ICCV OCT 19-23, 2025 HONOLULU HAWAII

Soft Local Completeness

Rethinking Completeness in XAI

Ziv Weiss Haddad*, Oren Barkan*

The Open University

Yehonatan Elisha, Noam Koenigstein

Tel Aviv University

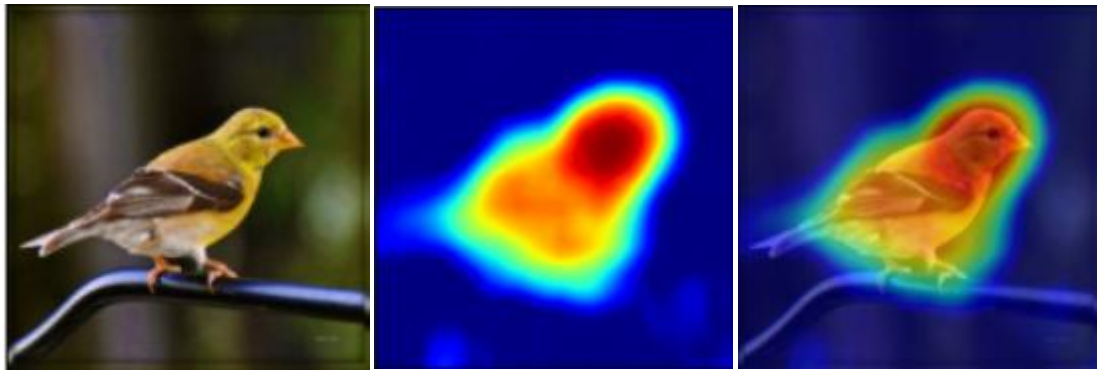
האוניברסיטה הפתוחה
THE OPEN UNIVERSITY OF ISRAEL
الجامعة المفتوحة

TEL AVIV UNIVERSITY

Explainable AI (XAI) via feature attribution

- This work focuses on XAI for image classification models $f : \mathbb{R}^n \rightarrow [0,1]^C$
- Task: Given an image $\mathbf{x} \in \mathbb{R}^n$ and a model prediction for class y, produce an *explanation* for the model prediction
- Explanation via attribution map $\mathbf{a}_{\mathbf{x}}^y \in \mathbb{R}^n$: highlight the input features that are responsible for the model prediction $y$or class

Input image | Attribution map | Attribution overlay



Attribution for class '*goldfinch*'

Input image | '*dog*' attribution | '*person*' attribution

Attributions for classes '*dog*' *and* '*person*'

Completeness

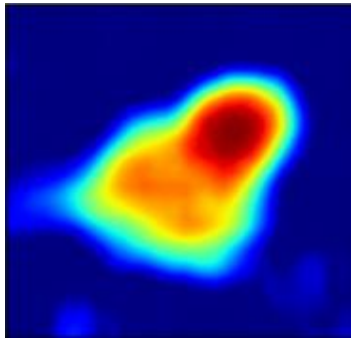
- Model response to input $\mathbf{x}$:
  - $r(\mathbf{x}) := f(\mathbf{x}) - f(\mathbf{b})$
  - $\mathbf{b} = \mathbf{0}_n$
- Completeness:
  - Satisfied when: $\sum_{i=1}^{n} \mathbf{a}_{\mathbf{x}}^{y}[i] = r(\mathbf{x})[y]$
  - "The explanation *completely* accounts for the response"
  - Can be superficially imposed by a simple normalization
- In this work, we rethink completeness as a **local and flexible guiding measure** rather than a **strict global requirement**

Soft Local Completeness

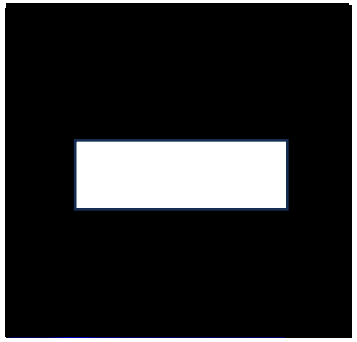
- **S**oft **LO**cal **C**omlpeteness (***SLOC***) aims to produce faithful explanations, by promoting completeness ***locally***, in a ***soft*** manner, across a large set of image subregions
- Definitions:
  - Mask: $\mathbf{m} \in \{0,1\}^n$
  - Masked Image: $\mathbf{x}^{\mathbf{m}} = \mathbf{x} \circ \mathbf{m} + (1 - \mathbf{m}) \circ \mathbf{b}$
  - Sub-map: $\mathbf{a}^y_{\mathbf{x}} \circ \mathbf{m}$
  - **Local Completeness**: $\mathbf{a}^y_{\mathbf{x}} \cdot \mathbf{m} = r(\mathbf{x}^{\mathbf{m}})[y]$
- Given a large set of masks: $\mathcal{M} \subset \{0,1\}^n$
  - Ideally, the attribution map would satisfy local completeness across all masks
  - However, it is likely that no such attribution map exists

Input image

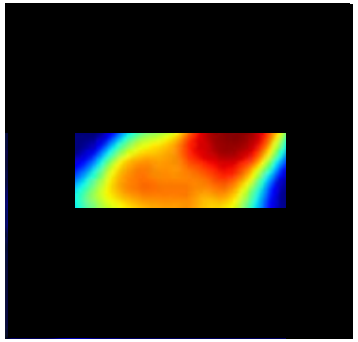
Attribution map



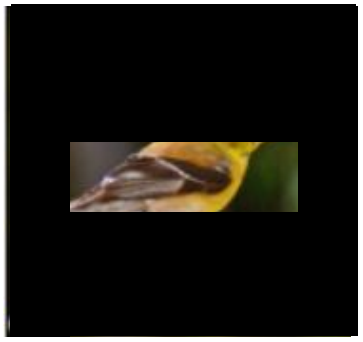
Mask



Sub-map



Masked image



Soft Local Completeness

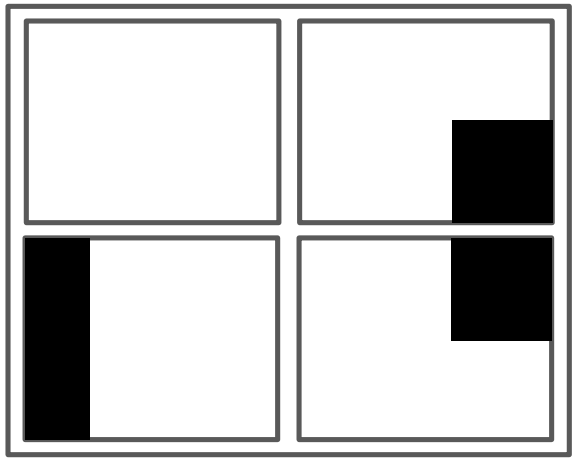
- *Completeness-Gap*: $r(\mathbf{x}^{\mathbf{m}})[y] - \mathbf{a}_{\mathbf{x}}^{y} \cdot \mathbf{m}$
- The completeness gap essentially measures to which extent the sub-map violates local completeness
- Our approach seeks to minimize the completeness gap for multiple sub-maps, simultaneously
- *SLOC* is a **model-agnostic black-box** method, facilitating optimization procedure that promotes completeness locally within subregions of the attribution map (sub-maps)

Method Motivation (Toy Example)

(a) Original image: 0.8

(c) Masked Image: 0.8

(e) Four masks



(b) Masked Image: 0.8

(d) Masked Image: 0.8

(f) Attribution Map

Enforcing local completeness across **all** four sub-maps implies **the shared central stripe**, visible in (a)-(d), **must alone account for the full 0.8 response**. This yields the attribution map shown in (f), highlighting the owl in red.

The attribution map satisfies local completeness for (a) and (c), where the model response is 0.8. Thus, the total attribution is 0.8, with zero contribution from the region masked in (c). The same holds for (b) and (d), implying that the unmasked central stripe must account for the full 0.8.

SLOC Optimization

Given a set of masks $\mathcal{M} \subset \{0,1\}^n$, we define:

Completeness-gap loss:

$$\mathcal{L}_c(\mathbf{a}_\mathbf{x}^y; \mathcal{M}) = \frac{1}{2|\mathcal{M}|} \sum_{\mathbf{m} \in \mathcal{M}} \frac{1}{|\mathbf{m}|} (\underbrace{r(\mathbf{x}^\mathbf{m})[y] - \mathbf{a}_\mathbf{x}^y \cdot \mathbf{m}}_{\text{completeness gap}})^2$$

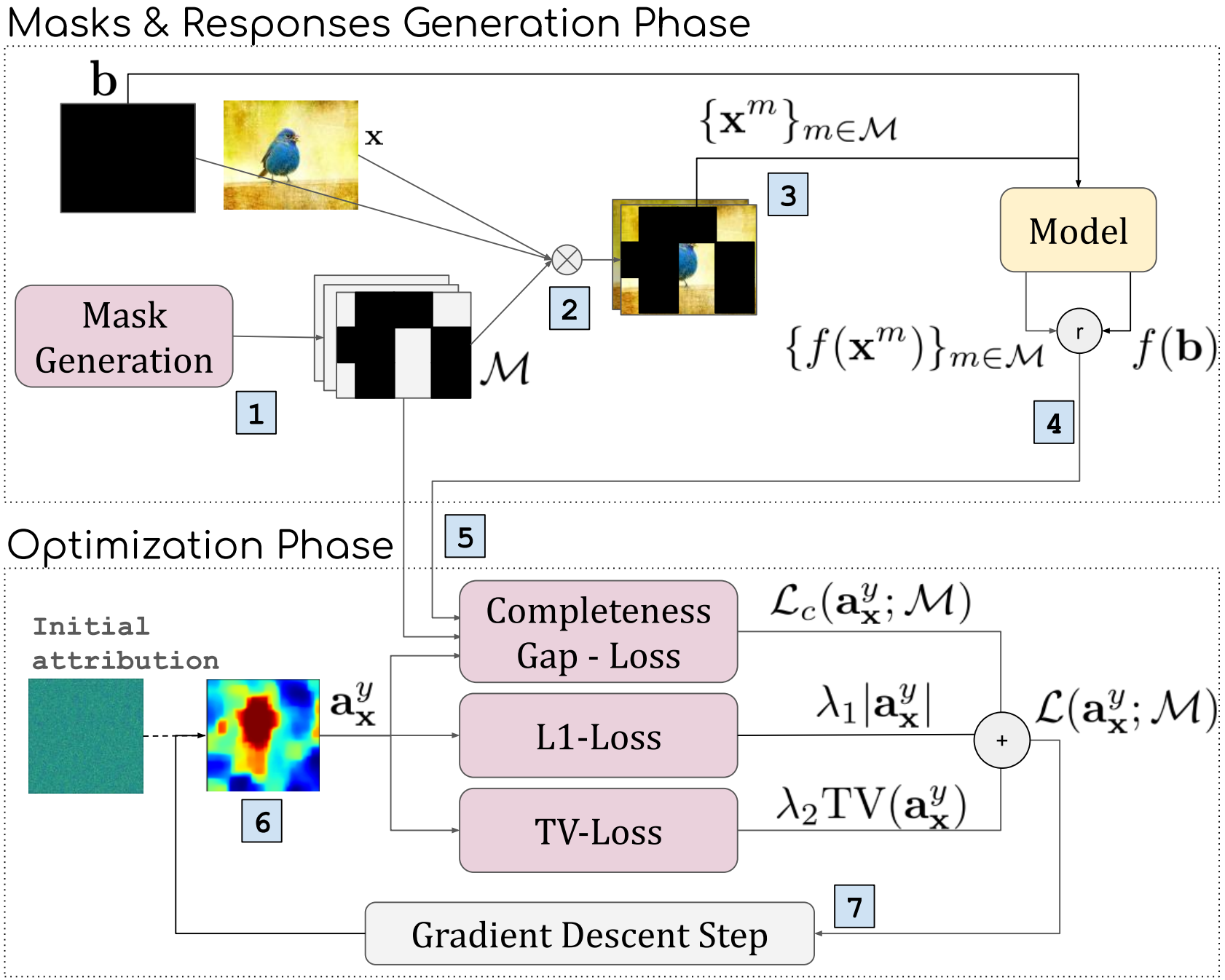
SLOC loss:

$$\mathcal{L}(\mathbf{a}_\mathbf{x}^y; \mathcal{M}) = \mathcal{L}_c(\mathbf{a}_\mathbf{x}^y; \mathcal{M}) + \lambda_1 |\mathbf{a}_\mathbf{x}^y| + \lambda_2 \mathrm{TV}(\mathbf{a}_\mathbf{x}^y)$$

where

$$\mathrm{TV}(\mathbf{a}_\mathbf{x}^y) := \sum_{i,j} (\mathbf{a}_\mathbf{x}^y[i,j] - \mathbf{a}_\mathbf{x}^y[i+1,j])^2 + (\mathbf{a}_\mathbf{x}^y[i,j] - \mathbf{a}_\mathbf{x}^y[i,j+1])^2$$

Optimized using gradient descent on $\mathcal{L}$ with respect to $\mathbf{a}_\mathbf{x}^y$



SLOC Method Diagram

Masks & Responses Generation Phase

(1) A set of M random masks is generated.
(2) Perturbed inputs are created by combining the original image with a baseline using these masks.
(3) These perturbed inputs are fed into the model to obtain corresponding outputs.
(4) The model response for each mask is computed by subtracting the output for the baseline from the output of the perturbed input.

Optimization Phase

(5)-(6) The resulting mask-response pairs are passed together with an initial random attribution to compute the completeness-gap.
(7) Gradient descent is used to iteratively update the attribution using the SLOC loss function, which includes the *completeness gap* loss, as well as TV and L1 regularization terms.

Note: The model *f* is used only **once**, during *Mask & Responses Generation Phase,* to compute responses via a **single** forward pass, hence SLOC is a **model-agnostic black-box method.**

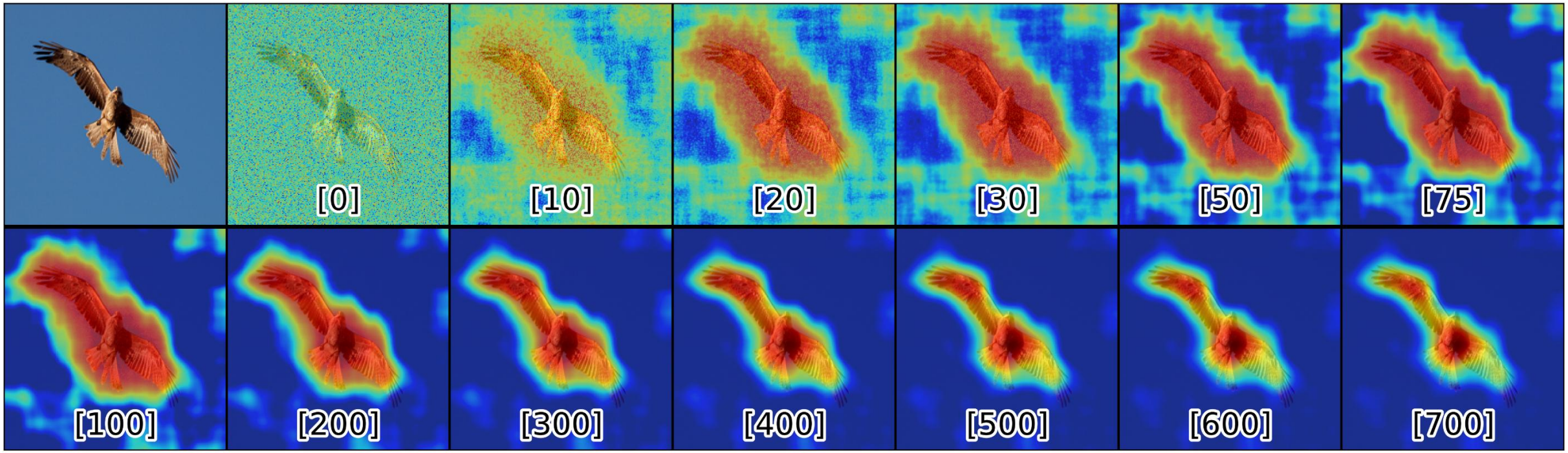
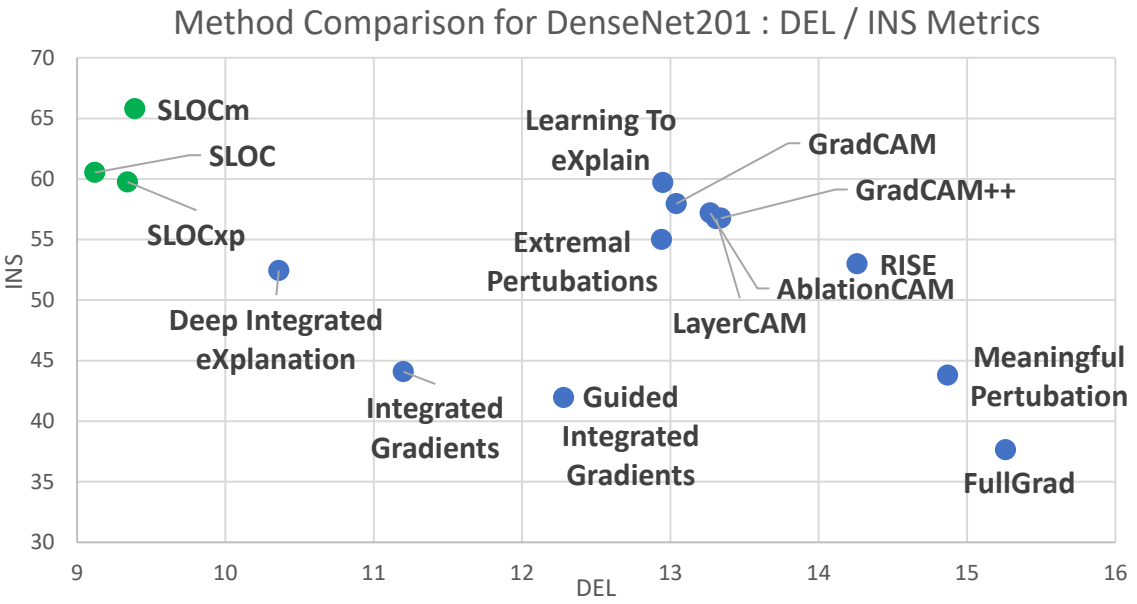


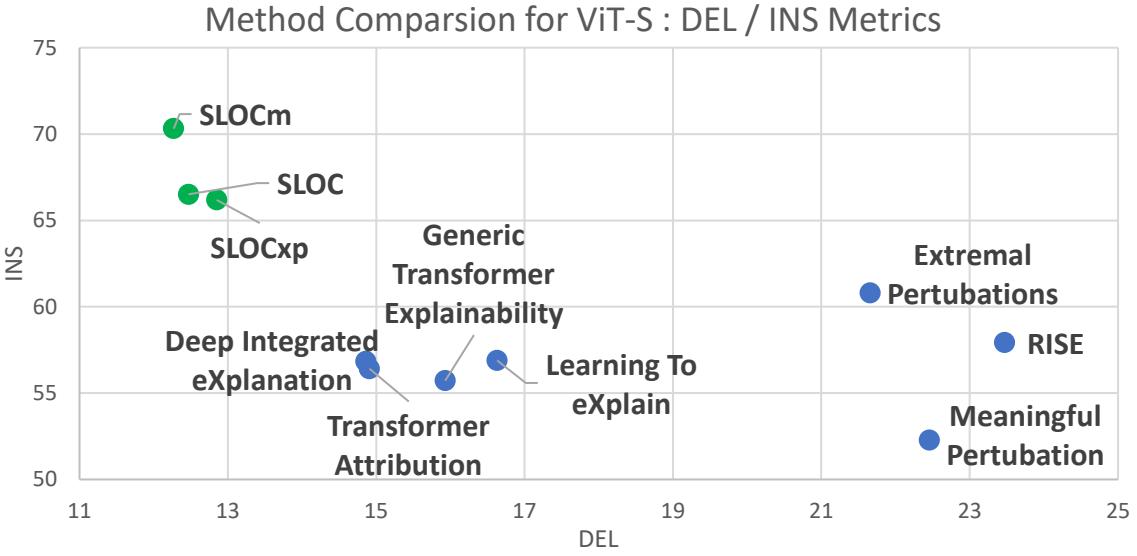
Figure 3. SLOC attribution maps across training steps. Faithful explanations emerge after a few hundred gradient updates, with an appropriate learning rate decay.

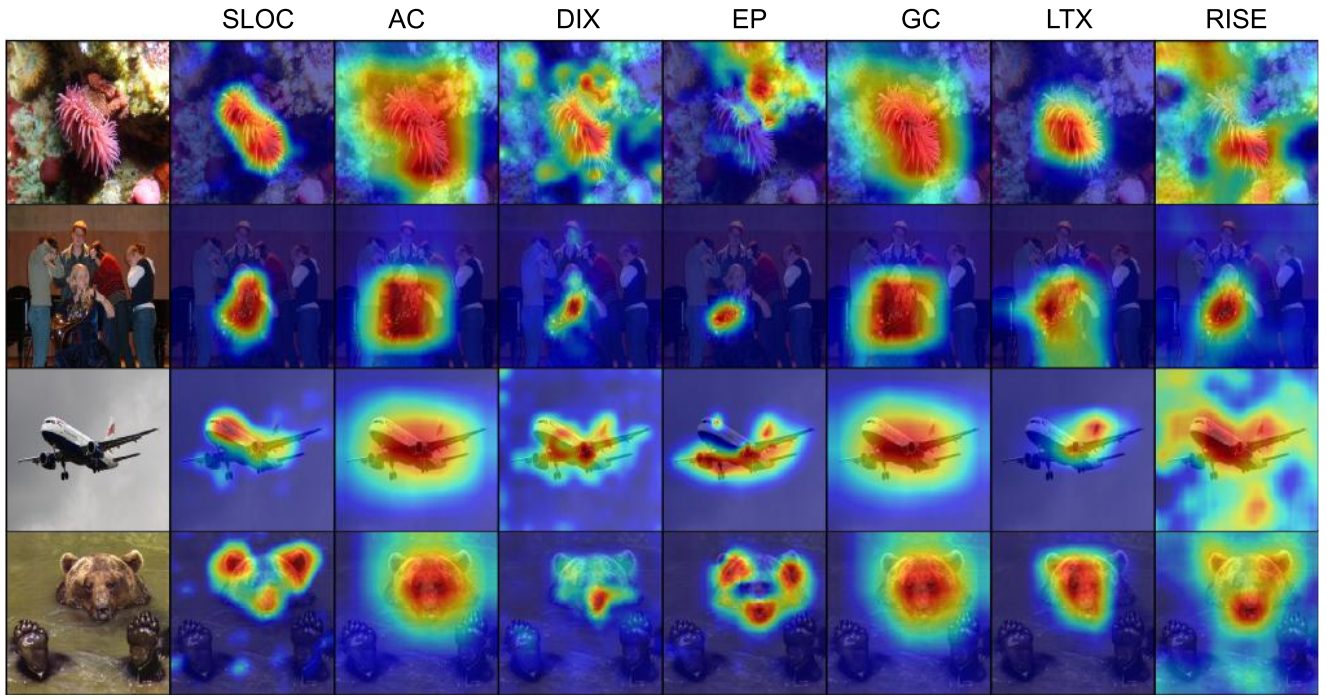
Experimental Setup

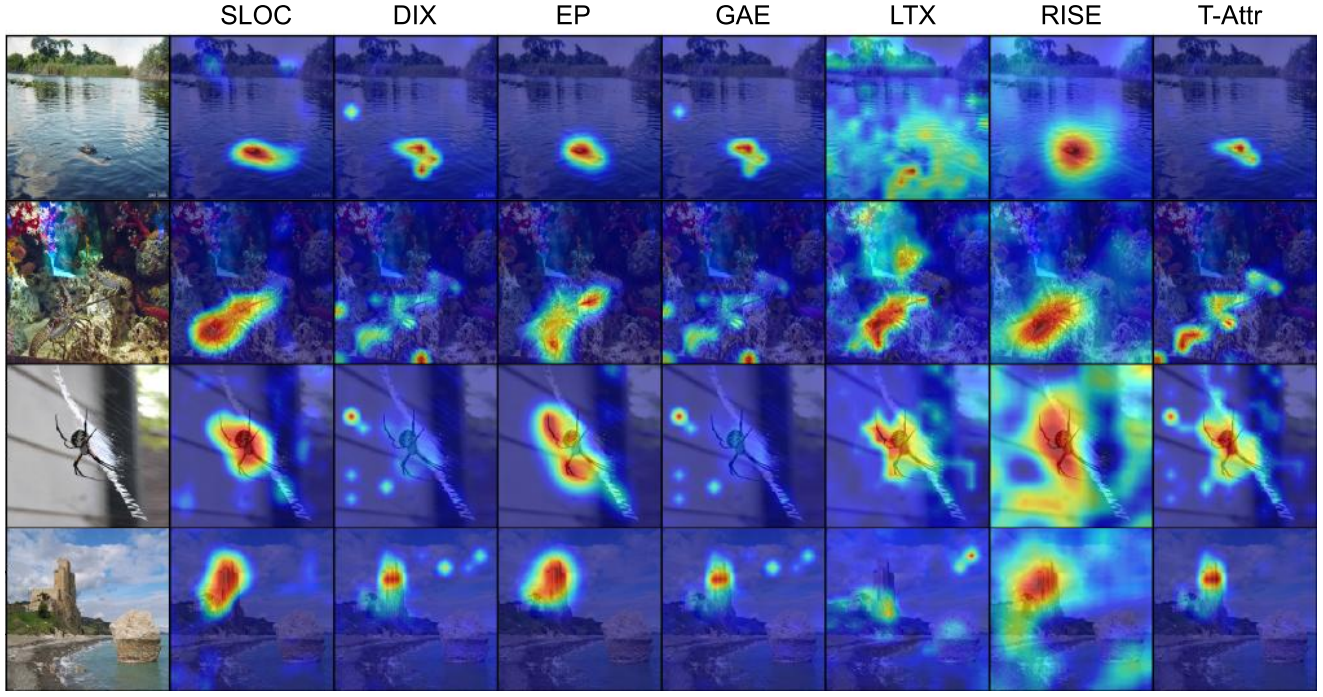
- Models: ResNet50, DenseNet201, ViT-Small, ViT-Base
- Datasets: ImageNet, ImageNet-Segmentation, Pascal-VOC, FunnyBirds
- Explanation evaluation protocols:
  - Faithfulness tests using the POS, NEG, DEL, INS, NPD, IDD, AIC, SIC metrics
  - ‘FunnyBirds’ evaluation using the Completeness, Correctness, Contrastivity metrics
  - Segmentation tests using the mIoU, mAP, Pixel Accuracy metrics

State-of-the-art Results









Thank you!

- Looking forward to seeing you at the poster session and continuing the discussion

Soft Local Completeness

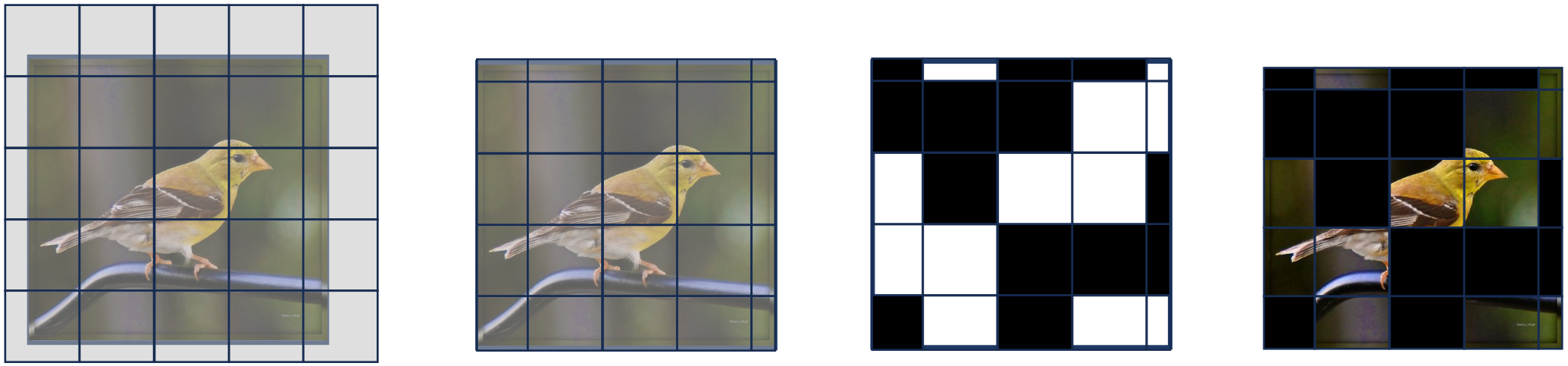
Rethinking Completeness in XAI

[GitHub]

Additional Slides

Implementation Details - Mask Generation

- We define a sufficiently large grid of patches of size LxL
- Offsets are drawn are drawn independently from {0..L-1}
- All the pixels of a patch are set to be either 0 or 1 with probability p



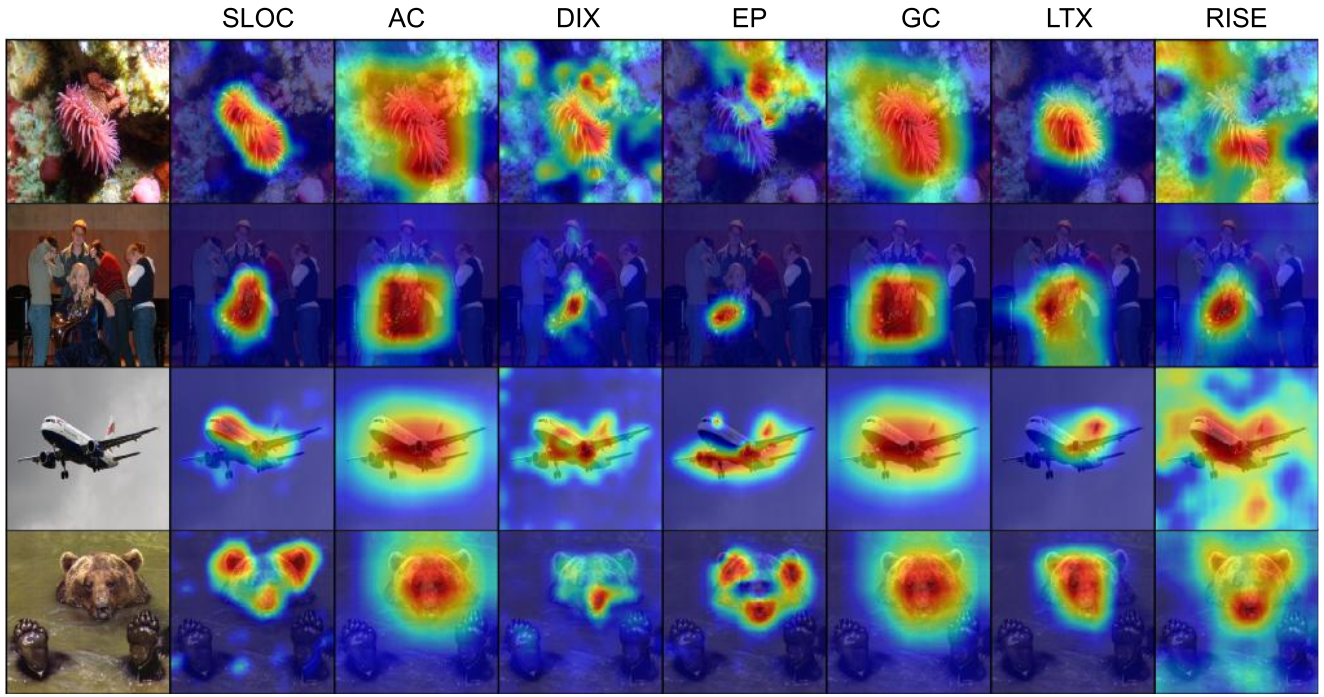
Results (IN)

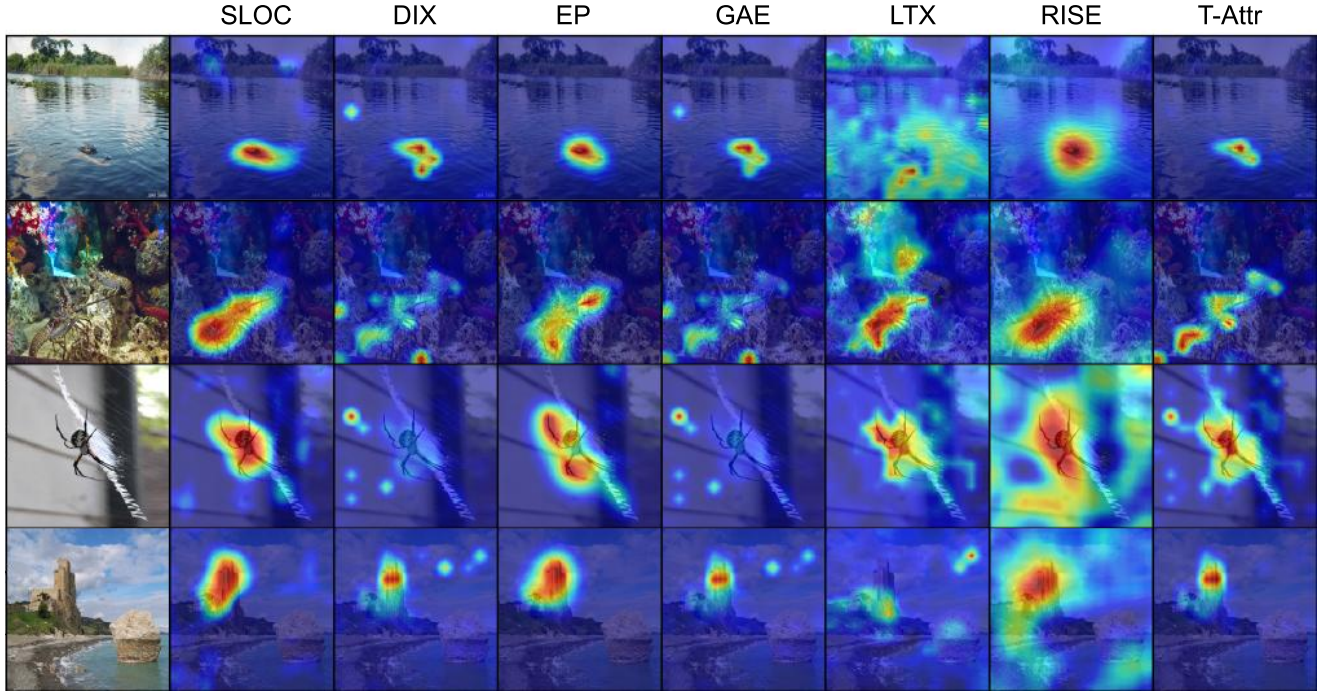
Method	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
$SLOC_m$	11.61	**76.54**	9.39	**65.79**	**64.93**	**56.4**	**79.21**	**78.35**
SLOC	**11.52**	70.97	**9.12**	60.53	59.45	51.41	77.79	77.72
$SLOC_{xp}$	11.78	70.20	9.34	59.75	58.43	50.41	77.33	76.80
AC	17.24	67.68	13.27	57.18	50.44	43.91	77.78	75.41
DIX	13.36	62.95	10.36	52.43	49.59	42.07	74.62	71.47
EP	16.12	65.68	12.94	55.0	49.55	42.06	77.38	74.97
FG	19.06	44.66	15.26	37.62	25.6	22.37	58.23	53.86
GC	16.88	68.54	13.04	57.95	51.66	44.91	78.38	76.01
GC++	17.36	67.29	13.34	56.75	49.93	43.41	78.04	75.62
GIG	14.81	49.96	12.28	41.93	35.16	29.65	61.12	57.6
IG	14.14	51.99	11.2	44.08	37.85	32.88	61.26	58.48
LC	17.28	67.27	13.31	56.71	49.99	43.4	77.95	75.42
LTX	16.24	71.09	12.95	59.69	54.84	46.74	78.92	76.25
MP	18.54	53.24	14.87	43.79	34.7	28.92	66.98	64.13
RISE	18.42	62.75	14.26	52.99	44.33	38.73	76.82	74.24

Table 1. Faithfulness results for all combinations of method and metric, using the DN model on the IN dataset.

Method	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
$SLOC_m$	**14.83**	**81.81**	**12.27**	**70.31**	**66.98**	**58.04**	**83.78**	**83.34**
SLOC	15.25	77.97	12.47	66.51	62.72	54.04	83.19	82.88
$SLOC_{xp}$	15.79	77.85	12.85	66.18	62.06	53.33	83.06	81.97
DIX	18.69	68.17	14.86	56.83	49.48	41.97	76.88	75.05
EP	27.37	72.52	21.66	60.8	45.14	39.14	79.45	77.56
GAE	19.98	66.93	15.93	55.72	46.95	39.79	75.42	73.9
LTX	20.84	68.5	16.63	56.89	47.66	40.27	74.22	71.56
MP	27.72	63.25	22.46	52.28	35.53	29.81	74.22	71.25
RISE	29.51	69.52	23.47	57.93	40.02	34.46	79.93	77.23
TATTR	19.06	67.52	14.91	56.41	48.46	41.5	78.03	75.69

Table 2. Faithfulness results for all combinations of method and metric, using the ViT-S model on the IN dataset.

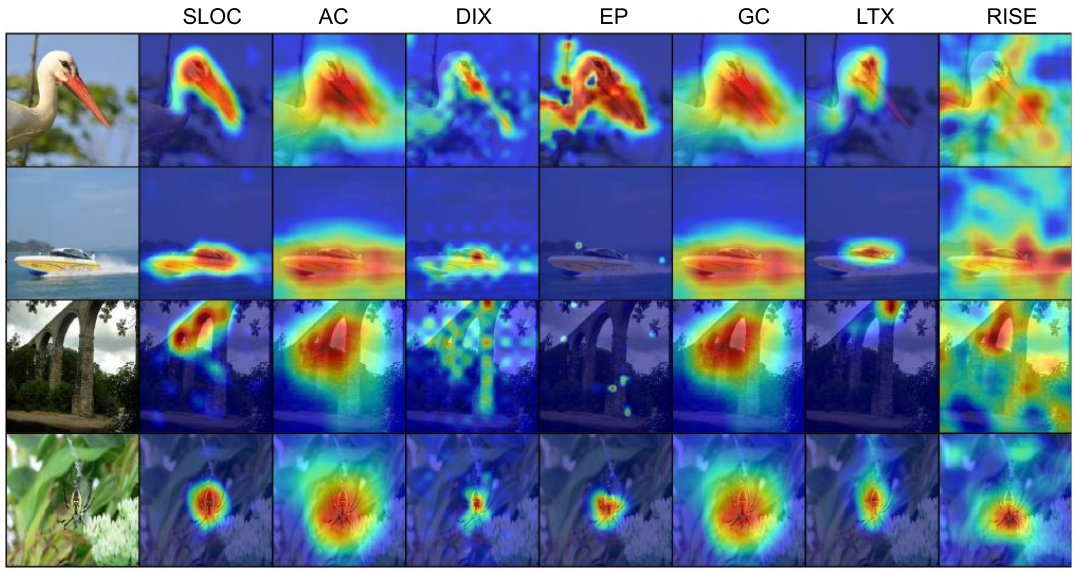




Additional Results (IN)

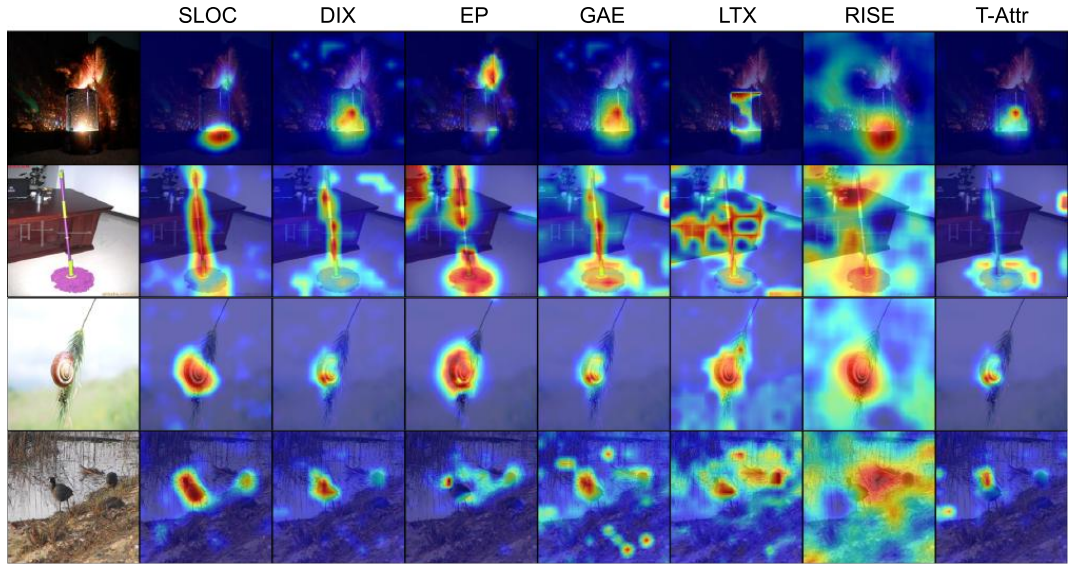
Method	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
$SLOC_m$	10.96	**75.84**	8.68	**64.19**	**64.87**	**55.5**	**79.44**	**78.13**
SLOC	10.65	69.56	8.32	58.33	58.91	50.01	78.5	77.04
$SLOC_{xp}$	10.74	68.62	8.41	57.41	57.88	49.01	77.75	76.78
AC	16.7	66.96	12.76	55.71	50.26	42.95	77.17	74.59
DIX	10.21	58.33	7.83	48.16	48.11	40.33	71.15	68.81
EP	14.9	66.41	11.5	54.51	51.51	43.01	75.06	73.96
FG	16.79	65.9	12.94	54.9	49.11	41.96	74.16	71.54
GC	16.37	68.04	12.56	56.65	51.67	44.1	77.33	75.1
GC++	16.81	66.85	12.85	55.54	50.04	42.68	76.82	74.54
GIG	**9.4**	45.28	**7.68**	37.71	35.89	30.03	57.52	54.51
IG	9.9	44.22	7.76	37.14	34.32	29.38	56.56	54.23
LC	17.04	66.58	13.0	55.25	49.54	42.25	76.46	74.33
LTX	14.98	69.88	11.7	57.74	54.91	46.03	76.69	74.29
MP	17.16	50.81	13.6	41.34	33.65	27.74	64.71	62.52
RISE	15.8	62.3	12.04	51.93	46.5	39.89	77.31	74.77

Table 7. Faithfulness results for all combinations of method and metric, using the RN model on the IN dataset.



Method	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
$SLOC_m$	**21.54**	**89.01**	**14.93**	**63.05**	**67.48**	**48.12**	**85.42**	81.82
SLOC	22.39	85.72	15.7	59.85	62.73	43.71	84.82	**81.86**
$SLOC_{xp}$	22.36	85.10	15.56	59.27	62.98	43.88	84.74	81.58
DIX	32.09	77.01	21.14	51.17	44.92	30.03	79.88	74.92
EP	41.24	82.95	25.18	58.99	41.72	33.81	81.59	77.32
GAE	33.16	76.88	21.95	51.12	43.72	29.17	79.02	74.6
LTX	28.3	80.74	18.8	55.52	52.44	36.72	79.65	74.98
MP	36.63	78.49	23.83	52.82	41.87	28.99	80.58	76.11
RISE	49.7	77.42	32.7	50.09	27.72	17.4	76.65	72.16
TATTR	32.8	77.24	21.49	51.56	44.44	30.07	80.04	74.79

Table 8. Faithfulness results for all combinations of method and metric, using the ViT-B model on the IN dataset.



Additional Results (VOC)

Method	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
$SLOC_m$	**7.31**	**69.2**	**5.07**	**48.83**	**61.89**	**43.76**	**75.63**	**79.09**
SLOC	7.32	63.56	5.09	44.58	56.25	39.5	74.96	77.84
$SLOC_{xp}$	7.61	63.23	5.22	43.72	55.65	38.48	73.69	77.59
DIX	10.19	48.66	6.74	31.73	38.47	24.99	64.51	68.66
EP	14.06	55.96	9.26	36.55	41.89	27.29	68.43	72.04
GAE	11.2	47.4	7.47	30.92	36.20	23.44	64.22	67.95
LTX	12.56	49.28	8.2	32.26	36.71	24.06	56.55	63.16
RISE	15.67	53.21	10.26	35.35	37.53	25.09	66.91	70.5
T-Attr	10.37	47.56	6.81	31.35	37.19	24.54	65.62	69.53

Table 3. Faithfulness results for combinations of method and metric, using the ViT-S model on the VOC dataset.

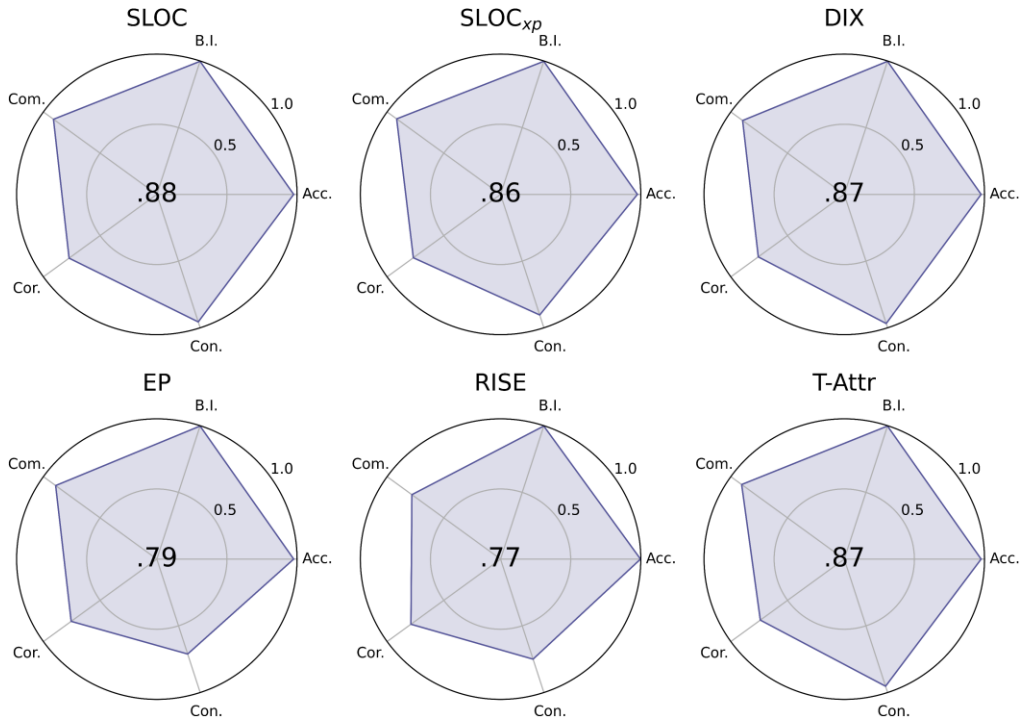
Method	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
$SLOC_m$	6.24	**61.32**	4.19	**45.15**	**55.09**	**40.96**	**68.6**	**71.89**
SLOC	**6.19**	53.4	**4.05**	39.01	47.2	34.97	65.63	70.92
$SLOC_{xp}$	6.40	52.61	4.08	38.11	45.74	34.04	66.65	70.58
AC	10.06	50.15	6.4	33.85	40.09	27.45	61.46	65.05
DIX	8.2	43.89	5.19	29.31	35.7	24.12	58.67	61.64
EP	9.08	49.12	5.91	33.1	40.04	27.19	63.7	65.53
FG	9.42	27.99	6.36	19.43	18.57	13.07	36.75	39.37
GC	9.85	51.83	6.24	34.89	41.99	28.64	63.16	65.68
GC++	10.11	49.43	6.5	33.03	39.31	26.54	61.42	64.15
GIG	6.47	30.37	4.3	21.46	23.9	17.16	41.59	44.56
IG	7.73	30.88	5.09	21.56	23.15	16.47	41.59	42.76
LC	10.14	49.29	6.52	32.94	39.15	26.42	61.25	64.39
LTX	8.98	54.84	5.92	36.93	45.86	31.01	61.37	65.11
RISE	9.38	43.75	6.02	30.78	34.38	24.76	61.99	64.67

Table 9. Faithfulness results for all combinations of method and metric, using the DN model on the VOC dataset.

FunnyBirds (ViT-B)

Method	Completeness↑	Correctness↑	Contrastivity↑	Overall↑
SLOC	0.91	0.77	0.96	**0.88**
$SLOC_{xp}$	**0.92**	0.77	0.90	0.86
DIX	0.9	0.76	**0.97**	0.87
EP	0.89	0.76	0.71	0.79
RISE	0.78	**0.79**	0.75	0.77
T-Attr	0.9	0.74	0.95	0.87

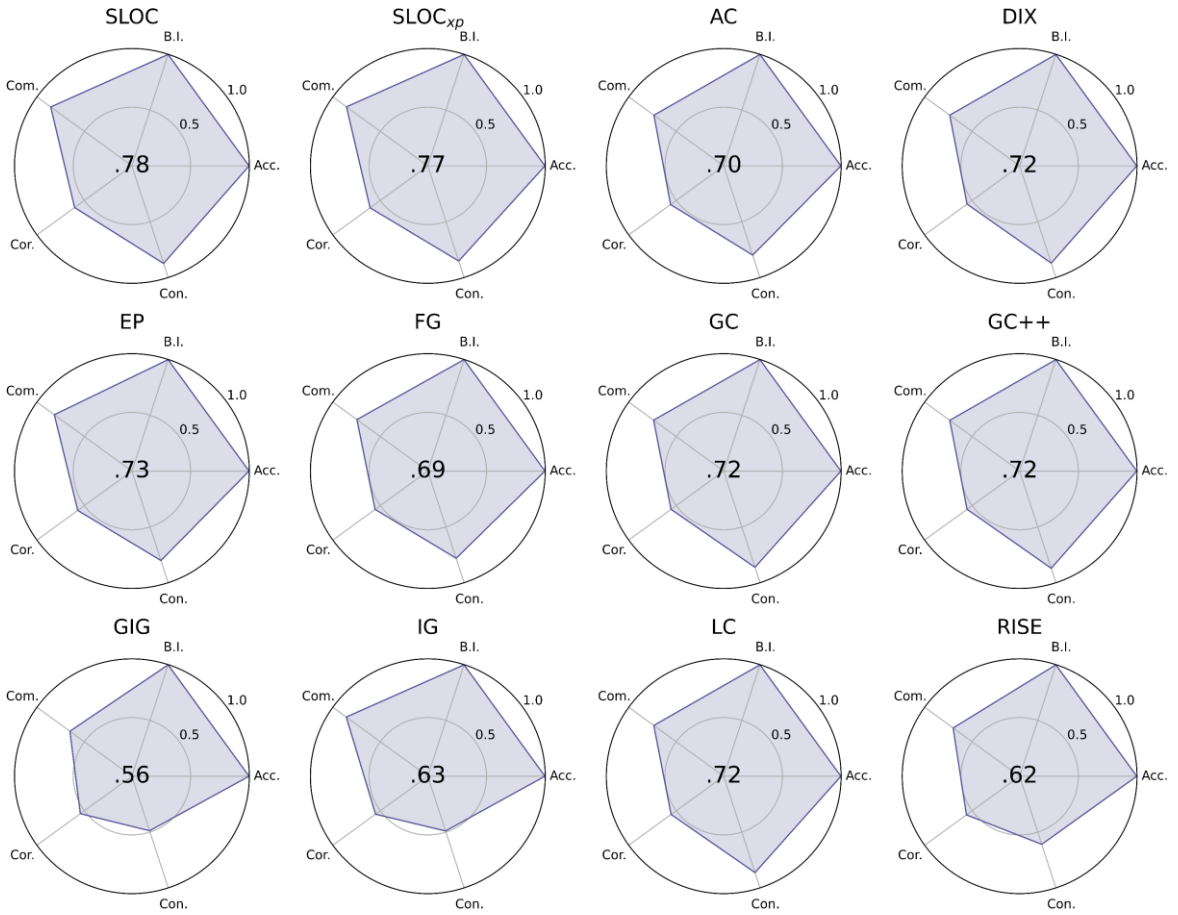
Table 11. FunnyBirds evaluation results for the ViT-B model.



FunnyBirds (RN)

Method	Completeness↑	Correctness↑	Contrastivity↑	Overall↑
SLOC	0.85	0.60	**0.87**	**0.78**
$SLOC_{xp}$	0.86	**0.61**	0.85	0.77
AC	0.73	0.56	0.80	0.70
DIX	0.74	0.55	**0.87**	0.72
EP	0.82	0.57	0.8	0.73
FG	0.75	0.56	0.78	0.69
GC	0.74	0.55	0.86	0.72
GC++	0.74	0.55	0.87	0.72
GIG	0.65	0.54	0.49	0.56
IG	**0.86**	0.55	0.49	0.63
LC	0.74	0.55	0.86	0.72
RISE	0.70	0.56	0.61	0.62

Table 10. FunnyBirds evaluation results for the RN model.



Segmentation results

Method	SLOC	AC	DIX	EP	GC	GC++	GIG	IG	LC	LTX	RISE
mIoU↑	0.56	0.55	**0.66**	0.52	0.55	0.56	0.51	0.48	0.55	0.56	0.51
mAP↑	0.80	**0.86**	0.84	0.76	0.85	0.85	0.78	0.76	0.85	0.83	0.79
PA↑	0.77	0.72	**0.82**	0.72	0.73	0.73	0.74	0.79	0.73	0.51	0.7

Table 6. Segmentation tests results for the RN model.

Method	SLOC	DIX	EP	GAE	LTX	MP	RISE	T-Attr
mIoU↑	0.52	0.63	0.50	0.61	0.56	0.55	0.50	0.68
mAP↑	0.76	0.81	0.76	0.79	0.81	0.74	0.75	0.83
PA↑	0.72	0.79	0.71	0.78	0.72	0.74	0.68	0.82

Table 13. Segmentation results for the ViT-S model.

Ablation Study

L	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
8	16.42	74.02	13.54	63.52	57.6	49.98	79.42	77.31
16	15.18	75.82	12.44	65.15	60.64	52.71	81.66	80.33
32	14.86	76.55	12.31	65.45	61.69	53.14	83.04	80.88
40	15.67	76.67	12.84	65.52	61.0	52.68	83.08	81.01
48	16.43	77.39	13.35	65.87	60.95	52.53	83.23	81.25
56	17.01	77.49	13.86	65.91	60.48	52.05	82.89	80.98
64	17.69	77.68	14.49	65.75	59.99	51.26	82.48	81.1
SLOC	15.7	77.38	12.83	66.1	61.68	53.26	83.19	81.36

Table 19. Faithfulness performance across different patch size settings.

$\|\mathcal{M}\|$	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
10	32.91	60.32	26.63	49.16	27.41	22.53	69.26	66.88
100	19.76	70.18	16.15	58.29	50.42	42.14	77.97	75.48
250	17.17	73.96	14.09	62.43	56.79	48.35	80.43	78.47
500	16.36	76.16	13.25	64.51	59.8	51.26	82.44	80.65
750	15.95	77.24	12.93	65.62	61.3	52.7	82.91	81.23
1000	15.7	77.38	12.83	66.1	61.68	53.26	83.19	81.36
1250	15.65	77.55	12.79	66.33	61.9	53.54	83.53	81.64
1500	15.54	78.06	12.77	66.7	62.52	53.93	83.25	81.66
2000	15.51	78.25	12.68	66.95	62.74	54.27	83.66	82.03

Table 20. Faithfulness performance for varying numbers of drawn masks.

T	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
25	20.12	75.85	16.41	63.68	55.73	47.27	80.1	77.64
50	17.03	79.86	13.75	68.47	62.83	54.72	83.26	81.77
75	17.19	80.26	13.99	68.9	63.07	54.9	83.9	82.7
100	17.24	80.2	14.12	68.8	62.95	54.68	84.21	82.54
200	16.52	79.1	13.54	67.86	62.57	54.32	83.41	82.44
250	16.19	78.76	13.28	67.59	62.57	54.31	83.71	82.26
500	15.34	77.86	12.59	66.8	62.52	54.21	83.43	82.11
750	14.88	77.46	12.26	66.36	62.58	54.11	83.82	81.97
1000	14.69	77.37	12.09	66.19	62.67	54.1	83.01	82.1

Table 22. Faithfulness performance for varying numbers of gradient update steps (iterations).

p	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
0.1	18.35	76.28	14.67	64.86	57.93	50.2	81.53	79.85
0.2	16.54	77.76	13.35	66.33	61.21	52.98	82.75	81.17
0.3	15.73	78.21	12.80	66.82	62.49	54.02	83.36	81.56
0.4	16.1	77.75	13.33	66.52	61.65	53.19	82.5	80.96
0.5	17.39	76.11	14.77	65.15	58.71	50.38	81.83	79.62
0.6	19.56	74.38	16.81	63.16	54.82	46.35	80.1	77.79
0.7	23.14	71.45	19.74	60.05	48.31	40.31	78.02	75.1
0.8	26.07	67.94	22.35	56.68	41.87	34.33	75.73	72.63
0.9	31.24	65.15	26.57	53.88	33.91	27.31	72.46	69.19
SLOC	15.35	77.87	12.59	66.76	62.52	54.17	83.87	82.20

Table 21. Faithfulness performance for varying patch probability.

Ablation Study (2)

λ_1	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
0.0	15.26	76.84	12.54	65.83	61.58	53.29	84.11	82.15
0.01	15.35	77.87	12.59	66.76	62.52	54.17	83.87	82.2
0.05	15.45	77.56	12.65	67.18	62.11	54.53	83.5	81.74
0.1	15.65	77.26	12.7	66.71	61.61	54.01	83.04	81.5
0.25	15.71	76.34	12.73	66.09	60.63	53.36	82.33	81.29
0.5	15.98	74.98	13.04	64.85	59.0	51.81	80.9	80.88
1.0	17.51	69.7	14.08	60.19	52.19	46.11	77.81	78.21

Table 15. Faithfulness evaluation. Ablation study on λ_1 - the coefficient of the L1 regularization term in Eq. 4.

λ_2	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
0.0	13.33	78.16	11.0	67.44	64.83	56.44	80.91	79.53
0.01	14.33	79.27	11.85	68.63	64.94	56.78	82.56	81.17
0.1	15.34	77.78	12.59	66.78	62.44	54.18	83.6	81.92
0.2	15.84	78.0	12.97	66.57	62.16	53.6	83.51	81.84
0.5	16.74	78.38	13.69	66.69	61.64	52.99	83.67	82.44
1.0	17.7	78.93	14.37	66.95	61.24	52.58	84.0	82.36

Table 16. Faithfulness evaluation. Ablation study on λ_2 - the coefficient of the TV regularization term in Eq. 4.

	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
$SLOC_{xTV}$	13.33	78.16	11.0	67.44	64.83	56.44	80.91	79.53
$SLOC_{xL1}$	15.26	76.84	12.54	65.83	61.58	53.29	84.11	82.15
$SLOC_{xL1xTV}$	13.89	73.60	11.35	62.95	59.70	51.60	77.80	76.28
SLOC	15.35	77.87	12.59	66.76	62.52	54.17	83.87	82.2

Table 14. Faithfulness evaluation. Ablation study on the regularization terms in Eq. 4.

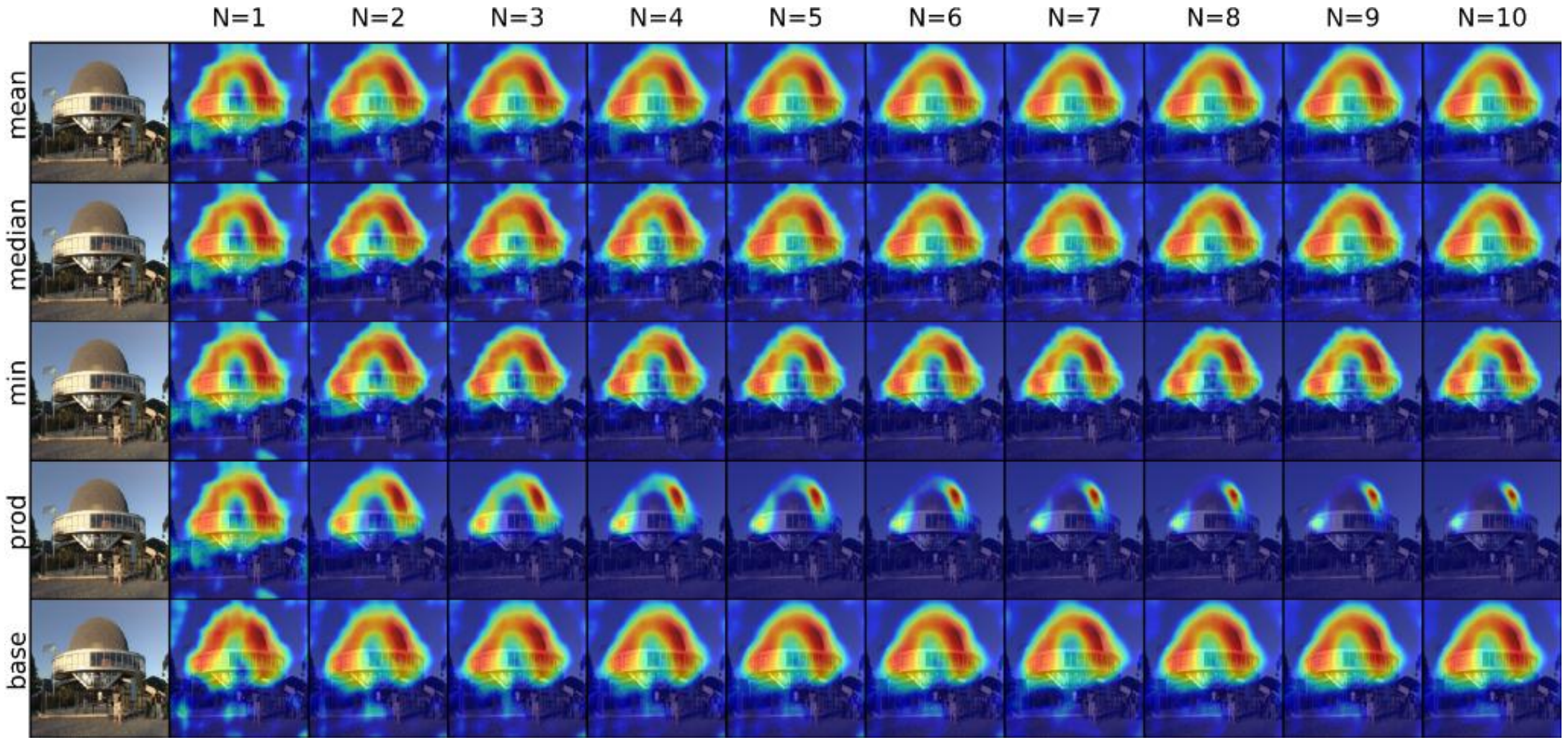
λ_1	0	0.01	0.05	0.1	0.25	0.5	0.1
mIoU↑	0.52	0.55	0.55	0.54	0.52	0.49	0.43
mAP↑	0.79	0.81	0.81	0.81	0.79	0.76	0.71
PA↑	0.71	0.74	0.75	0.74	0.72	0.70	0.64

Table 17. Segmentation evaluation. Ablation study on λ_1 - the coefficient of the L1 regularization term in Eq. 4, using the RN model.

λ_2	0	0.01	0.1	0.2	0.5	0.1
mIoU↑	0.54	0.54	0.55	0.55	0.55	0.55
mAP↑	0.8	0.8	0.8	0.81	0.81	0.82
PA↑	0.73	0.73	0.74	0.74	0.74	0.74

Table 18. Segmentation evaluation. Ablation study on λ_2 - the coefficient of the TV regularization term in Eq. 4, using the RN model.

Aggregation



Aggregation (2)

N	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
1	15.79	77.49	12.95	66.16	61.7	53.22	83.21	81.19
2	15.35	78.05	12.59	66.64	62.7	54.05	83.5	81.96
3	15.4	78.13	12.52	66.74	62.74	54.22	83.74	82.41
4	15.34	78.03	12.5	66.77	62.69	54.26	83.75	82.19
5	15.4	78.12	12.49	66.88	62.72	54.39	83.58	82.18
6	15.41	78.18	12.52	66.87	62.77	54.35	84.08	82.31
7	15.37	78.22	12.48	66.85	62.85	54.37	83.84	82.2
8	15.46	78.24	12.48	66.88	62.77	54.4	84.03	82.26
9	15.41	78.38	12.47	66.91	62.97	54.44	83.77	82.45
10	15.4	78.44	12.45	66.94	63.05	54.49	83.96	82.6

Table 23. Evaluating the effect of combining N generated attributions by **mean aggregation** to produce the final attribution map.

N	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
1	15.79	77.49	12.95	66.16	61.7	53.22	82.77	81.36
2	15.4	78.76	12.68	67.51	63.36	54.83	83.58	81.68
3	15.42	79.73	12.68	68.51	64.31	55.83	83.61	81.96
4	15.27	80.13	12.66	68.84	64.86	56.18	83.75	82.17
5	15.39	80.24	12.66	69.18	64.86	56.52	83.67	82.53
6	15.41	80.42	12.68	69.41	65.02	56.73	83.84	82.37
7	15.48	80.5	12.69	69.66	65.02	56.96	83.63	82.27
8	15.35	80.84	12.66	69.68	65.49	57.02	84.34	82.38
9	15.58	80.84	12.75	69.87	65.26	57.12	84.11	82.55
10	15.36	80.7	12.74	69.82	65.34	57.08	83.9	82.74

Table 24. Evaluating the effect of combining N generated attributions by **median aggregation** to produce the final attribution map.

N	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
1	15.79	77.49	12.95	66.16	61.7	53.22	82.62	81.57
2	15.4	78.76	12.68	67.51	63.36	54.83	83.09	81.6
3	15.48	79.06	12.57	67.9	63.57	55.34	83.19	81.61
4	15.38	79.07	12.58	68.04	63.69	55.46	83.34	81.81
5	15.37	79.08	12.53	68.14	63.72	55.6	83.41	81.74
6	15.3	79.19	12.47	68.26	63.89	55.8	83.5	81.93
7	15.27	79.2	12.49	68.19	63.93	55.7	83.1	81.88
8	15.43	79.21	12.51	68.24	63.78	55.73	83.28	81.89
9	15.29	79.28	12.5	68.4	64.0	55.9	83.63	81.99
10	15.29	79.65	12.49	68.51	64.36	56.02	82.98	82.11

Table 25. Evaluating the effect of combining N generated attributions by **minimum aggregation** to produce the final attribution map.

N	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
1	15.79	77.49	12.95	66.16	61.7	53.22	82.75	81.23
2	15.48	77.24	12.67	65.75	61.76	53.07	83.12	81.89
3	15.3	77.01	12.58	65.48	61.71	52.9	83.3	81.62
4	15.35	76.77	12.59	65.27	61.43	52.67	83.16	81.56
5	15.38	76.86	12.6	65.22	61.47	52.61	82.81	81.67
6	15.58	76.86	12.67	65.02	61.28	52.35	83.31	81.59
7	15.67	76.87	12.72	64.88	61.2	52.16	82.73	81.36
8	15.67	76.46	12.72	64.62	60.79	51.9	82.5	81.34
9	15.7	76.63	12.75	64.63	60.93	51.88	82.63	81.31
10	15.8	76.16	12.74	64.48	60.37	51.73	82.76	80.91

Table 26. Evaluating the effect of combining N generated attributions by **multiplying** them element-wise to produce the final attribution map.

N	POS↓	NEG↑	DEL↓	INS↑	NPD↑	IDD↑	AIC↑	SIC↑
1	15.38	77.54	12.74	66.16	62.16	53.42	83.48	81.18
2	15.16	77.73	12.51	66.71	62.57	54.2	83.58	81.99
3	15.1	78.3	12.53	67.11	63.21	54.57	84.06	82.0
4	15.15	78.25	12.42	67.12	63.1	54.7	83.54	82.22
5	15.13	78.0	12.37	67.23	62.87	54.86	83.94	82.36
6	15.17	78.18	12.43	67.3	63.01	54.87	83.83	82.34
7	15.17	78.11	12.42	67.15	62.94	54.73	84.13	82.29
8	15.15	78.22	12.38	67.32	63.07	54.94	83.92	82.49
9	15.29	78.41	12.38	67.33	63.13	54.95	84.03	82.46
10	15.11	78.24	12.33	67.31	63.13	54.98	83.89	82.6

Table 27. Single-run experiment. Each row reports faithfulness results obtained by SLOC using $N|\mathcal{M}|$ sampled masks for varying values of N (to match the total number of masks used across N attributions in the aggregation experiments).