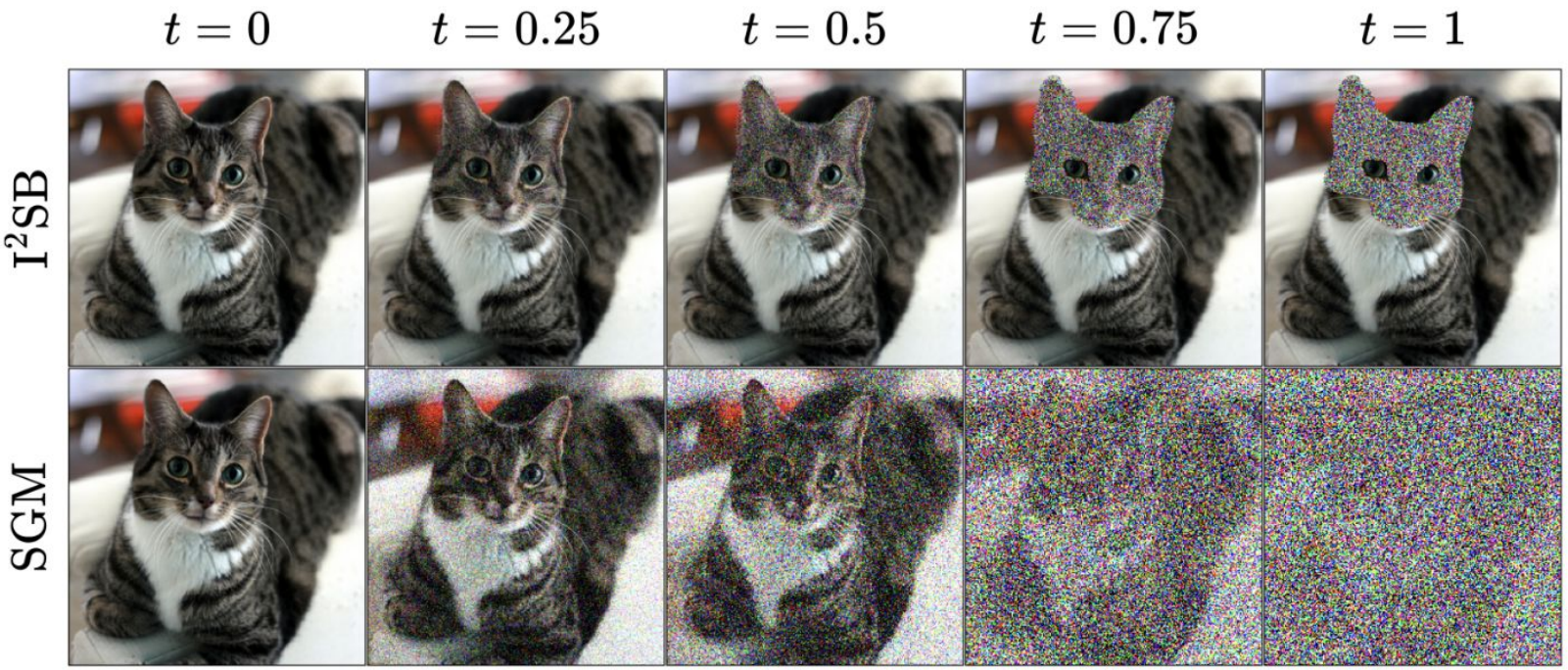
1. Conflicting gradients from two sources.
2. Does not utilize the information available from the outside of the region.

Can we map images with indicated regions directly to RVCEs?

Use Image-to-Image Schrödinger Bridges instead



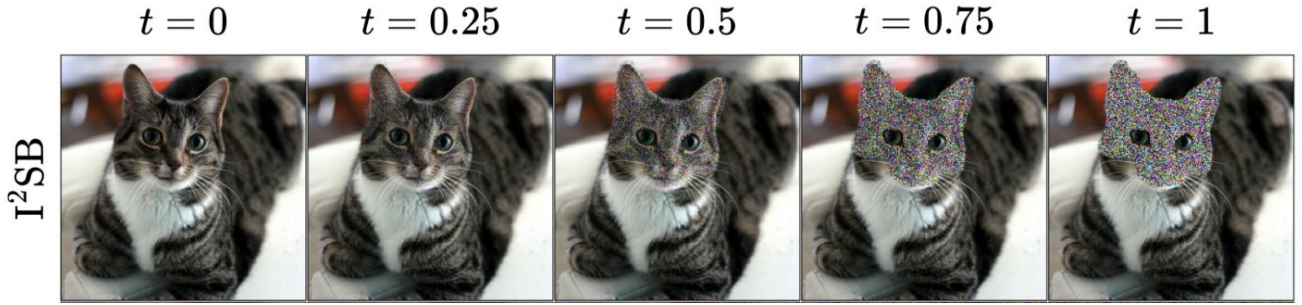
Map masked images directly



Gradient alignment problem

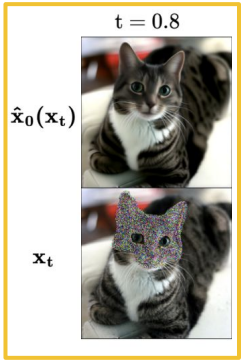
$$\mathrm{d}\mathbf{x}_t = (\mathbf{F}_t(\mathbf{x_t}) - \beta_t[\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t, t)])\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\bar{\mathbf{w}}$$

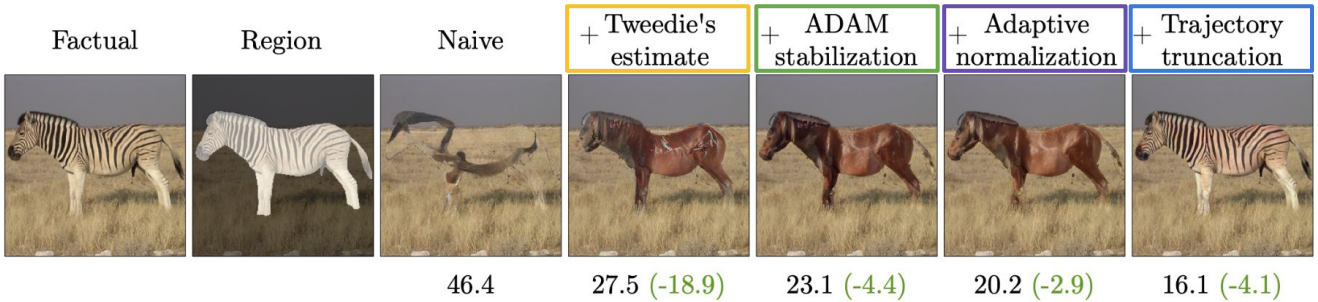
Never seen by a non-robust classifier



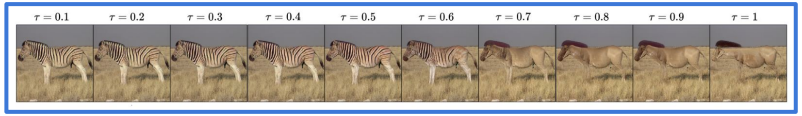
Region-constrained Counterfactual Schrödinger Bridge

Aligning the classifier's gradients





$$\hat{\mathbf{x}}_0(\mathbf{x}_t) := \mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t] = \mathbf{x}_t + \sigma_t^2 \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t)$$



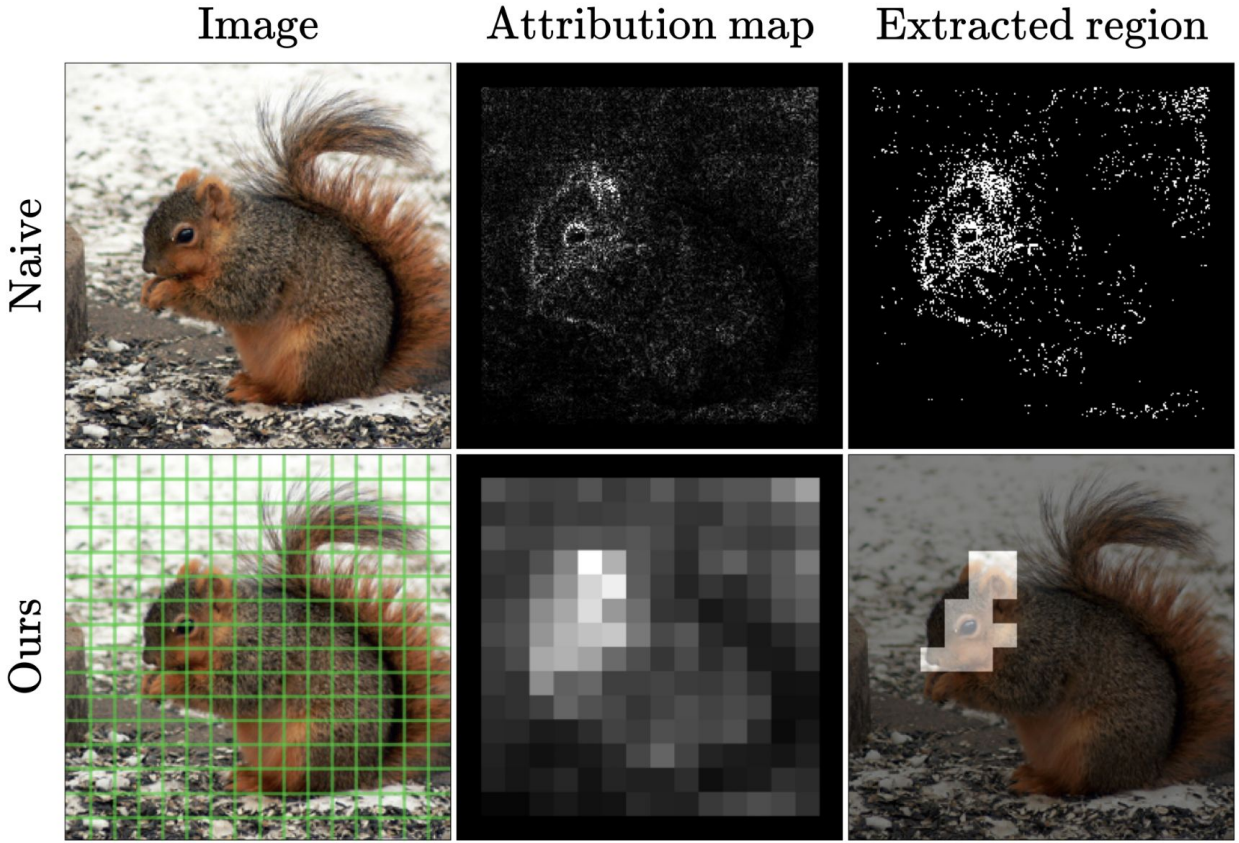
Algorithm 4 ADAM Update Rule

Input: Gradient at step n $\mathbf{g}_n$, hyperparameters $\alpha, \epsilon, \beta_1, \beta_2$ (set to PyTorch (Paszke et al., 2019) defaults)
$\mathbf{m}_n = \beta_1 \mathbf{m}_{n-1} + (1-\beta_1)\mathbf{g}_n$ # update biased first moment estimate
$\mathbf{v}_n = \beta_2 \mathbf{v}_{n-1} + (1-\beta_2)\mathbf{g}_n^2$ # update biased second moment estimate
$\hat{\mathbf{m}}_n = \mathbf{m}_n / (1-\beta_1^n)$ # compute bias-corrected first moment
$\hat{\mathbf{v}}_n = \mathbf{v}_n / (1-\beta_2^n)$ # compute bias-corrected second moment
$\bar{\mathbf{g}}_n = \alpha \hat{\mathbf{m}}_n / (\sqrt{\hat{\mathbf{v}}_n} + \epsilon)$ # update gradient
return $\bar{\mathbf{g}}_n$ # return updated gradient

Algorithm 3 RCSB

Input: Number of steps N, binary region mask $\mathbf{R}$, trajectory truncation τ, classifier scale s, input image $\mathbf{x}^*$, trained $\mathbf{s}_\psi(\cdot,\cdot)$, trained classifier $f(\mathbf{y} \mid \cdot)$, target class y
$\mathbf{x}_1 = (\mathbf{1} - \mathbf{R}) \odot \mathbf{x}^* + \mathbf{R} \odot \mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$
Discretize truncated timeline $0 = t_0 < t_1 < \cdots < t_N = \tau$
$\mathbf{x}_N \sim q(\mathbf{x}_N | \mathbf{x}_0, \mathbf{x}_1)$ # sample from analytic posterior (Eq. (15))
for $n = N$ **to** 1 **do**
  Predict $\hat{\mathbf{x}}_0(\mathbf{x}_n)$ using $\mathbf{s}_\psi(\mathbf{x}_n, t_n)$
  $\mathbf{g}_n = \nabla_{\mathbf{x}_n} \log f(y \mid \hat{\mathbf{x}}_0)$
  $\bar{\mathbf{g}}_n = \text{ADAM}(\mathbf{g}_n)$
  if $n == N$ **then**
    $g = \|\bar{\mathbf{g}}_N\|_2$ # register norm of the first gradient
  end if
  $\bar{\mathbf{x}}_n = \mathbf{x}_n + s \frac{\bar{\mathbf{g}}_n}{g}$
  $\mathbf{x}_{n-1} = \mu_{n-1} \hat{\mathbf{x}}_0 + \bar{\mu}_{n-1} \bar{\mathbf{x}}_n$
end for
return $\mathbf{x}_0$

But what about the regions?



Results

New state-of-the-art on ImageNet

Method	FID	sFID	S^3	COUT	FR
Zebra – Sorrel					
ACE l_1	84.5	122.7	**0.92**	−0.45	47.0
ACE l_2	67.7	98.4	0.90	−0.25	81.0
LDCE-cls	84.2	107.2	0.78	−0.06	88.0
LDCE-txt	82.4	107.2	0.71	−0.21	81.0
DVCE	33.1	43.9	0.62	−0.21	57.8
RCSB^C	13.0	20.4	0.82	0.70	**99.7**
RCSB^B	9.51	17.4	0.86	0.72	97.4
RCSB^A	**8.0**	**16.2**	0.88	**0.74**	94.7

Method	FID	sFID	S^3	COUT	FR
Cheetah – Cougar					
ACE l_1	70.2	100.5	0.91	0.02	77.0
ACE l_2	74.1	102.5	0.88	0.12	95.0
LDCE-cls	71.0	91.8	0.62	0.51	100.0
LDCE-txt	91.2	117.0	0.59	0.34	98.0
DVCE	46.9	54.1	0.70	0.49	99.0
RCSB^C	30.2	39.2	0.87	0.79	100.0
RCSB^B	23.4	32.4	0.90	0.85	99.9
RCSB^A	**17.2**	**26.6**	**0.92**	**0.92**	**100.0**

Method	FID	sFID	S^3	COUT	FR
Egyptian Cat – Persian Cat					
ACE l_1	93.6	156.7	0.85	0.25	85.0
ACE l_2	107.3	160.4	0.78	0.34	97.0
LDCE-cls	102.7	140.7	0.63	0.52	99.0
LDCE-txt	121.7	162.4	0.61	0.56	99.0
DVCE	46.6	59.2	0.59	0.60	98.5
RCSB^C	41.1	56.3	0.79	0.82	100.0
RCSB^B	31.3	48.1	0.84	0.87	100.0
RCSB^A	**23.0**	**40.0**	**0.87**	**0.92**	**100.0**

- Extreme realism indicated by 2-4x better FID and sFID
- Nearly maximal sparsity (COUT) for the first time on ImageNet
- Explanations do not deviate far from the original image (S^3)
- Flip Rate reaches the upper bound

CelebA

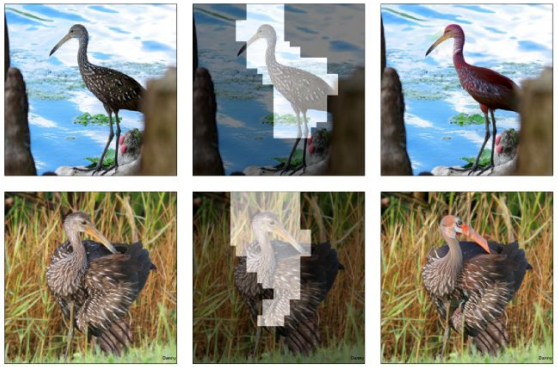
	Smile								Age							
Method	FID	sFID	FVA	FS	MNAC	CD	COUT	FR	FID	sFID	FVA	FS	MNAC	CD	COUT	FR
DiVE	29.4	-	97.3	-	-	-	-	-	33.8	-	98.1	-	4.58	-	-	-
DiVE^{100}	36.8	-	73.4	-	4.63	2.34	-	-	39.9	-	52.2	-	4.27	-	-	-
STEEX	10.2	-	96.9	-	4.11	-	-	-	11.8	-	97.5	-	3.44	-	-	-
ACE ℓ_1	1.27	3.97	99.9	0.87	2.94	1.73	0.78	97.6	1.45	4.12	99.6	0.78	3.20	2.94	0.72	96.2
ACE ℓ_2	1.90	4.56	99.9	0.87	2.77	1.56	0.62	84.3	2.08	4.62	99.6	0.80	2.94	2.82	0.56	77.5
DiME	3.17	4.89	98.3	0.73	3.72	2.30	0.53	97.0	4.15	5.89	95.3	0.67	3.13	3.27	0.44	99.0
FastDiME	4.18	6.13	99.8	0.76	3.12	1.91	0.44	99.0	4.82	6.76	99.2	0.74	2.65	3.80	0.36	98.6
FastDiME-2	3.33	5.49	99.9	0.77	3.06	1.89	0.44	99.4	4.04	6.01	99.6	0.75	2.63	3.80	0.37	99.3
FastDiME-2+	3.24	5.23	99.9	0.79	2.91	2.02	0.41	98.9	3.60	5.59	99.7	0.77	2.44	3.76	0.32	98.7
RCSB	2.98	4.79	100.0	0.91	2.24	2.78	0.87	99.8	2.94	4.94	99.9	0.88	2.14	3.63	0.81	99.3

and CelebA-HQ

	Smile								Age							
Method	FID	sFID	FVA	FS	MNAC	CD	COUT	FR	FID	sFID	FVA	FS	MNAC	CD	COUT	FR
DiVE	107.0	-	35.7	-	7.41	-	-	-	107.5	-	32.3	-	6.76	-	-	-
STEEX	21.9	-	97.6	-	5.27	-	-	-	26.8	-	96.0	-	5.63	-	-	-
DiME	18.1	27.7	96.7	0.67	2.63	1.82	0.65	97.0	18.7	27.8	95.0	0.66	2.10	4.29	0.56	97.0
ACE ℓ_1	3.21	20.2	100.0	0.89	1.56	2.61	0.55	95.0	5.31	21.7	99.6	0.81	1.53	5.4	0.40	95.0
ACE ℓ_2	6.93	22.0	100.0	0.84	1.87	2.21	0.60	95.0	16.4	28.2	99.6	0.77	1.92	4.21	0.53	95.0
LDCE	13.6	25.8	99.1	0.76	2.44	1.68	0.34	-	14.2	25.6	98.0	0.73	2.12	4.02	0.33	-
FastDiME-2+	16.51	31.4	99.9	0.87	1.43	4.16	0.28	87.1	26.0	40.3	99.6	0.81	3.15	4.36	0.31	92.6
RCSB	3.04	20.0	100.0	0.93	1.22	3.22	0.83	98.9	4.92	27.3	100.0	0.96	1.47	5.16	0.797	99.4

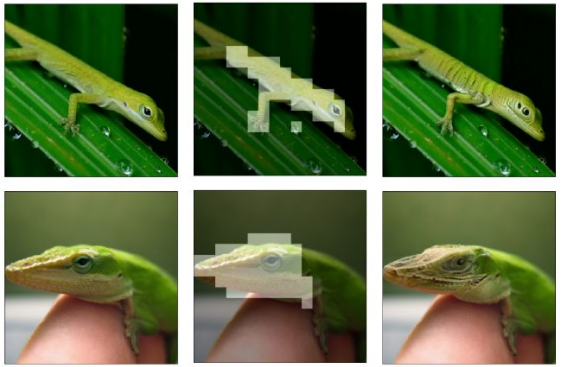
Infills of the most important regions

Limpkin → Flamingo

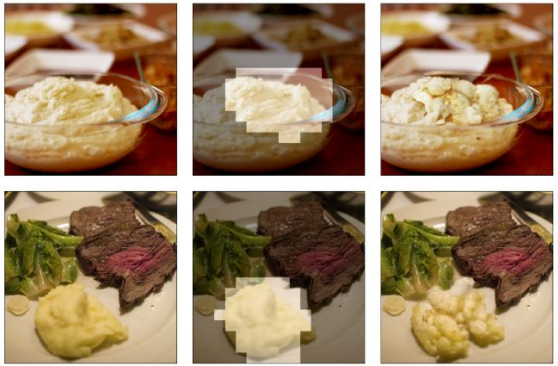


Guacamole → Cabbage

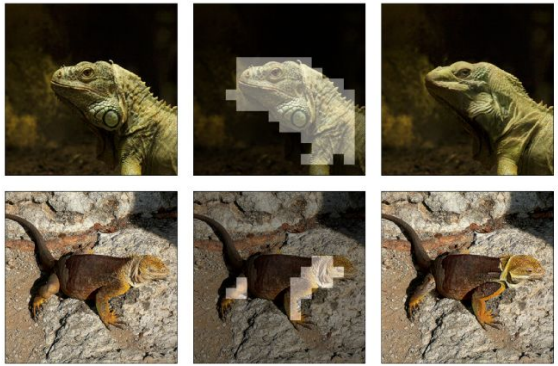
American Chameleon → Common Iguana



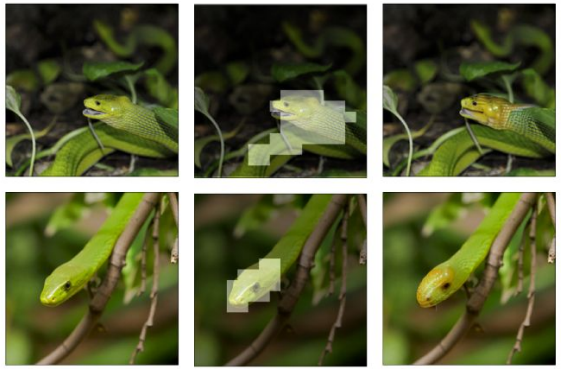
Mashed Potato → Cauliflower



Common Iguana → American Chameleon

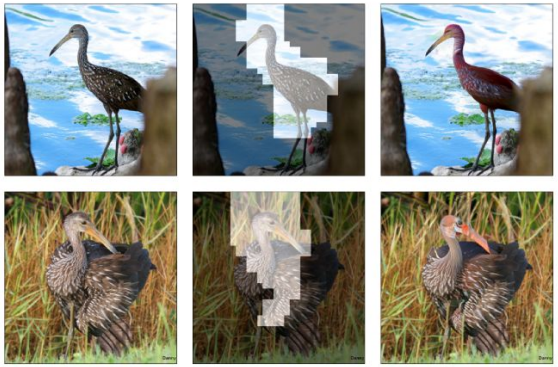


Green Mamba → Indian Cobra



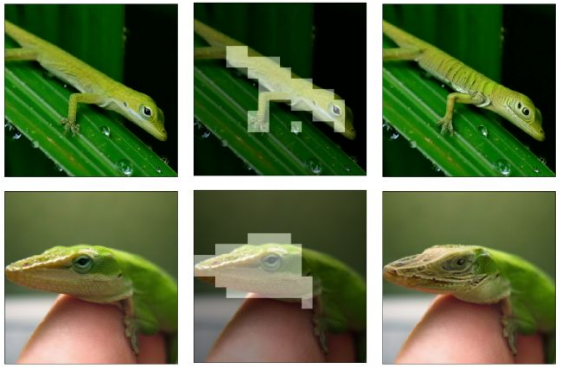
Infills of the most important regions

Limpkin → Flamingo

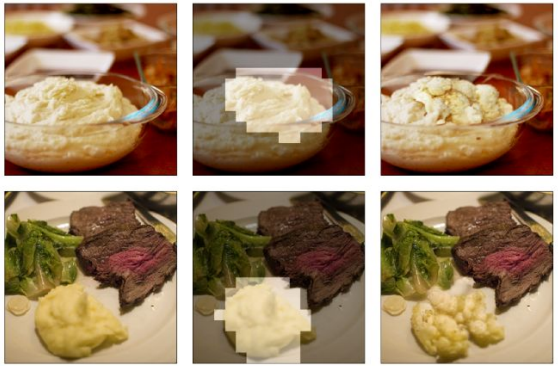


Guacamole → Cabbage

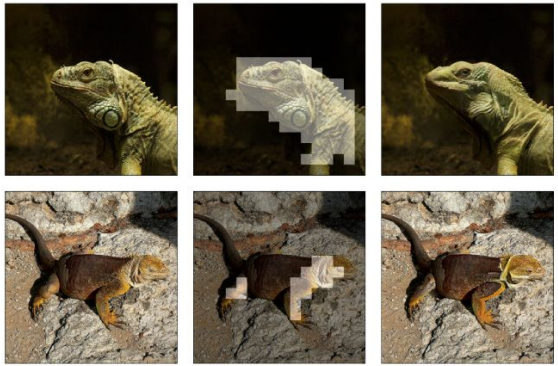
American Chameleon → Common Iguana



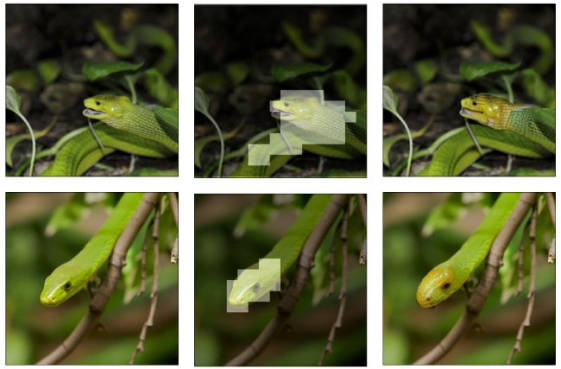
Mashed Potato → Cauliflower



Common Iguana → American Chameleon



Green Mamba → Indian Cobra



on various datasets

Smiling $\rightarrow$ Not smiling

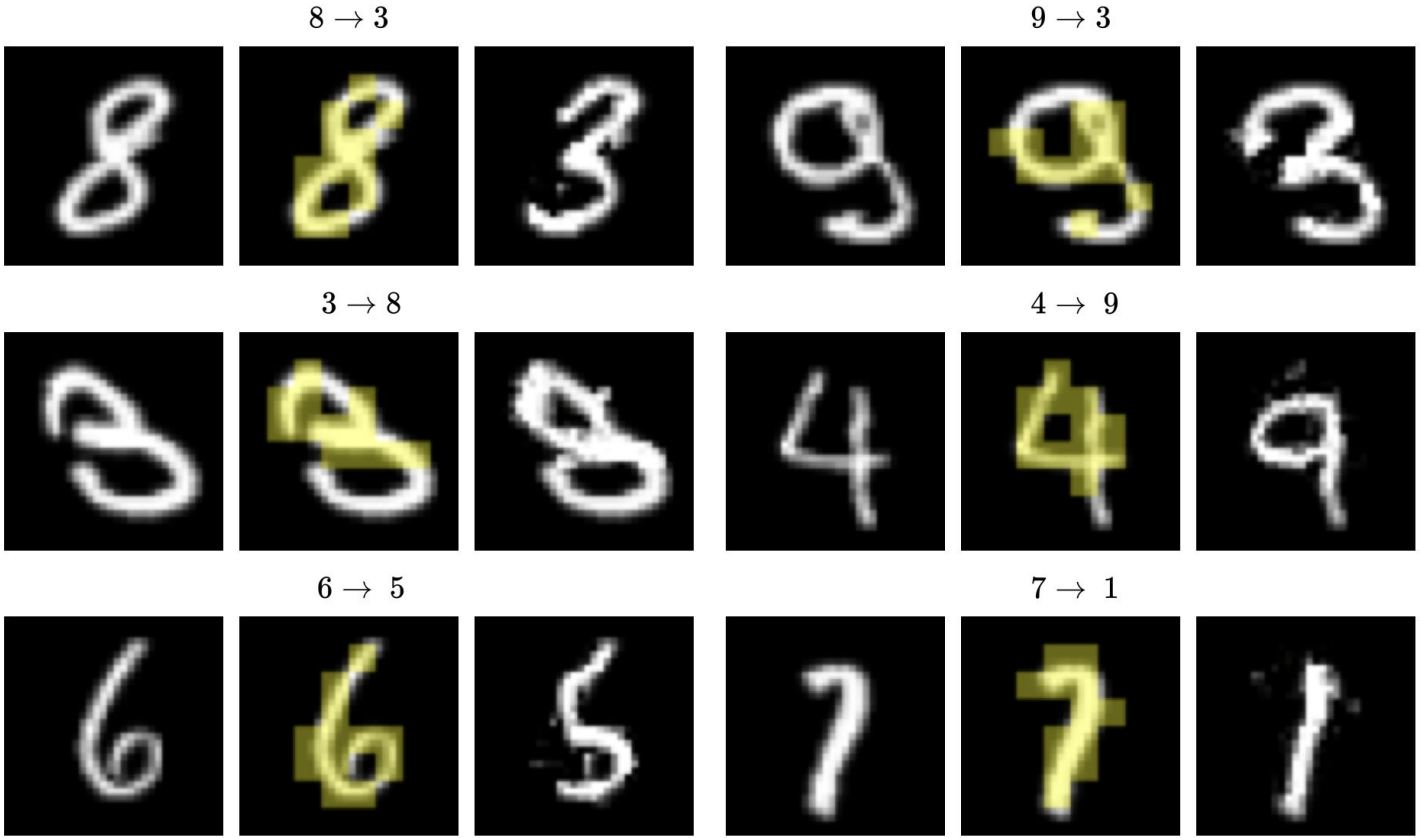
Not smiling $\rightarrow$ Smiling

on various datasets

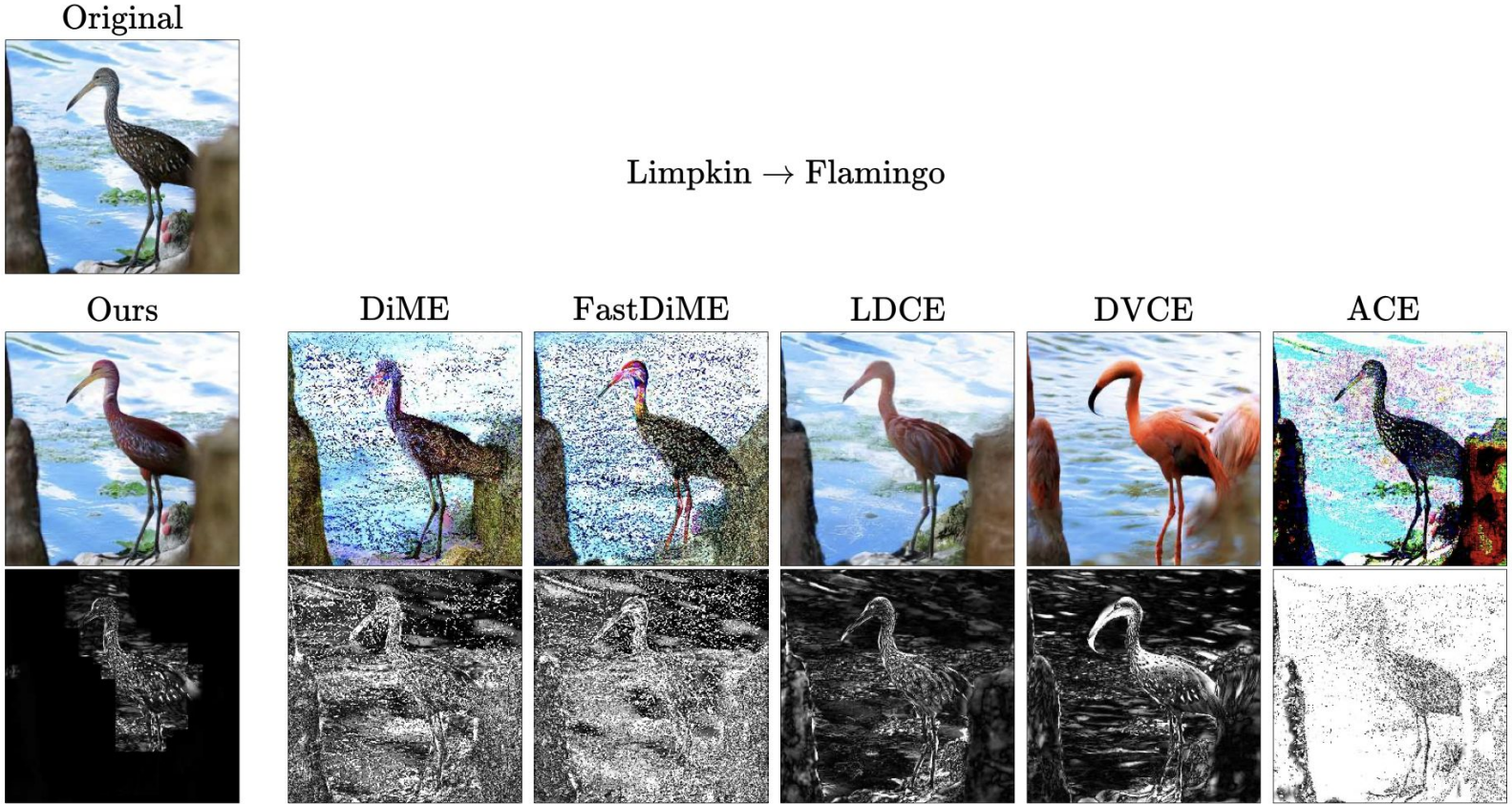
Young → Old

Old → Young

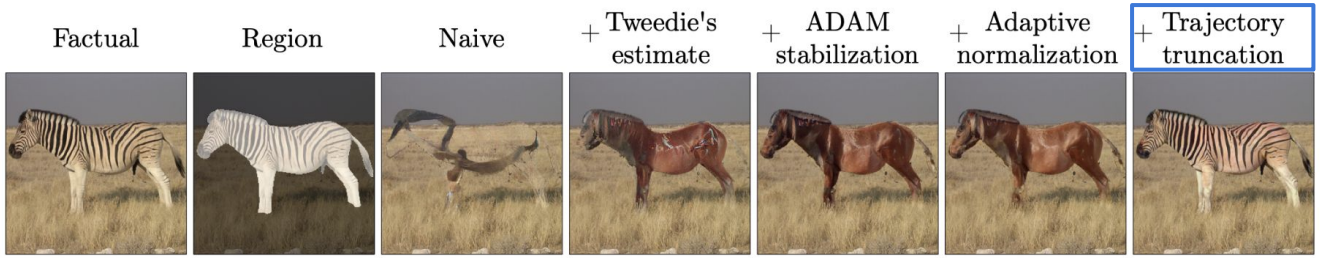
on various datasets

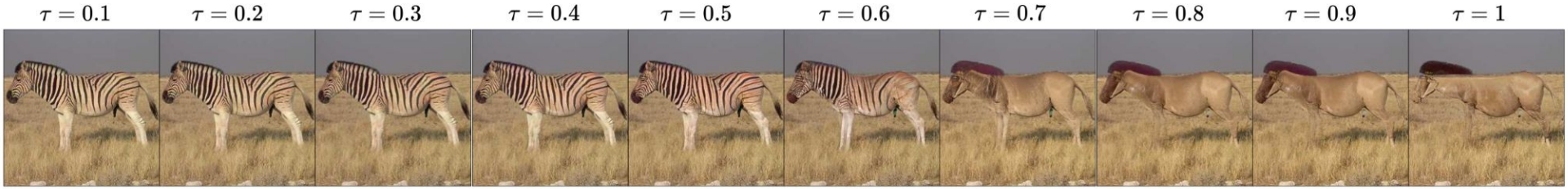


Outperforms other approaches qualitatively



Gives control over content preservation

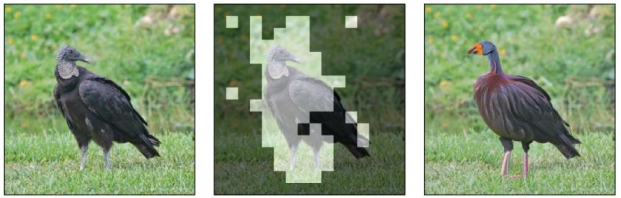




Allows for modifying shape, texture and color

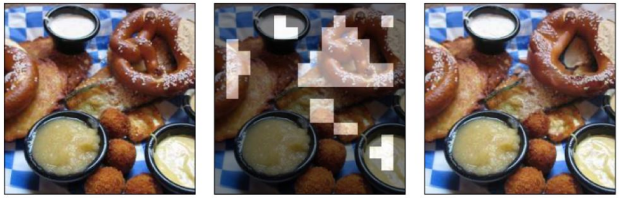
White stork → Black stork

Vulture → Flamingo



Spoon → Ladle

Pretzel → Bagel



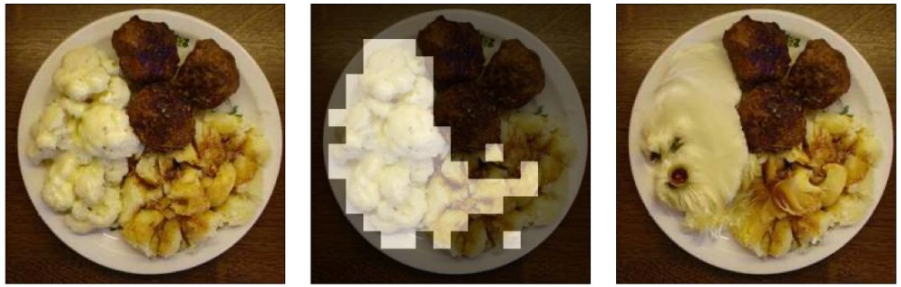
Hatchet → Hammer

Paperknife → Wooden spoon

	Pretzel → Bagel					**Hatchet → Hammer**					**Paperknife → Wooden spoon**				
Method	FID	sFID	S^3	COUT	FR	FID	sFID	S^3	COUT	FR	FID	sFID	S^3	COUT	FR
DVCE	34.3	43.9	0.59	0.37	77.4	31.2	39.8	0.66	0.43	92.8	29.1	35.4	0.69	0.41	88.2
$RCSB^A$	11.4	22.9	0.86	0.84	97.2	9.8	15.4	0.91	0.89	97.8	9.9	18.2	0.86	0.88	98.9

Works even when some things do not make sense!

Cauliflower → Maltese dog



Honeycomb → Hamster

Missile → Hamster

Purse → Hamster

Effective with various attribution methods and classifiers

Zebra – Sorrel					
Attribution method	FID	sFID	S^3	COUT	FR
LRP	**7.5**	**15.5**	**0.87**	0.62	93.6
InputXGradient	9.0	16.8	0.87	0.73	97.8
DeepLift	9.2	17.0	0.87	0.73	97.9
Integrated Gradients	9.5	17.4	0.86	0.72	97.4
GradientShap	10.5	18.5	0.87	0.74	97.4
LIME	12.9	20.7	0.85	0.55	88.4
GuidedBackprop	13.8	21.49	0.86	0.72	96.5
Occlusion	13.9	21.7	0.86	0.50	86.0
GradCAM	14.1	22.15	0.85	0.52	87.1
GuidedGradCAM	15.1	22.5	0.86	0.71	96.1
Saliency	15.2	23.0	0.86	**0.75**	**98.4**

	Zebra – Sorrel				
Classifier	FID	sFID	S^3	COUT	FR
ClipZeroShot	4.13	12.76	0.90	0.93	100.0
ConvNeXtBase	15.69	23.55	0.82	0.84	99.8
MadryResNet50	47.49	55.22	0.65	−0.19	36.2
RBDeiT	10.00	17.76	0.83	0.70	94.0
RBXCiT	16.04	23.45	0.79	0.46	83.6
SwinB	3.20	12.19	0.94	0.50	88.0
VGG16	7.29	15.39	0.88	0.84	98.0
VGG16_BN	5.44	13.53	0.91	0.87	99.9
ViTB16	8.60	16.84	0.86	0.80	98.9

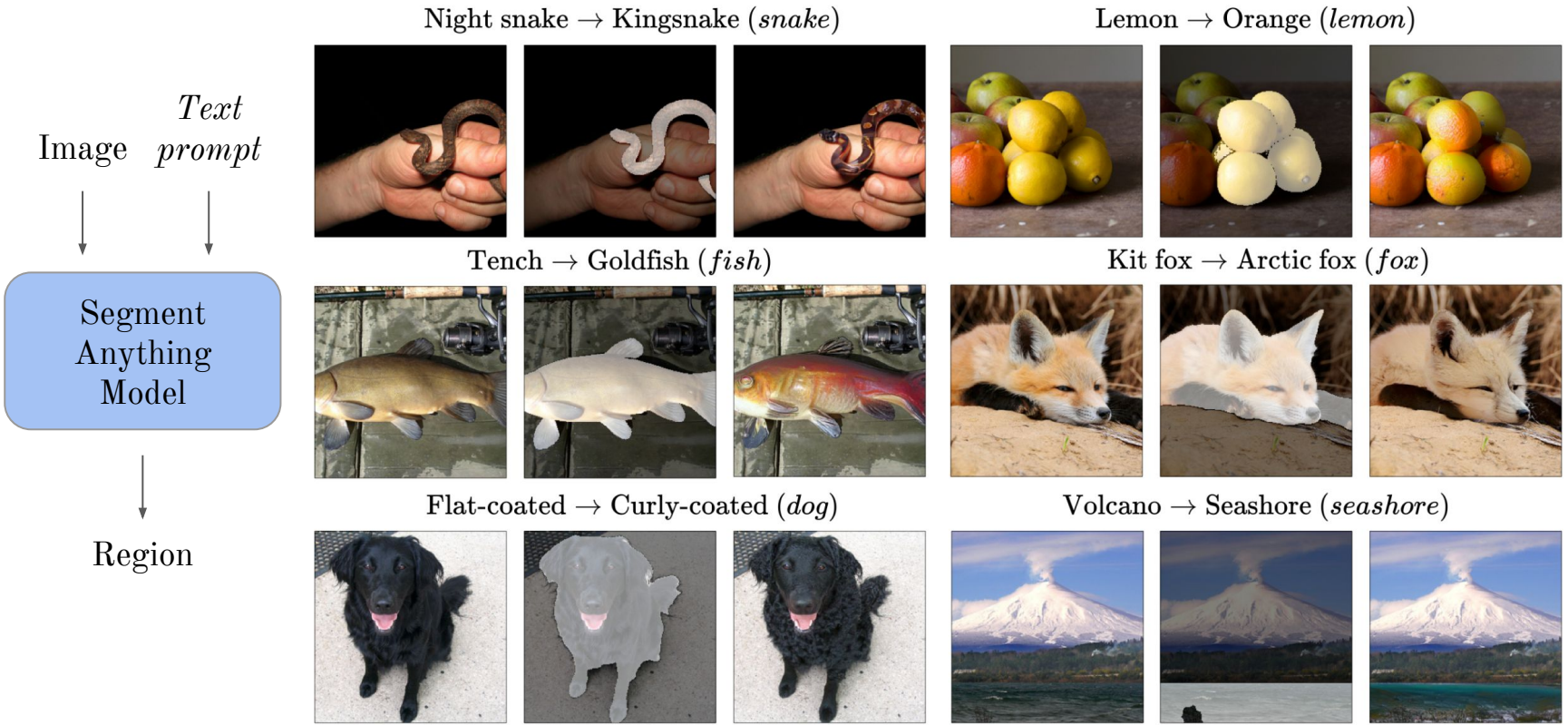
Can explore extremely low attribution values

q	Zebra – Sorrel					Cheetah – Cougar					Egyptian Cat – Persian Cat				
Metric	FID	sFID	S^3	COUT	FR	FID	sFID	S^3	COUT	FR	FID	sFID	S^3	COUT	FR
0.9	5.2	13.8	0.87	0.62	91.6	6.1	17.1	0.92	0.88	99.8	13.5	31.4	0.86	0.89	99.9
0.8	4.6	13.3	0.87	0.53	88.0	4.4	15.7	0.91	0.80	99.1	11.0	28.6	0.85	0.84	99.3
0.7	3.9	12.7	0.88	0.48	85.7	3.5	15.1	0.91	0.74	97.4	9.2	26.9	0.85	0.79	98.5
0.6	3.6	12.4	0.89	0.45	82.4	2.9	14.6	0.92	0.68	94.3	7.8	25.5	0.85	0.73	96.1
0.5	3.6	12.4	0.89	0.40	78.5	2.7	14.5	0.92	0.65	93.1	6.9	24.7	0.86	0.69	94.6
0.4	3.7	12.3	0.89	0.38	77.7	2.8	14.6	0.92	0.65	93.4	6.6	24.5	0.87	0.65	93.1
0.3	3.9	12.6	0.89	0.40	79.2	3.1	14.9	0.92	0.69	95.7	7.5	25.3	0.86	0.67	94.2

Outperforms in terms of efficiency

Inpainting method	NFE					
	U-Net		Classifier		Other	
	forward	backward	forward	backward	forward	backward
RCSB	100	100	100	100	0	0
LDCE	191	191	191	191	0	0
DVCE	200	200	1600	1600	1600	1600
ACE	520	500	25	25	0	0
DDRM	200	200	200	200	0	0
MCG	1000	1000	1000	1000	0	0
RePaint	2410	2410	2410	2410	0	0

Allows for *exact* counterfactual reasoning

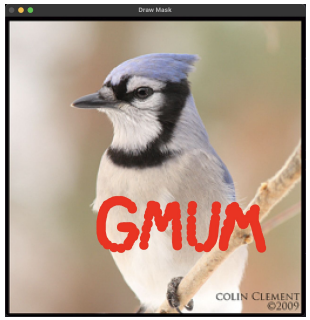


while preserving performance

Metric	FID	sFID	S^3	COUT	FR	FID	sFID	S^3	COUT	FR	FID	sFID	S^3	COUT	FR
Task	**Zebra – Sorrel**					**Cheetah – Cougar**					**Egyptian Cat – Persian Cat**				
A	Exact regions obtained with LangSAM and prompts: zebra / horse, cheetah / cougar, cat respectively														
Values	32.8	41.5	0.87	0.74	98.9	37.2	50.6	0.91	0.84	99.4	52.0	82.8	0.81	0.84	99.2

Can be adapted to human-explanation interaction

User input



Region

while also preserving performance

Metric	FID	sFID	S^3	COUT	FR	FID	sFID	S^3	COUT	FR	FID	sFID	S^3	COUT	FR
Task	**Zebra – Sorrel**					**Cheetah – Cougar**					**Egyptian Cat – Persian Cat**				
B	Regions based on freeform masks with the area in the indicated range														
$10 - 20\%$	6.7	15.0	0.85	0.85	87.6	9.0	19.1	0.89	0.72	96.6	12.4	29.6	0.80	0.73	96.9
$20 - 30\%$	7.8	15.8	0.84	0.53	92.2	11.6	21.3	0.88	0.71	99.6	17.7	34.0	0.78	0.74	99.3

despite the masks being random

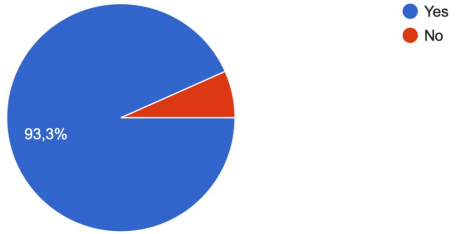
Sorrel $\rightarrow$ Zebra

Image	Region	Explanation	Image	Region	Explanation
					
					

and is preferred by the users

Do you think that the possible interaction with explanations from type B can be useful in obtaining a better understanding of the model?

15 odpowiedzi



Type A

Image | Explanation | Difference

Prediction: jay | Prediction: bulbul

Difference

Prediction: guacamole | Prediction: cabbage

Type B

Image | Explanation | Difference

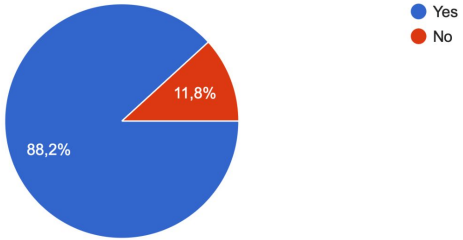
Prediction: jay | Prediction: bulbul

Difference

Prediction: guacamole | Prediction: cabbage

Do you find the explanations of type B more useful than type A?

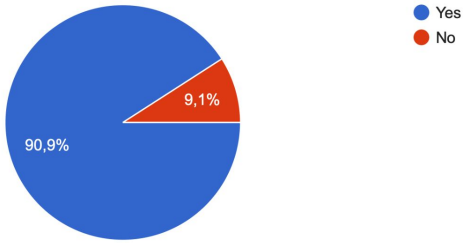
17 odpowiedzi



in various scenarios ;)

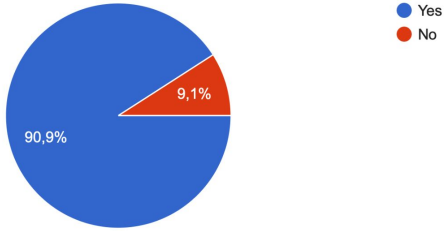
Do you judge RVCEs as helpful in understanding why the model predicted the wrong class in the beginning?

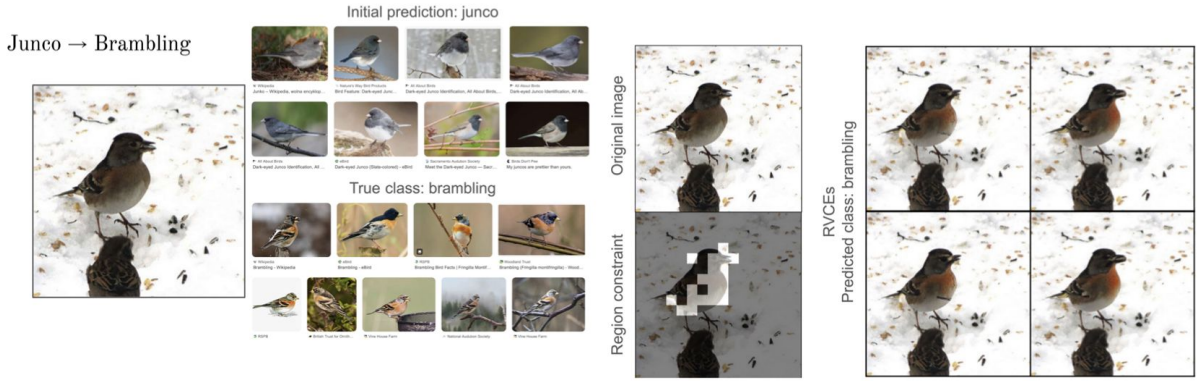
11 odpowiedzi



Does it seem like RVCEs indicate semantic features that were missing in the beginning for the model to provide a correct prediction?

11 odpowiedzi





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. Dhariwal and Nichol, *Diffusion Models Beat GANs on Image Synthesis*, NeurIPS 2021
4. Sobieski and Biecek, *Global Counterfactual Directions*, ECCV 2024,
5. Sobieski et al., *Rethinking Visual Counterfactual Explanations Through Region Constraint*, arXiv 2024.