Supplementary material for: Explaining through Transformer Input Sampling

1. Transformer Input Sampling Parameters

Using the quantitative experiments described in the paper for faithfulness metrics, we compared the main parameters of TIS: the number of tokens sampled and the number of masks. We first compared token masking ratios of 0.5, 0.25, and 0.125, respectively corresponding to 98, 49, and 24 sampled tokens (n_k) out of 196 tokens using 2048 masks. The results can be seen in Table 1 and 2 for the faithfulness metrics, showing that the ratio of 0.5 performs best for TIS. We then explored the impact of the reduction of the number of masks (N_m) using the best ratio (0.5) on the faithfulness metrics with 2048, 1024, 512, 256, and 128 masks. The results are in Table 3 and 4 for the faithfulness metrics, and Table 5 shows the results for the Pointing Game metric. For faithfulness and localization (Pointing Game), the higher the number of masks, the better the metrics. However, decreasing the number of masks still provides good results on the metrics, even with as low as 128 masks. It is therefore a tradeoff between speed and quality, the choice being left to the user.

2. Inference Time

We conducted an analysis of the mean inference time for explainability methods applied to the 2000 test images of the ILSVRC2012 dataset using the ViT model, as shown in Tab. 6. Overall, perturbation-based methods are slower due to the computation and prediction of mask-based images they employ, compared to methods that use gradients or relevance to obtain saliency maps. This characteristic is not new in the state-of-the-art literature.

However, we found that TIS can yield comparable inference time results to gradient-based methods when using as few as 128 masks. Additionally, TIS demonstrates only a slight degradation in performance, as indicated in Figure 3 and Figure 4. In Tab. 7, we present the inference time while varying TIS hyperparameters: the number of masks and the token ratio. The results demonstrate that TIS can be employed with inference times superior to many gradient-based methods (except for Rollout and Chefer2) when the token ratio is reduced to 0.125 with 128 masks.

3. Sparseness metric

The Sparseness metric gives a theoretical score of 1 for a saliency map with an infinite number of pixels of which only one differs from 0 and conversely gives a score of 0 if all pixels of a map have the same value. Overall, we consider Sparseness as an additional indicator to compare metrics using another property, and not as a ranking to follow to find the best method, as obtaining 0 or 1 is not desirable for a saliency map.

The results are shown in Table 8. Most of the methods have close sparseness scores (range of 0.2 - 0.3) and TIS is on the high end of this range, highlighting slightly more pixels than the other methods. At the top of the list are Attention Rollout and Chefer1 with much narrower areas highlighted (range of 0.45 - 0.6 sparseness). At the bottom of the list, we find Integrated Gradients being very close to a score of 0.

4. Additional Visual Comparisons

In order to present visual results in a broader scope, we display in this Section additional visualization maps of our proposed TIS method (Figures 2 and 3), along with an additional class disagreement visualization (Figure 1).

In Figures 2 and 3, more complete comparisons with ViT-CX [12], the Transition Attention Maps (TAM) [13], the two methods from Chefer [3, 4], Attention rollout [1], the token (BT-T) and head (BT-H) methods from Bidirectional Transformers [5], RISE [7], Integrated Gradient [10] and SmoothGrad [9]. We randomly sampled 12 images from the subset of images used in our paper, displaying the visualization maps in Figure 2 and Figure 3 for ViT [6] and DeiT [11], respectively. Figure 1 illustrates an additional class disagreement, showcasing the disparity between the target and predicted classes using TiS on ViT applied to a snake image. This demonstration emphasizes that the disagreement presented in the main paper was not an isolated case.

[1] The best result is in bold.

Token ratio	Model	Insertion ↑				Deletion ↓				Insertion - Deletion ↑			
		Mean	Blur	Black	Rand	Mean	Blur	Black	Rand	Mean	Blur	Black	Rand
0.125	ViT	0.44	0.61	0.42	0.40	0.15	0.43	0.15	0.14	0.29	0.18	0.28	0.26
0.25	ViT	0.47	0.62	0.46	0.43	0.14	0.43	0.14	0.13	0.33	0.19	0.32	0.30
0.5	ViT	**0.53**	**0.67**	**0.51**	**0.48**	**0.10**	**0.38**	**0.09**	**0.09**	**0.43**	**0.29**	**0.41**	**0.39**

Table 1: Results of Insertion and Deletion metrics and their difference on 5000 images from the ImageNet Validation set [8] using TIS for the ViT-Base model [6] with different ratios of tokens and 2048 masks [1]

Token ratio	Model	Insertion ↑				Deletion ↓				Insertion - Deletion ↑			
		Mean	Blur	Black	Rand	Mean	Blur	Black	Rand	Mean	Blur	Black	Rand
0.125	DeiT	0.56	0.62	0.56	0.53	0.18	0.43	0.18	0.16	0.38	0.18	0.38	0.38
0.25	DeiT	0.55	0.62	0.55	0.52	0.20	0.44	0.20	0.17	0.35	0.18	0.35	0.35
0.5	DeiT	**0.58**	**0.65**	**0.58**	**0.55**	**0.15**	**0.39**	**0.15**	**0.14**	**0.43**	**0.26**	**0.42**	**0.41**

Table 2: Results of Insertion and Deletion metrics and their difference on 5000 images from the ImageNet Validation set [8] using TIS for the DeiT-Base [11] model with different ratios of tokens and 2048 masks [1]

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 1

[2] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *ICML*, 2020. 4

[3] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 1

[4] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 1

[5] Jiamin Chen, Xuhong Li, Lei Yu, Dejing Dou, and Haoyi Xiong. Beyond intuition: Rethinking token attributions inside transformers. *Transactions on Machine Learning Research*, 2022. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 3, 4, 5

[7] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *BMC*, 2018. 1

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 2, 3, 5, 6

[9] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 1

[10] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. 1

[11] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 1, 2, 3, 4, 6

[12] Weiyan Xie, Xiao-Hui Li, Caleb Chen Cao, and Nevin L Zhang. Vit-cx: Causal explanation of vision transformers. *arXiv preprint arXiv:2211.03064*, 2022. 1

[13] Tingyi Yuan, Xuhong Li, Haoyi Xiong, Hui Cao, and Dejing Dou. Explaining information flow inside vision transformers using markov chain. In *eXplainable AI approaches for debugging and diagnosis.*, 2021. 1

[14] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 3

Masks	Model	Insertion ↑				Deletion ↓				Insertion - Deletion ↑			
		Mean	Blur	Black	Rand	Mean	Blur	Black	Rand	Mean	Blur	Black	Rand
2048	ViT	**0.53**	**0.67**	**0.51**	**0.48**	**0.10**	**0.38**	**0.09**	**0.09**	**0.43**	**0.29**	**0.41**	**0.39**
1024	ViT	0.52	0.66	0.50	0.47	**0.10**	0.39	0.10	**0.09**	0.42	0.28	0.40	0.38
512	ViT	0.51	0.66	0.49	0.47	**0.10**	0.39	0.10	**0.09**	0.41	0.27	0.39	0.37
256	ViT	0.50	0.65	0.48	0.46	**0.10**	0.40	0.10	**0.09**	0.40	0.25	0.38	0.36
128	ViT	0.49	0.64	0.48	0.45	0.11	0.40	0.11	0.10	0.39	0.24	0.37	0.35

Table 3: Results of Insertion and Deletion metrics and their difference on 5000 images from the ImageNet Validation set [8] using TIS for the ViT-Base model [6] with a token ratio of 0.5 and different numbers of masks [1]

Masks	Model	Insertion ↑				Deletion ↓				Insertion - Deletion ↑			
		Mean	Blur	Black	Rand	Mean	Blur	Black	Rand	Mean	Blur	Black	Rand
2048	DeiT	**0.58**	**0.65**	**0.58**	**0.55**	**0.15**	0.39	**0.15**	**0.14**	**0.43**	**0.26**	**0.42**	**0.41**
1024	DeiT	0.57	**0.65**	0.57	0.54	**0.15**	0.40	**0.15**	**0.14**	0.42	0.25	**0.42**	**0.41**
512	DeiT	0.57	0.64	0.57	0.54	**0.15**	0.40	**0.15**	**0.14**	0.41	0.24	0.41	0.40
256	DeiT	0.56	0.63	0.56	0.53	0.16	0.41	0.16	**0.14**	0.40	0.23	0.40	0.39
128	DeiT	0.55	0.63	0.55	0.52	0.16	0.41	0.16	**0.14**	0.39	0.22	0.39	0.38

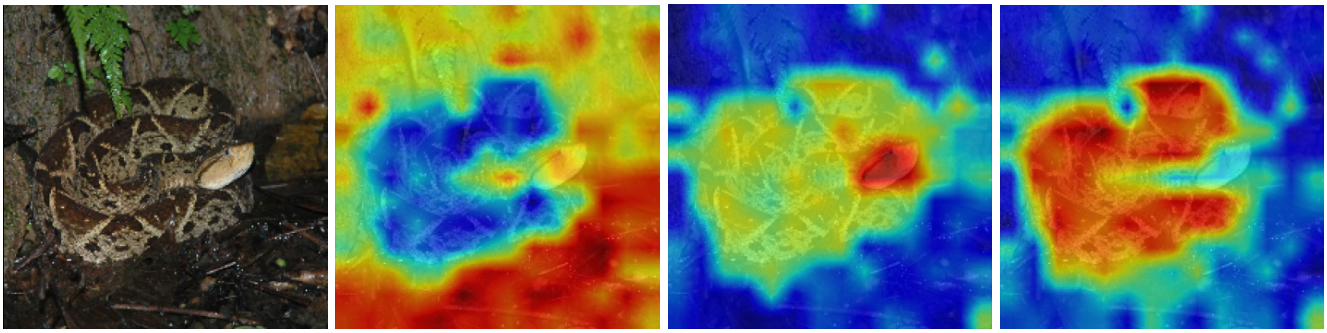
Table 4: Results of Insertion and Deletion metrics and their difference on 5000 images from the ImageNet Validation set [8] using TIS for the DeiT-Base [11] model with a token ratio of 0.5 and different numbers of masks [1]

Token ratio	Masks	DeiT	ViT
0.125	2048	0.745	0.573
0.25	2048	0.732	0.684
0.5	2048	**0.829**	**0.824**
0.5	1024	0.825	0.823
0.5	512	0.820	0.818
0.5	256	0.813	0.800
0.5	128	0.807	0.784

Table 5: Results of the Pointing Game metric [14] for the ViT [6] and DeiT [11] model with different parameters of TIS [11] [1]

Method	Inference (s)
ViT-CX	0.81
TAM	0.61
TiS 0.5 128	0.69
TiS 0.5 256	1.01
TiS 0.5 512	1.07
TiS 0.5 1024	2.05
TiS 0.5 2048	3.98
Chefer1	0.2
Chefer2	0.05
Att. Rollout	0.06
BT H	0.59
BT T	0.63
RISE	15.42
IntegratedGrad	0.52
SmoothGrad	0.54

Table 6: Mean inference time of the explainability methods applied to the 2000 test images for the ViT model using an NVIDIA GeForce RTX 3070.



(a) Original image (b) target: “hognose snake” 0.008 (c) “night snake” 0.556 (d) “horned viper” 0.100

Figure 1: Maps generated by TIS on ViT using the target class and the two higher predicted classes. In a minority of cases, a mismatch between the predicted classes and a low certainty target class leads to a negative saliency map for the target class. In this example, the model gives a certainty of 0.008 “hognose snake”, the target class of snake breed. In comparison, the TIS result for the breeds predicted with higher confidence are well formed, we can see that the head is the main element in this image

Masks \ Token ratio	0.125	0.25	0.5
128	0.31	0.42	0.69
256	0.40	0.61	1.01
512	0.41	0.62	1.07
1024	0.71	1.13	2.05
2048	1.31	2.24	3.98

Table 7: Mean inference time (in seconds) of TiS depending on the number of masks and the token ratio, applied to the 2000 test images for the ViT model using an NVIDIA GeForce RTX 3070.

Method	DeiT	ViT
TIS (ours)	0.296	0.311
ViT-CX	0.256	0.256
TAM	0.335	0.320
Chefer1	0.457	0.585
Chefer2	0.312	0.334
Attention Rollout	0.565	0.496
BT H	0.344	0.366
BT T	0.353	0.395
RISE	0.309	0.317
Integrated Gradient	0.015	0.028
SmoothGrad	0.301	0.314

Table 8: Results of the Sparseness metric [2] for the ViT [6] and DeiT model [11]. We consider Sparseness as an indicative metric and not as a ranking to be achieved, as obtaining 0 or 1 is not desirable for a saliency map.

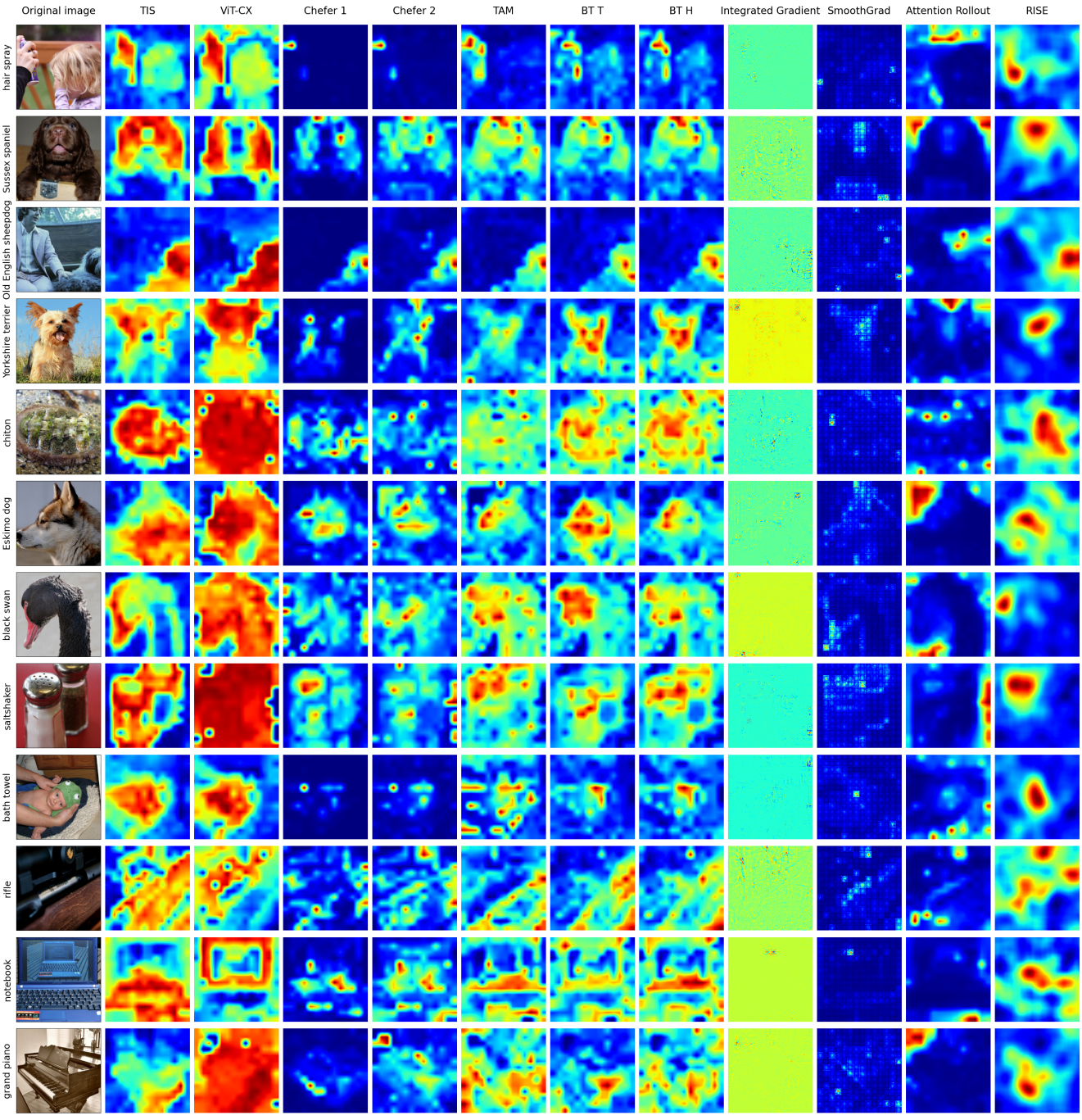


Figure 2: Comparison of the explainability methods for the ViT-Base model [6] on 12 random images from the ImageNet Validation set [8]

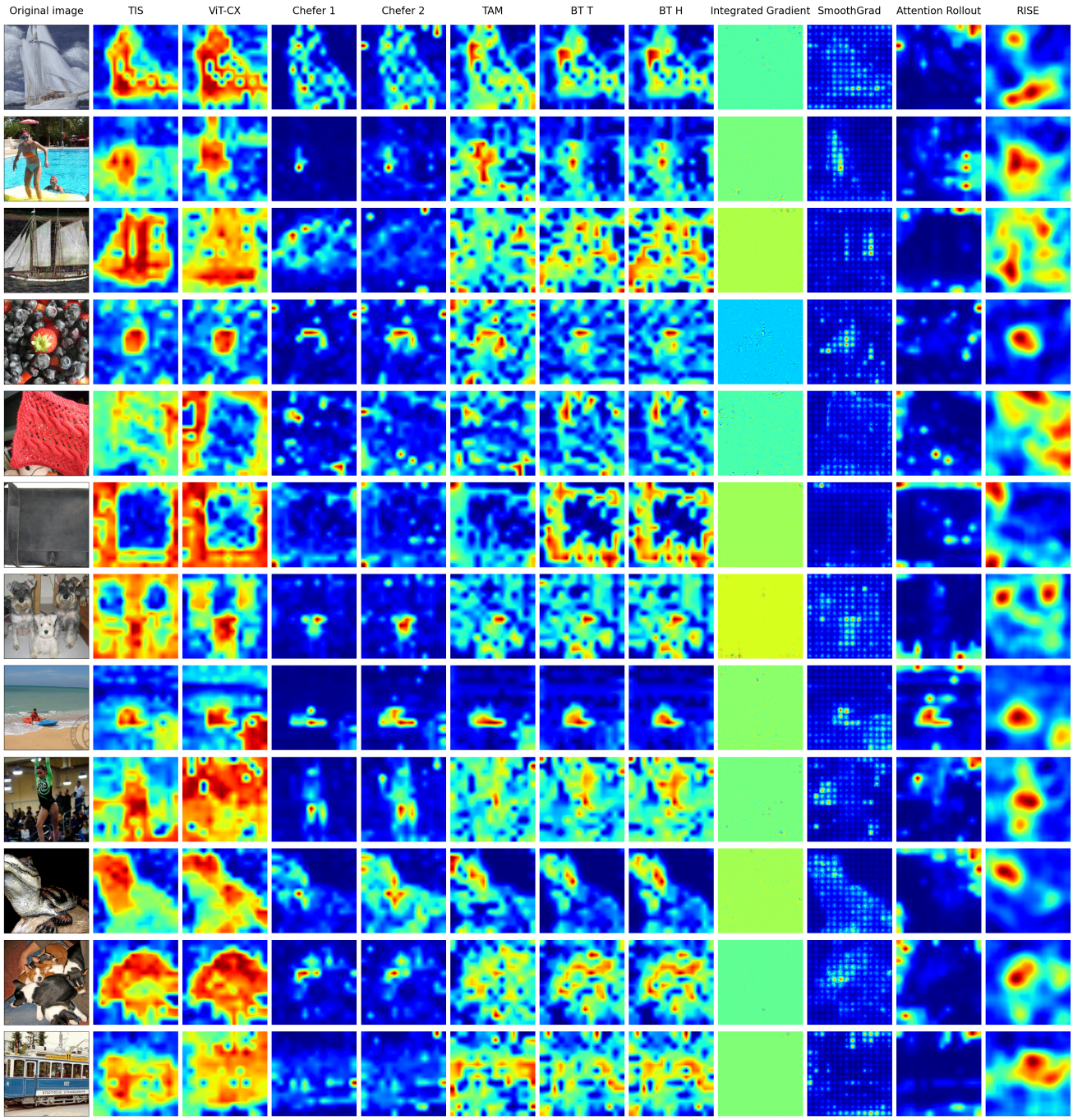


Figure 3: Comparison of the explainability method for the DeiT-Base model [11] on 12 random images from the ImageNet Validation set [8]