SkipPLUS: Skip the First Few Layers to Better Explain Vision Transformers

Faridoun Mehri[1] Mohsen Fayyaz[2] Mahdieh Soleymani Baghshah[1] Mohammad Taher Pilehvar[3]

[1]Sharif University of Technology, Iran [2]University of Tehran, Iran [3]Cardiff University, UK

faridoun.mehri@ce.sharif.edu* mohsen.fayyaz77@ut.ac.ir
soleymani@sharif.edu pilehvarmt@cardiff.ac.uk

Abstract

Despite their remarkable performance, the explainability of Vision Transformers (ViTs) remains a challenge. While forward attention-based token attribution techniques have become popular in text processing, their suitability for ViTs hasn't been extensively explored. In this paper, we compare these methods against state-of-the-art input attribution methods from the Vision literature, revealing their limitations due to improper aggregation of information across layers. To address this, we introduce two general techniques, PLUS and SkipPLUS, that can be composed with any input attribution method to more effectively aggregate information across layers while handling noisy layers. Through comprehensive and quantitative evaluations of faithfulness and human interpretability on a variety of ViT architectures and datasets, we demonstrate the effectiveness of PLUS and SkipPLUS, establishing a new state-of-the-art in white-box token attribution. We conclude with a comparative analysis highlighting the strengths and weaknesses of the best versions of all the studied methods. The code used in this paper is freely available at https://github.com/NightMachinery/SkipPLUS-CVPR-2024.

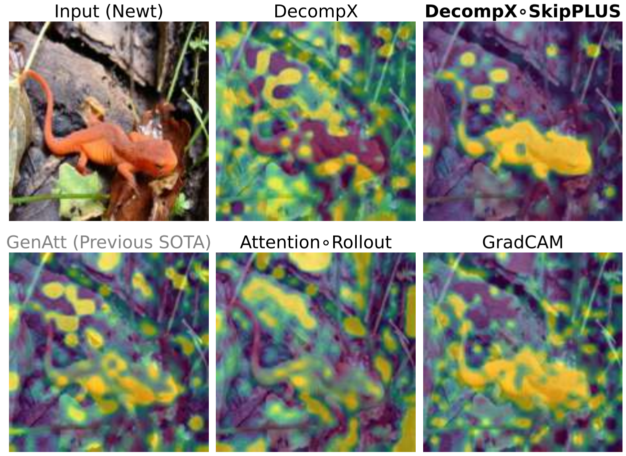

Figure 1. This figure presents a qualitative comparison of the composition of the proposed SkipPLUS method with the Forward Attention-Based Token Attribution technique DecompX, using the EVA Large model. Brighter shades signify an increased positive contribution of features towards the prediction of the target class. The results illustrate that DecompX-SkipPLUS effectively concentrates on the target class ("Newt" in this image) while minimizing noise. In contrast, GenAtt, the previous state-of-the-art method, exhibits suboptimal performance on EVA Large, falling behind baselines such as GradCAM. For additional qualitative examples, including multi-class instances, refer to Figs. 5 and 6, as well as the appendices.

1. Introduction

Transformers currently dominate various NLP tasks and are gaining significant popularity in the field of computer vision [20, 23, 53, 57, 71]. Despite their remarkable success, a crucial challenge remains in comprehending their inner workings, which poses risks in real-world deployments. Consequently, there is an increasing demand for research that explains the outputs of Transformers [15, 41, 48, 60].

Input attribution methods are techniques designed to quantify the influence of individual input features, or groups of them, on a model's output [6, 38, 43, 44, 61, 68, 69, 80]. Input attribution methods can assist in understanding a model's decision locally for a single input considered in isolation. They also act as foundational elements for more advanced explanation techniques. For instance, in concept-based explanation methods like CRAFT [25], attribution methods are employed for two main purposes: to quantify the impact of each activated concept and to identify the specific input features responsible for activating these concepts.

Early works used the raw self-attention weights of the last layer as a token attribution map [8, 11, 32]. Recent studies have questioned the reliability of this approach, given

*You can also reach me at feraidoonmehri@gmail.com.

This work was published at the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

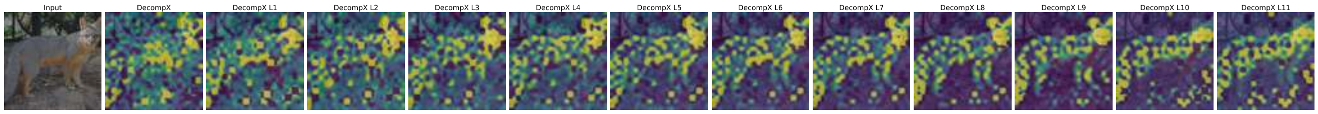

Figure 2. Starting DecompX from Different Layers

that self-attention is only a small part of a Transformer block [15, 35, 64, 75]. Forward Attention-Based Token Attribution methods [1, 26–28, 47] try to address this issue by incorporating other components of the block. These methods have been primarily developed and evaluated on text Transformers and their effectiveness in Vision Transformers has not been thoroughly investigated.

In this paper, we conduct a comprehensive study comparing Forward Attention-Based Token Attribution methods against popular interpretation techniques used in the Vision literature. Our findings reveal that, in their original formulation, Forward Attention-Based Token Attribution methods do not perform well for Vision Transformers, mainly due to their improper aggregation of information across layers.

To address this limitation, we propose PLUS and SkipPLUS, simple techniques that can be applied to a wide range of input attribution methods, without imposing any specific prerequisites. We perform extensive quantitative evaluations on a variety of ViT architectures and datasets to assess the faithfulness and human interpretability of the proposed method. Our results show that composing SkipPLUS with recent text-based interpretation techniques can significantly enhance their performance, surpassing all the widely-used techniques for interpreting vision models. Encouraged by this success and the composability of these techniques, we compose PLUS and SkipPLUS with other attribution methods. This integration enhances the effectiveness of many traditional methods without compromising performance when it proves unhelpful. (See Fig. 4 and Sec. C in the appendices for further details.)

Finally, we provide a comparative analysis of the best versions of all the studied methods, highlighting their strengths and weaknesses.

2. Related Work

Owing to space constraints, this section will briefly introduce only the most fundamental methods. A comprehensive overview of additional methods is provided in the appendix, detailed in Sec. E.

2.1. Gradient-Based Methods

Input×Gradient (IxG). IxG [2, 66, 67] multiplies the input values by their corresponding gradients. Let x_i be a spatial feature of the input, where $x_{i,j}$ represents the j-th channel of x_i. The Input×Gradient attribution for the spatial feature x_i with respect to the target class c is computed as follows, where y_c is the output of the model for the target class c:

$$\text{Input}\times\text{Gradient}_i = \sum_j \frac{\partial y_c}{\partial x_{i,j}} \cdot x_{i,j}, \tag{1}$$

2.2. Forward Attention-Based Token Attribution Methods

Rollout. Rollout [1] is a technique used to aggregate the attribution maps from different self-attention layers in Transformer models. While Rollout was originally developed for aggregating attention weights across layers, it can be applied to any attribution method that produces a 2D From-To attribution matrix for each layer. The Rollout method linearly multiplies the attribution maps from each layer.

Given the attribution map $\mathbf{T}_i$ at the i-th layer, the Rollout operation updates the accumulated attribution map $\mathbf{R}_i$ that has aggregated information up to the i-th layer. The update rule for the accumulated attribution map is as follows:

$$\mathbf{R}_0 = \mathbf{I}, \quad \mathbf{R}_{i+1} = (0.5\mathbf{I} + 0.5\overline{\mathbf{T}}_i)\mathbf{R}_i \tag{2}$$

where $\mathbf{I}$ is the identity matrix, and $\overline{\mathbf{T}}_i$ is the normalized attribution map of the i-th layer. The normalization ensures that the sum of each row in the attribution map is one, mirroring the behavior of attention weights. Some methods forgo the $0.5\mathbf{I}$ term and/or the normalization step in the Rollout operation.

3. Methodology

3.1. Progressive Layer Unification Through Summation (PLUS)

Let f be a model with L layers (numbered from 0 to $L-1$), and A be an attribution method that takes a model and an input, and produces an attribution map. Given an input x_0 to the model, we define Progressive Layer Unification Through Summation (PLUS) as:

$$\text{PLUS}(f, A)(x_0) = \sum_{l=0}^{L-1} A(f_l)(x_l) \tag{3}$$

where f_l is the sub-network of f starting from layer l (and going until the end of the model), and x_l is the intermediate output of the model at layer $l-1$ when the input is x_0. Note that the layer numbers are zero-based, so f_0 is equivalent to the full model f. In essence, PLUS varies the start layer of the underlying attribution method A from the first to the last layer and aggregates the resulting attribution maps through summation.

3.2. SkipPLUS

SkipPLUS is a variant of PLUS that starts the aggregation from the middle layer instead of the first layer. Formally, let $m = \lceil \frac{L}{2} \rceil$ be the middle layer of the model f. Then, SkipPLUS is defined as:

$$\text{SkipPLUS}(f, A)(x_0) = \sum_{l=m}^{L-1} A(f_l)(x_l) \tag{4}$$

In other words, SkipPLUS skips the first m layers and only aggregates the attribution maps starting from the middle layer m.

For a discussion on the justification and insights behind these methods, refer to Sec. 5.1.

3.3. FullGrad+

FullGrad [70] extends the Input×Gradient (IxG) method [38] by computing attribution maps not only for the original input but also for each bias term in the network. The final attribution map is obtained by summing the IxG attribution map of the input with the bias attribution maps.

Directly applying PLUS or SkipPLUS on FullGrad would lead to multiple sums of the bias attribution maps of the later layers, as the bias attribution maps of layer l also appears in FullGrad(f_k) for k less than l. To avoid this repetition, we define FullGrad+ as follows:

$$\begin{aligned}&\text{FullGrad+} \circ \text{PLUS}(f)(x_0) = \\ &\qquad \sum_{l=0}^{L-1} \text{IxG}(f_l)(x_l) + \sum_{l=0}^{L-1} \sum_{b \in B_l} \text{IxG}(f_b)(b)\end{aligned} \tag{5}$$

$$\begin{aligned}&\text{FullGrad+} \circ \text{SkipPLUS}(f)(x_0) = \\ &\qquad \sum_{l=m}^{L-1} \text{IxG}(f_l)(x_l) + \sum_{l=m}^{L-1} \sum_{b \in B_l} \text{IxG}(f_b)(b)\end{aligned} \tag{6}$$

where IxG$(f_l)(x_l)$ is the Input×Gradient attribution map of the sub-network f_l with input x_l, and IxG$(f_b)(b)$ is the Input×Gradient attribution map of the sub-network f_b with a bias term b from layer l as the input. f_b is the sub-network of f starting from the bias term b and going until the end of the model. B_l denotes the set of all bias terms in layer l. FullGrad+ aggregates the input attribution maps of each layer along with the attribution maps of all bias terms in each layer, ensuring no repetition occurs. Refer to Fig. 10 (in the appendices) for a quantitative evaluation of FullGrad+ versus FullGrad, and Fig. 6 for a qualitative comparison.

3.4. Special Cases of PLUS in Prior Work

The methods described below can be viewed as special cases of PLUS, where PLUS is composed with a previously existing method.

GradSAM. GradSAM [4] is equivalent to composing GenAtt [10] with PLUS, instead of using Rollout. (cf. Fig. 9 in the appendices)

CAT. Class Activation Tokens [56] is equivalent to IxG∘PLUS. (cf. Fig. 11 in the appendices)

AttCAT. We can define an attention-enhanced variant of IxG, AttIxG, by multiplying IxG with AttnFrom:

$$\texttt{AttnFrom}_j = \frac{1}{H \times N} \sum_{h=1}^{H:=\text{Heads}} \sum_{i=1}^{N:=\text{Tokens}} \text{RawAttn}_{h,i,j}$$

Note that attention weights have three dimensions: heads, to, from.

Attentive Class Activation Tokens [56, AttCAT] would then be equivalent to AttIxG∘PLUS. (cf. Fig. 11 in the appendices)

LayerCAM. LayerCAM [36] was introduced for ReLU CNN networks, where it is equivalent to applying a normalization process on the layer-wise attribution maps obtained from GradCAMElementWise [31], followed by the PLUS aggregation method. The normalization step is proposed because earlier layers tend to have smaller attribution maps compared to later layers. By normalizing the maps, LayerCAM ensures that each layer contributes more equally to the final attribution map. However, this approach is not suitable for ViTs, as we explicitly want to avoid giving earlier layers the same impact on the final attribution map as later layers (cf. Fig. 2, also supported by our preliminary quantitative evaluations).

4. Experimental Setup

4.1. Faithfulness Evaluation Metrics

Modern literature favors evaluations for input attribution methods that are collectively called faithfulness, which intuitively measures how well the attribution scores reflect the true contribution of each input feature to the target output. Although several metrics have been proposed to quantify faithfulness, we adopt the most comprehensive approach, which involves computing the area under the curve (AUC) for the deletion and insertion operations, considering the changes in accuracy and the target probability [14, 27, 47, 50].

The deletion accuracy curve is obtained by progressively removing input features in order of decreasing attribution scores and measuring the model's accuracy at each step. A faithful attribution method should result in a steep drop in performance as the most important features are removed first. The deletion accuracy scores are normalized using the formula $100-x$, where x is the original score, so that higher scores always indicate better performance.

Similarly, the deletion AOPC curve is generated by gradually removing input features in order of decreasing attribution scores and evaluating the change in the target output probability at each step. A faithful attribution method should lead to a rapid decrease in the target probability as the most important features are removed first.

Conversely, the insertion accuracy curve is generated by gradually adding input features in order of decreasing attribution scores and evaluating the model's performance at each step. A faithful attribution method should lead to a rapid increase in performance as the most important features are added first. Sec. A.1 in the appendices explains these metrics in more detail.

True Token Masking. Instead of simply overlaying a color mask, we choose to completely exclude the masked patches from the model's input [16]. At the same time, we preserve accurate positional encodings for the unmasked patches. (cf. Sec. A.2 in the appendices)

4.2. Human Interpretability Evaluation

Although lacking a strong theoretical justification, human interpretability evaluations serve as effective sanity checks and provide a quantitative measure that aligns with intuitive inferences drawn from qualitative examples of attribution methods. Following the zero-shot segmentation setup proposed by [11], we report the Average Precision (AP) metric. This evaluation requires a dataset with ground truth labels for the target class. Notably, AP is invariant to shift and scale transformations, mirroring the properties of our faithfulness metrics.

4.3. Models, Task, and Datasets

We assess three models on two datasets. First, we employ EVA Large (Patch Size 14) [23], a top-performing ViT model in the `timm` library [76], pretrained on image-text reconstruction and finetuned on ImageNet [18]. Second, we use ViT Base (Patch Size 8) [20], pretrained and finetuned on ImageNet, choosing a model size (Base) and a patch size (8) to maximally differ from our previous choice of EVA Large (Patch Size 14); this aligns with prior work [11, 78] that evaluate attribution methods on the vanilla ViT. Third, following [16], we use MURA ViT Base (Patch 16), trained on the MURA dataset. MURA [58] contains bone X-rays labeled as normal or abnormal.

All models serve as image classifiers, and the attribution target is set to the ground truth label for a better assessment of class discriminativity [11]. We use ImageNet due to its prevalence in prior work, the availability of high-quality ViT finetunes in `timm`, and its challenging 1000-class setting. We randomly select 5000 images from the ImageNet-1k validation set and 2000 images from the MURA training set, using fixed seeds for reproducibility.

For segmentation evaluations, we use ImageNet-S [30] ground truth segmentation maps, which encompasses 919 distinct classes, with a random subset of 5000 images from the validation set. The target is set to the class with the largest area in the ground truth segmentation map. Token attribution methods generate token-level rather than pixel-level maps. Consequently, we apply nearest interpolation to upscale these token-level maps to pixel-level.

5. Results

5.1. Justification for PLUS and SkipPLUS

Fig. 3 shows that Forward Attention-Based Token Attribution methods, most of which use Rollout, are not competitive with the previous SOTA in ViT-specific attribution methods. This underperformance can be partially attributed to the vanishing attributions problem, which arises from the multiplication of small numbers in each layer, resulting in nearly zero values in the final aggregated output.

The severity of the vanishing attributions problem varies depending on the combination of the model and the attribution method used. For example, ALTI-Rollout [27] suffers from severe vanishing attributions on ViT Base due to the presence of ReLU operations in the layer-wise maps, resulting in almost all-zero attribution maps (Fig. 5) and performance worse than the random baseline. However, it performs competitively on EVA Large. GlobEnc-Rollout [45], which does not involve ReLU operations, avoids vanishing attributions on ViT Base but exhibits the issue on EVA Large (cf. D.2 in the appendices). Other Rollout-based methods, such as GenAtt [10], do not encounter the vanishing attributions problem on any of the evaluated models.

DecompX [47], a successor to older methods such as GlobEnc and ALTI, does not need to produce layer-wise attribution maps and aggregate them separately. However, Fig. 2 shows that when DecompX starts from the initial layers and propagates attribution scores to the end of the network, it results in noisy attribution maps. Simply starting later can significantly boost the performance, but choosing this optimal layer can be challenging. As many leading

EVA2 and EVA Giant outperform EVA Large, but our fork of `timm` did not support them. We also lacked the resources for evaluating on the Giant variants. (cf. https://github.com/huggingface/pytorch-image-models/blob/main/results/results-imagenet-real.csv)

Transformer-specific methods (e.g., GenAtt) aggregate information across multiple layers using Rollout, one might be tempted to use Rollout on DecompX. However, this is not (naively) possible, as Rollout requires a 2D From-To attribution matrix for each layer, while DecompX simply produces a 1D From-Target vector with a single target.

These observations lead us to propose Progressive Layer Unification Through Summation (PLUS). PLUS involves varying the start layer of the underlying attribution method from the very first layer to the very last and aggregating the resultant attribution maps through summation. While PLUS successfully enhances the performance of Forward Attention-Based Token Attribution methods, surpassing previous white-box state-of-the-art methods from ViT and CNN literature, we might wonder what happens if we drop the noisy layers altogether from our aggregated output. We conduct evaluations dropping the starting layers one by one from the layers considered in PLUS (cf. Sec. B in the appendices). We see that two points emerge naturally to start the aggregation from: one is the very first layer, and the other is the middle layer. We name this latter variant SkipPLUS, which achieves state-of-the-art performance (cf. Fig. 4).

As PLUS and SkipPLUS have no constraints on the base attribution method they wrap around, we also evaluate their compositions with many other methods (cf. Sec. C in the appendices). This investigation leads to improving several methods. Notably, even when composing a method with PLUS does not help, it usually does not degrade performance considerably either. This makes PLUS and SkipPLUS robust methods that can be used with other methods without thorough evaluations.

5.2. Comprehensive Benchmark of White-Box Attribution Methods on Vision Transformers

Having enhanced several underperforming and underrated methods, we now present a thorough and modern benchmark of the best versions of all the methods studied in Fig. 4. DecompX∘SkipPLUS outperforms all other methods, including its original version DecompX, by a significant margin, except in the insertion faithfulness tests on EVA Large, where it remains competitive with the best method, AttIxG∘PLUS. In general, the top-performing methods are compositions of PLUS and SkipPLUS, with some even incorporating classic methods such as IxG∘SkipPLUS.

Another interesting observation is the high performance of the random baseline on the insertion faithfulness tests, which signifies the robustness of Vision Transformers to random token omissions. In insertion tests, the goal is to insert patches so that the model reaches the correct target class faster; as the model is robust against random omissions, this happens quickly. However, the random baseline does not fare well in the deletion faithfulness tests. In deletion tests, the aim is to adversarially delete strategic tokens to change the model's decision. Here, the model's robustness acts against the random baseline.

In general, we observe a trade-off between insertion performance and deletion, and a positive correlation between deletion performance and segmentation performance (AP). This trade-off can also be seen when selecting which layers to drop from PLUS in Sec. C in the appendices. Prior work has also reported similar trade-offs [61]. However, a strong method such as DecompX∘SkipPLUS manages to achieve almost optimal performance in all metrics, highlighting the possibility of attaining high performance despite the trade-offs.

6. Conclusion

We conducted a comprehensive evaluation of white-box token attribution methods for Vision Transformers (ViTs). We compared Forward Attention-Based Token Attribution methods, originally developed for text Transformers, against state-of-the-art input attribution methods from the ViT and CNN literature. Our analysis revealed the limitations of these methods due to improper aggregation of information across layers.

To address these limitations, we introduced Progressive Layer Unification Through Summation (PLUS) and SkipPLUS, two general techniques that can be combined with any input attribution method to more effectively aggregate information across layers while handling noisy layers. Through extensive quantitative evaluations of faithfulness and human interpretability on various ViT architectures and datasets, we demonstrated the effectiveness of PLUS and SkipPLUS. We also conducted thorough qualitative comparisons, including an analysis of multi-class qualitative examples to assess class discriminativity. Our comprehensive approach, combining quantitative and qualitative analyses, establishes a new state-of-the-art in white-box token attribution.

Future work could explore the application of these techniques to other domains (e.g., text Transformers) and investigate combining them with other explainability methods to further improve the interpretability of Transformers.

Acknowledgements

I thank my parents for their unwavering support, which made this work possible.

Figure 3. Composing PLUS and SkipPLUS with Forward Attention-Based Token Attribution methods is helpful in increasing performance across all metrics.

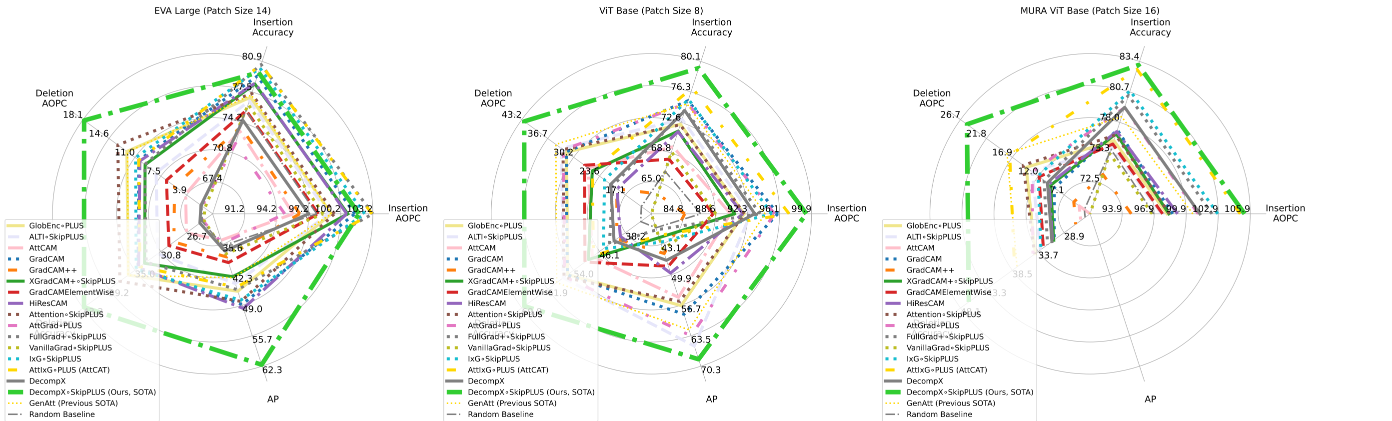

Figure 4. Best versions of methods compared against each other. The axis hidden under the legend corresponds to Deletion Accuracy. (cf. Sec. 5.2 for analysis)

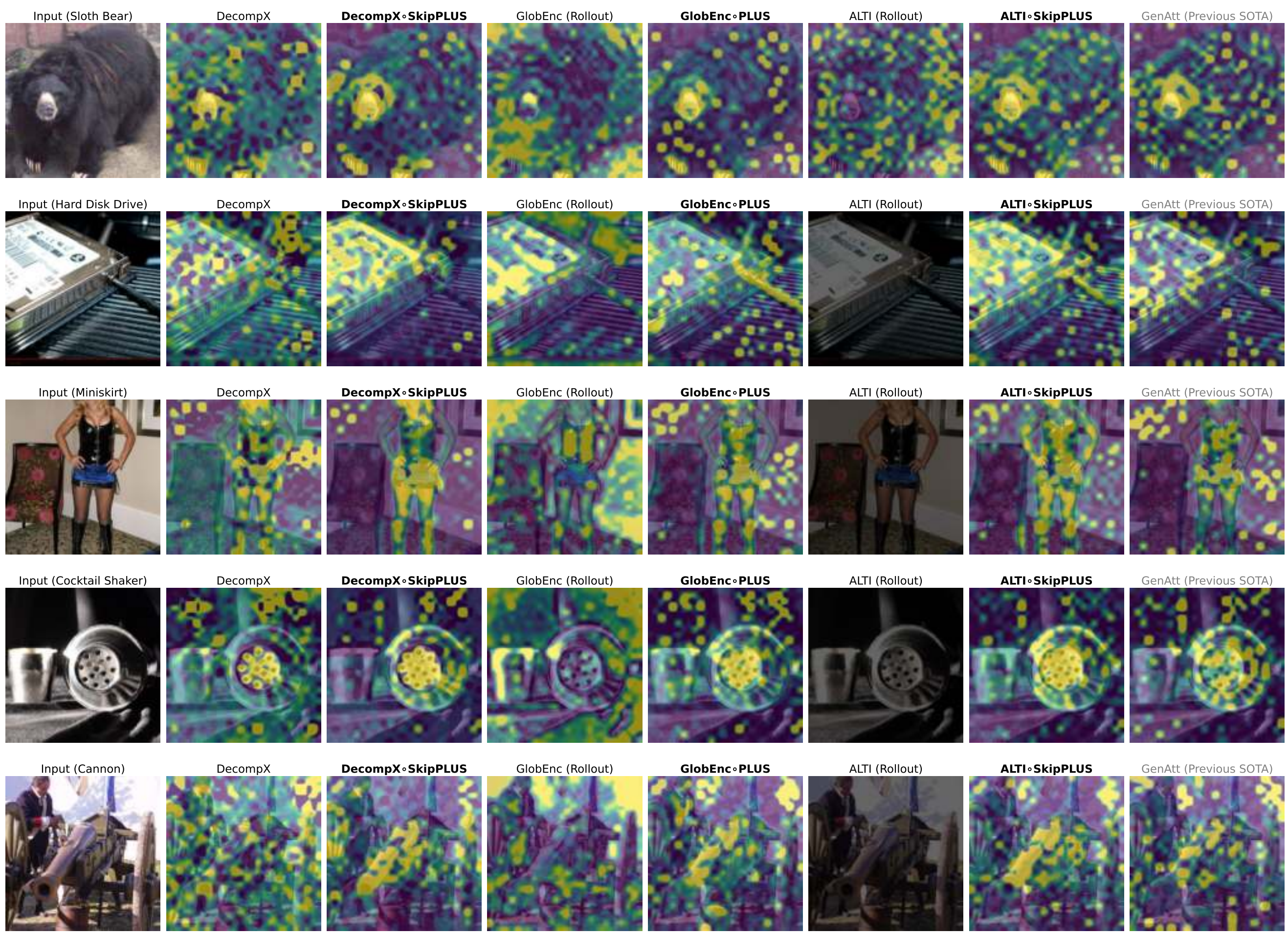

Figure 5. The SkipPLUS method can be applied in conjunction with any attribution technique to improve its performance. In contrast, the Rollout aggregation approach is not robust; its multiplicative properties frequently lead to suboptimal interactions with the ReLU operation in ALTI, resulting in attribution maps that are largely composed of zero values. Further qualitative examples are provided in the appendices. The model employed in this figure is ViT Base (Patch Size 8).

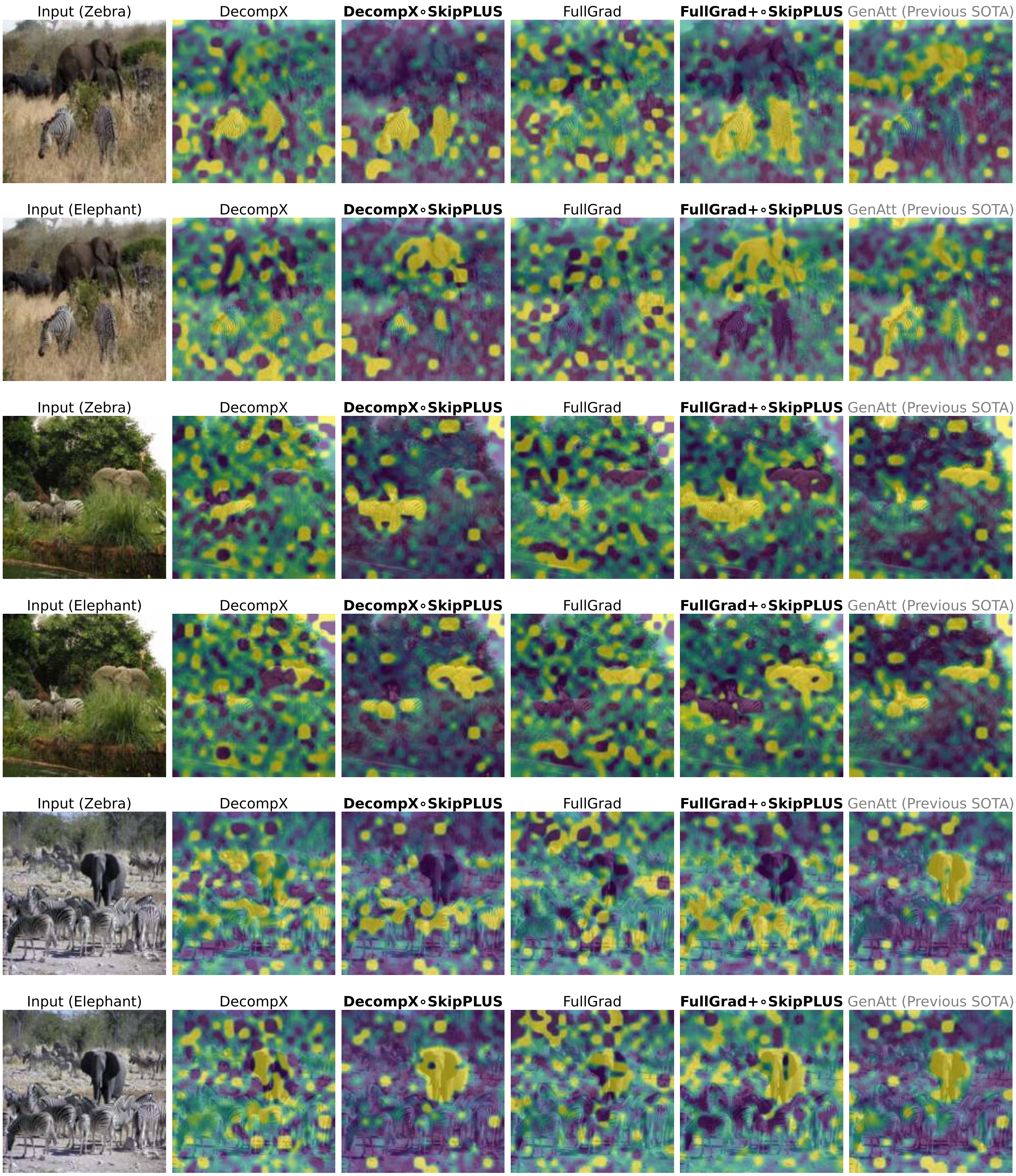

Figure 6. This figure presents a preliminary qualitative evaluation of the class discriminativity of various attribution methods applied to the EVA Large model. Theoretically class-insensitive methods such as ALTI and GlobEnc have been excluded from this analysis. The images selected for this evaluation are among the few suitable instances in the COCO 2017 training set [42] that contain both zebras and elephants within the same frame, with the animals mostly visible and not cropped out. We chose zebras and elephants because prior work, such as [34], has also used these animals in their evaluations. Additionally, ImageNet has a single class for zebras and three classes for elephants (we chose "African Elephant" as the target class here), which is in contrast to most other animals that can have tens of different fine-grained ImageNet classes.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, 2020. Association for Computational Linguistics. 2, 34

[2] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus H. Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2017. 2, 34

[3] Christopher J. Anders, David Neumann, Talmaj Marinc, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Xai for analyzing and unlearning spurious correlations in imagenet. 2020. 34

[4] Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. Grad-sam: Explaining transformers via gradient self-attention maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, page 2882–2887, New York, NY, USA, 2021. Association for Computing Machinery. 3, 35

[5] Daniel Becking, Maximilian Dreyer, Wojciech Samek, Karsten Müller, and Sebastian Lapuschkin. Ecqx: Explainability-driven quantization for low-bit and sparse dnns. *ArXiv*, abs/2109.04236, 2021. 34

[6] Alexander Binder, Sebastian Bach, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for deep neural network architectures. 2016. 1, 34

[7] Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations*, 2020. 35

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 1

[9] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2017. 34

[10] 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 (ICCV)*, pages 397–406, 2021. 3, 4, 35

[11] 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 (CVPR)*, pages 782–791, 2021. 1, 4, 34

[12] Hila Chefer, Idan Schwartz, and Lior Wolf. Optimizing relevance maps of vision transformers improves robustness. *ArXiv*, abs/2206.01161, 2022. 34

[13] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ArXiv*, abs/2301.13826, 2023. 34

[14] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 3, 14

[15] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, 2019. Association for Computational Linguistics. 1, 2

[16] Ian Covert, Chanwoo Kim, and Su-In Lee. Learning to estimate shapley values with vision transformers. *ArXiv*, abs/2206.05282, 2022. 4, 14

[17] Mayukh Deb, Björn Deiseroth, Samuel Weinbach, Patrick Schramowski, and Kristian Kersting. Atman: Understanding transformer predictions through memory efficient attention manipulation. *ArXiv*, abs/2301.08110, 2023. 36

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[19] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, 2020. Association for Computational Linguistics. 14

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 4

[21] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. 2020. 34

[22] Sami Ede, Serop Baghdadlian, Leander Weber, An Thai Nguyen, Dario Zanca, Wojciech Samek, and Sebastian Lapuschkin. Explain to not forget: Defending against catastrophic forgetting with xai. In *International Cross-Domain Conference on Machine Learning and Knowledge Extraction*, 2022. 34

[23] Yuxin Fang, Wen Wang, Binhui Xie, Quan-Sen Sun, Ledell Yu Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369, 2022. 1, 4

[24] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid

Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Juergen Gall. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, 2021. 34

[25] Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien Colin, R'emi Cadene, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2711–2721, 2022. 1, 34

[26] Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta Ruiz Costa-jussà. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. *ArXiv*, abs/2205.11631, 2022. 2

[27] Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 3, 4, 14, 35

[28] Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta Ruiz Costa-jussà. Explaining how transformers use context to build predictions. *ArXiv*, abs/2305.12535, 2023. 2

[29] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *ArXiv*, abs/2008.02312, 2020. 34

[30] Shangqi Gao, Zhong-Yu Li, Ming-Hsuan Yang, Mingg-Ming Cheng, Junwei Han, and Philip H. S. Torr. Large-scale unsupervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:7457–7476, 2021. 4

[31] Jacob Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021. 3, 34, 35

[32] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. In *AAAI Conference on Artificial Intelligence*, 2020. 1

[33] Yi Huang and Adams Wai-Kin Kong. Transferable adversarial attack based on integrated gradients. *ArXiv*, abs/2205.13152, 2022. 34

[34] Brian Kenji Iwana, Ryohei Kuroki, and Seiichi Uchida. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4176–4185, 2019. 8

[35] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2

[36] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. 3, 35

[37] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. *ArXiv*, abs/2308.12964, 2023. 34

[38] Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. Investigating the influence of noise and distractors on the interpretation of neural networks. *CoRR*, abs/1611.07270, 2016. 1, 3, 34

[39] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online, 2020. Association for Computational Linguistics. 35

[40] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4547–4568, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 35

[41] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China, 2019. Association for Computational Linguistics. 1

[42] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 8

[43] QING LYU, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey. *ArXiv*, abs/2209.11326, 2022. 1, 34

[44] Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55:1 – 42, 2021. 1, 34

[45] Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 258–271, Seattle, United States, 2022. Association for Computational Linguistics. 4, 35

[46] A. Modarressi, Hosein Mohebbi, and Mohammad Taher Pilehvar. Adapler: Speeding up inference by adaptive length reduction. In *Annual Meeting of the Association for Computational Linguistics*, 2022. 34

[47] Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. DecompX: Explaining transformers decisions by propagating token decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2649–2664, Toronto, Canada, 2023. Association for Computational Linguistics. 2, 3, 4, 14, 35

[48] Hosein Mohebbi, Jaap Jumelet, Michael Hanna, A. Alishahi, and Willem Zuidema. Transformer-specific interpretability. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2024. 1

[49] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recogn.*, 65(C):211–222, 2017. 34

[50] Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana, 2018. Association for Computational Linguistics. 3, 14

[51] Fahimeh Hosseini Noohdani, Parsa Hosseini, Arian Yazdan Parast, Hamidreza Yaghoubi Araghi, and Mahdieh Soleymani Baghshah. Decompose-and-compose: A compositional approach to mitigating spurious correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 34

[52] Paul Novello, Thomas Fel, and David Vigouroux. Making sense of dependence: Efficient black-box explanations using dependence measure. *ArXiv*, abs/2206.06219, 2022. 36

[53] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1

[54] Roni Paiss, Hila Chefer, and Lior Wolf. No token left behind: Explainability-aided image classification and generation. In *Computer Vision – ECCV 2022*, pages 334–350, Cham, 2022. Springer Nature Switzerland. 34

[55] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *ArXiv*, abs/1806.07421, 2018. 36

[56] Yao Qiang, Deng Pan, Chengyin Li, Xin Li, Rhongho Jang, and Dongxiao Zhu. AttCAT: Explaining transformers via attentive class activation tokens. In *Advances in Neural Information Processing Systems*, 2022. 3, 35

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[58] Pranav Rajpurkar, Jeremy A. Irvin, Aarti Bagul, Daisy Yi Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, C. Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and A. Ng. Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. *ArXiv*, abs/1712.06957, 2017. 4

[59] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. 36

[60] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2021. 1

[61] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109: 247–278, 2021. 1, 5, 34

[62] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 34

[63] Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2591–2600, 2019. 34

[64] Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, 2019. Association for Computational Linguistics. 2

[65] Wei Shi, Wentao Zhang, Weishi Zheng, and Ruixuan Wang. Pami: partition input and aggregate outputs for model interpretation. *ArXiv*, abs/2302.03318, 2023. 36

[66] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *ArXiv*, abs/1605.01713, 2016. 2, 34

[67] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, 2017. 2, 34

[68] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. 1, 34

[69] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014. 1, 34

[70] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Neural Information Processing Systems*, 2019. 3

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1

[72] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy, 2019. Association for Computational Linguistics. 34

[73] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr (Peter) Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 111–119, 2019. 36

[74] Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. Beyond explaining: Opportunities and challenges of xai-based model improvement. *Inf. Fusion*, 92: 154–176, 2022. 34

[75] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, 2019. Association for Computational Linguistics. 2

[76] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 4

[77] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1158–1167, 2020. 34

[78] Weiyan Xie, Xiao hui Li, Caleb Chen Cao, and Nevin L.Zhang. Vit-cx: Causal explanation of vision transformers. In *International Joint Conference on Artificial Intelligence*, 2022. 4, 36

[79] Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane ling Wang, and Michael I. Jordan. Ml-loo: Detecting adversarial examples with feature attribution. In *AAAI Conference on Artificial Intelligence*, 2019. 34

[80] Jianming Zhang, Zhe L. Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126: 1084–1102, 2016. 1, 34

[81] Jianping Zhang, Weibin Wu, Jen tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R. Lyu. Improving adversarial transferability via neuron attribution-based attacks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14973–14982, 2022. 34

Appendix

Table of Contents

A Further Experimental Details **14**
- A.1. Faithfulness Evaluation Metrics . . . 14
- A.2 True Token Masking . . . 14

B Is SkipPLUS's Choice of Skipping the First Half of the Network Optimal? **15**

C Composing PLUS and SkipPLUS With Other Methods **16**

D Qualitative Results **19**
- D.1. ViT Base (Patch Size 8) . . . 19
- D.2 EVA Large (Patch Size 14) . . . 24

E Related Work **34**
- E.1. Gradient-Based Methods . . . 34
  - Vanilla Gradients. . . . 34
  - Input×Gradient (IxG). . . . 34
  - E.1.1 CAM Methods . . . 34
    - GradCAM. . . . 34
    - AttCAM. . . . 34
    - XGradCAM. . . . 34
  - E.1.2 Gradient-Based Rollout Methods . . . 34
    - TransAtt. . . . 34
    - GenAtt. . . . 35
  - E.1.3 Special Cases of PLUS . . . 35
    - GradSAM. . . . 35
    - CAT. . . . 35
    - AttCAT. . . . 35
    - LayerCAM. . . . 35
- E.2. Forward Attention-Based Token Attribution Methods . . . 35
  - Attention×Input_Norm (AttIN). . . . 35
  - GlobEnc & ALTI. . . . 35
  - DecompX. . . . 35
- E.3. Black-Box Methods . . . 36
  - LIME . . . 36
  - RISE . . . 36
  - PAMI . . . 36
  - ScoreCAM . . . 36
  - ViT-CX . . . 36
  - AtMan . . . 36
  - HSIC . . . 36

A. Further Experimental Details

A.1. Faithfulness Evaluation Metrics

Modern literature favors evaluations for input attribution methods that are collectively called faithfulness, which intuitively measures how well the attribution scores reflect the true contribution of each input feature to the target output. Although several metrics have been proposed to quantify faithfulness, we adopt the most comprehensive approach, which involves computing the area under the curve (AUC) for the deletion and insertion operations, considering the changes in accuracy and the target probability [14, 27, 47, 50].

The deletion accuracy curve is obtained by progressively removing input features in order of decreasing attribution scores and measuring the model's accuracy at each step. A faithful attribution method should result in a steep drop in performance as the most important features are removed first. The deletion accuracy scores are normalized using the formula $100 - x$, where x is the original score, so that higher scores always indicate better performance.

Similarly, the deletion AOPC curve is generated by gradually removing input features in order of decreasing attribution scores and evaluating the change in the target output probability at each step. A faithful attribution method should lead to a rapid decrease in the target probability as the most important features are removed first.

Using the input images $\mathbf{x}_i$, the perturbed input $\tilde{\mathbf{x}}_i^k$ is created by deleting k% of the most significant patches from $\mathbf{x}_i$. Subsequently, Area Over the Perturbation Curve [19, AOPC] evaluates the mean alteration in the predicted class probability across the entire validation dataset using the following formula:

$$\text{AOPC}(k) = \frac{1}{N} \sum_{i=1}^{N} p\left(\hat{y}|\mathbf{x}_i\right) - p\left(\hat{y}|\tilde{\mathbf{x}}_i^k\right).$$

Here, N represents the total number of instances, $\hat{y}$ stands for the predicted class, and $p(\hat{y}|\cdot)$ denotes the probability of the predicted class.

Conversely, the insertion accuracy curve is generated by gradually adding input features in order of decreasing attribution scores and evaluating the model's performance at each step. A faithful attribution method should lead to a rapid increase in performance as the most important features are added first.

Likewise, the insertion AOPC curve is obtained by progressively adding input features in order of decreasing attribution scores and measuring the change in the target output probability at each step. A faithful attribution method should result in a steep increase in the target probability as the most important features are added first. Similar to deletion accuracy, the insertion AOPC scores are also normalized using $100 - x$, ensuring that higher scores consistently represent better performance.

A.2. True Token Masking

Instead of simply overlaying a color mask, we choose to completely exclude the masked patches from the model's input [16]. At the same time, we preserve accurate positional encodings for the unmasked patches. We term this strategy *True Token Masking*. The conventional method of using the color black (or simply zeroing the tokens in text-based Transformers) for patch masking encounters several issues:

- If a patch is predominantly black, painting it black does not effectively eliminate its informational content. For instance, a black drawing on a white background would remain mostly unchanged.
- Patches might serve computational functions, such as acting as a scratchpad for the model's internal processes. Masking these with black does not prevent the model from using them for such purposes.
- Introducing a black mask can create artifacts in the image, potentially leading to out-of-distribution data, which affects the model's performance.

B. Is SkipPLUS's Choice of Skipping the First Half of the Network Optimal?

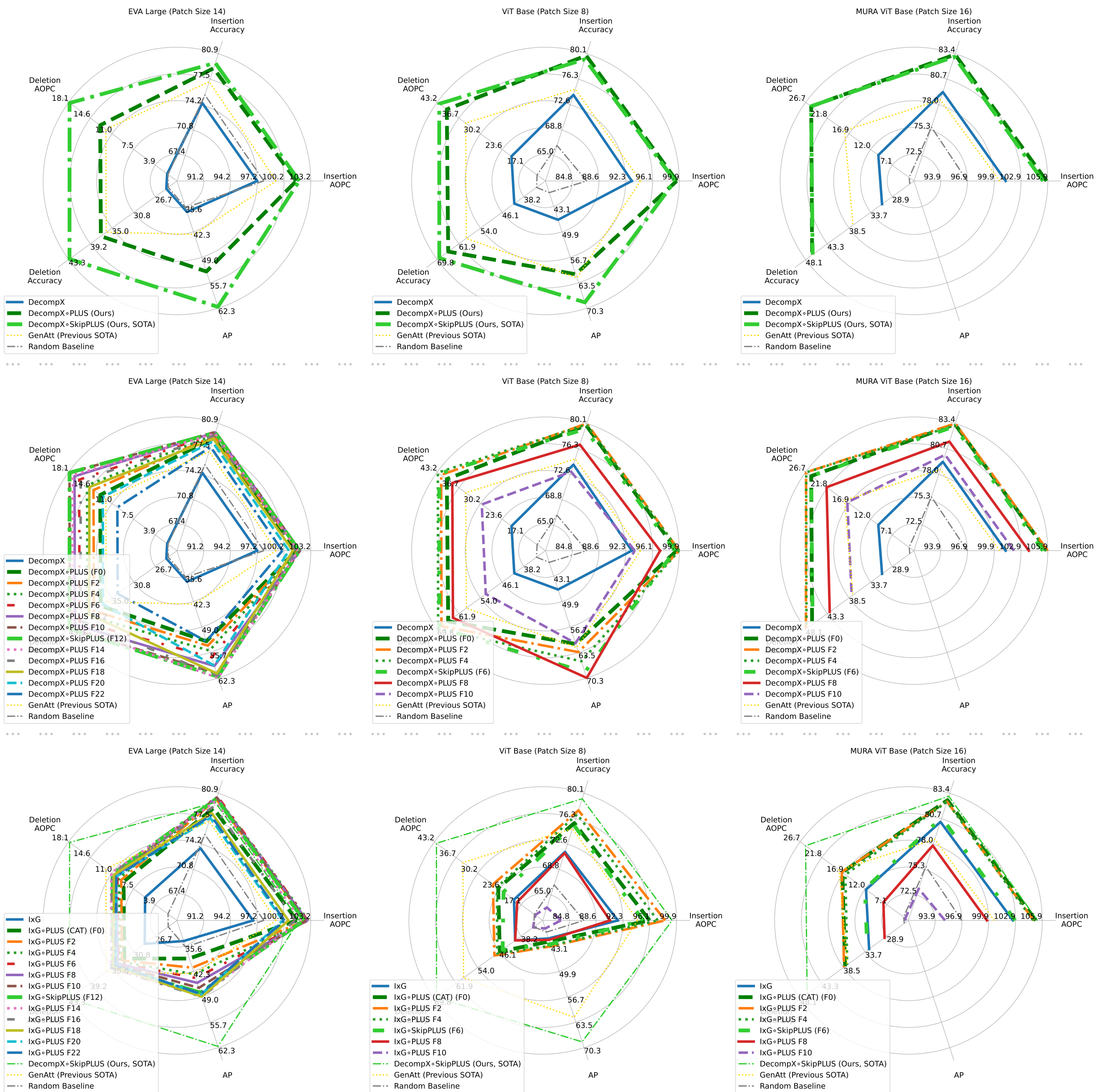

Figure 7. Either PLUS (starting the attribution method from all layers and aggregating information) or SkipPLUS (starting the atribution method from layers in the latter half of the network and aggregating information) are always near the optimal selection of the cutoff layer.

C. Composing PLUS and SkipPLUS With Other Methods

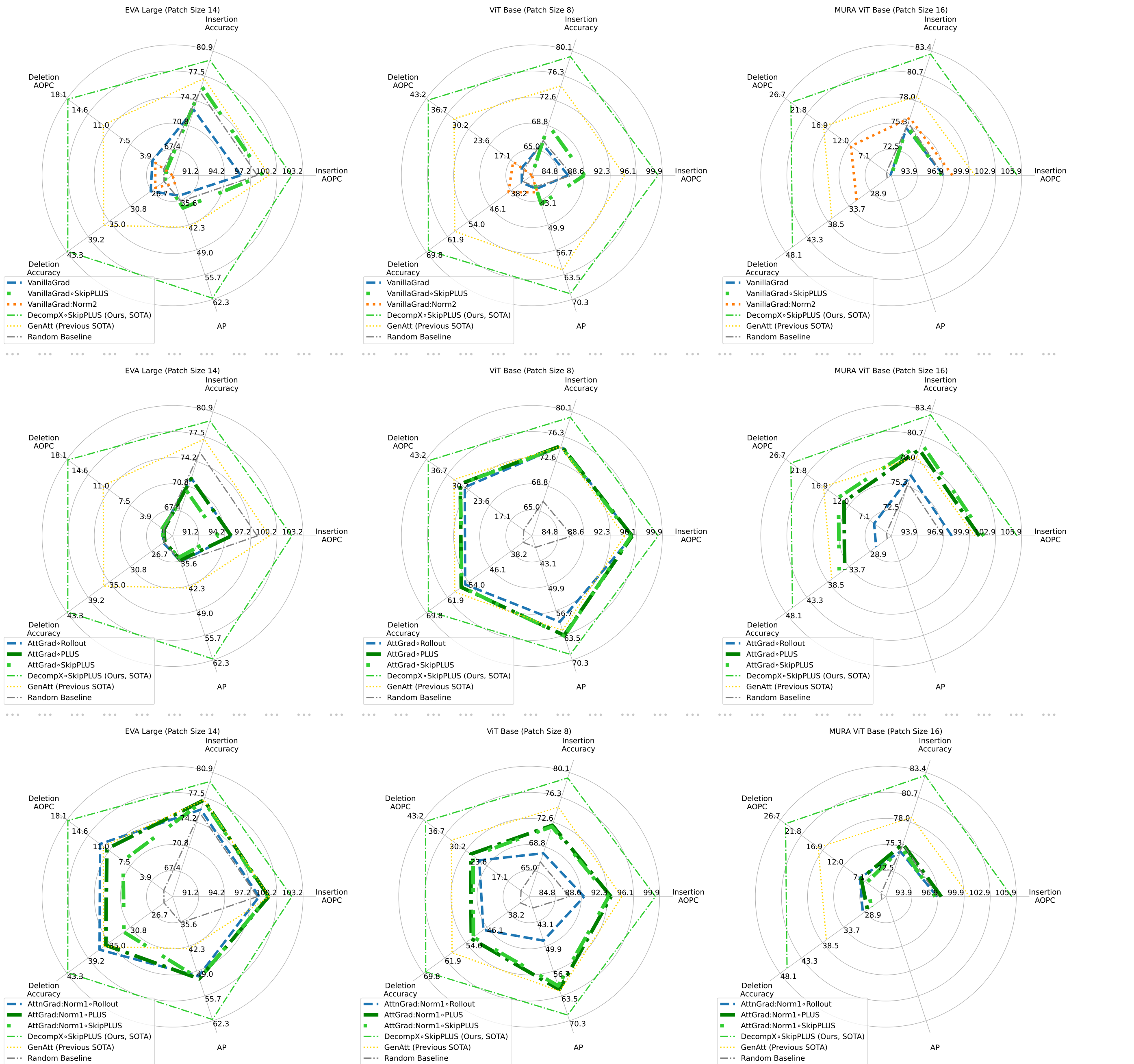

Figure 8. Evaluating the composition of PLUS and SkipPLUS with the Vanilla Gradients of tokens and attention weights.

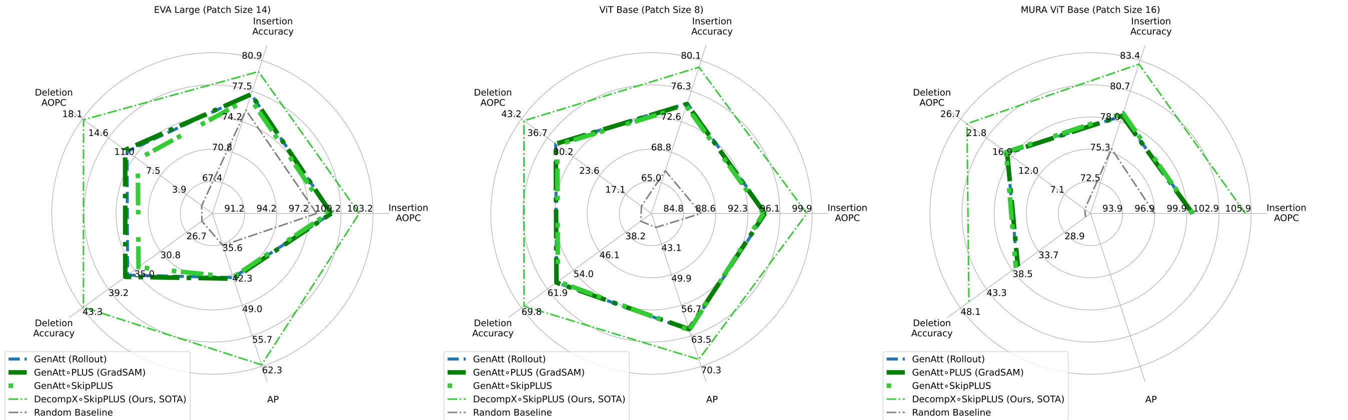

Figure 9. GenAtt, the previous state-of-the-art baseline, is not enhanced by our proposed PLUS or SkipPLUS methods, but it is also not harmed.

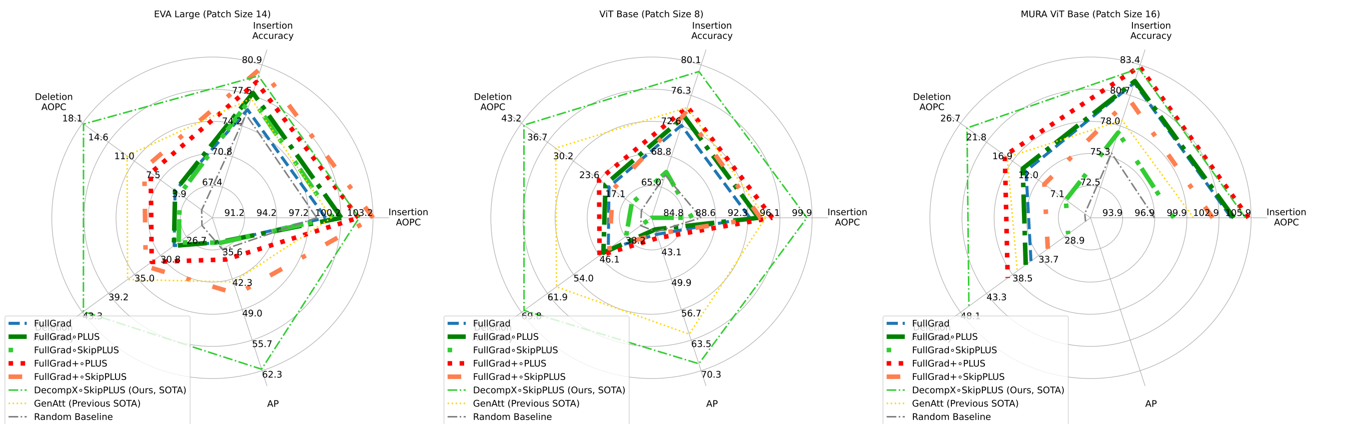

Figure 10. Evaluating the composition of PLUS and SkipPLUS with FullGrad.

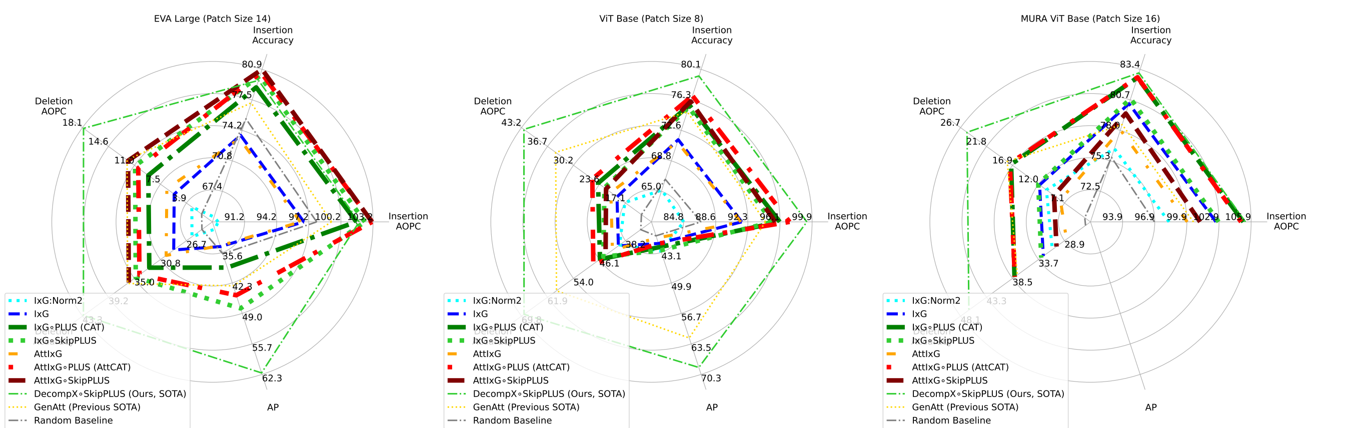

Figure 11. Evaluating the composition of PLUS and SkipPLUS with InputxGradient (IxG) and an attention-enhanced variant, AttIxG. The PLUS compositions of these methods are also known as CAT and AttCAT, respectively.

Figure 12. Evaluating the composition of PLUS and SkipPLUS with CAM methods from CNNs.

D. Qualitative Results

D.1. ViT Base (Patch Size 8)

Figure 13. Additional qualitative examples demonstrating the application of SkipPLUS on ViT Base (Patch Size 8). The images presented were selected at random.

Figure 14. Additional qualitative examples demonstrating the application of SkipPLUS on ViT Base (Patch Size 8). The images presented were selected at random.

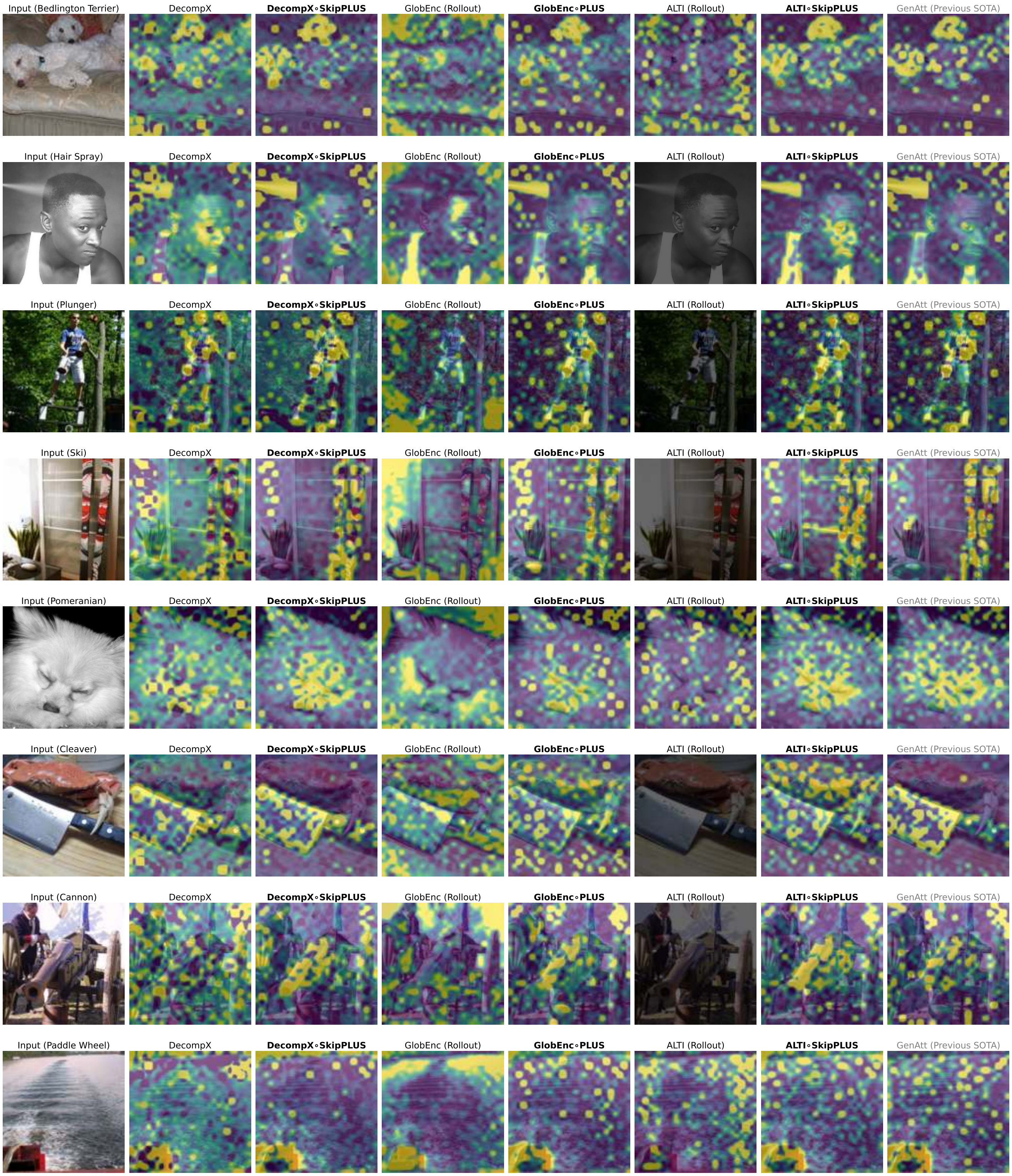

Figure 15. Additional qualitative examples demonstrating the application of SkipPLUS on ViT Base (Patch Size 8). The images presented were selected at random.

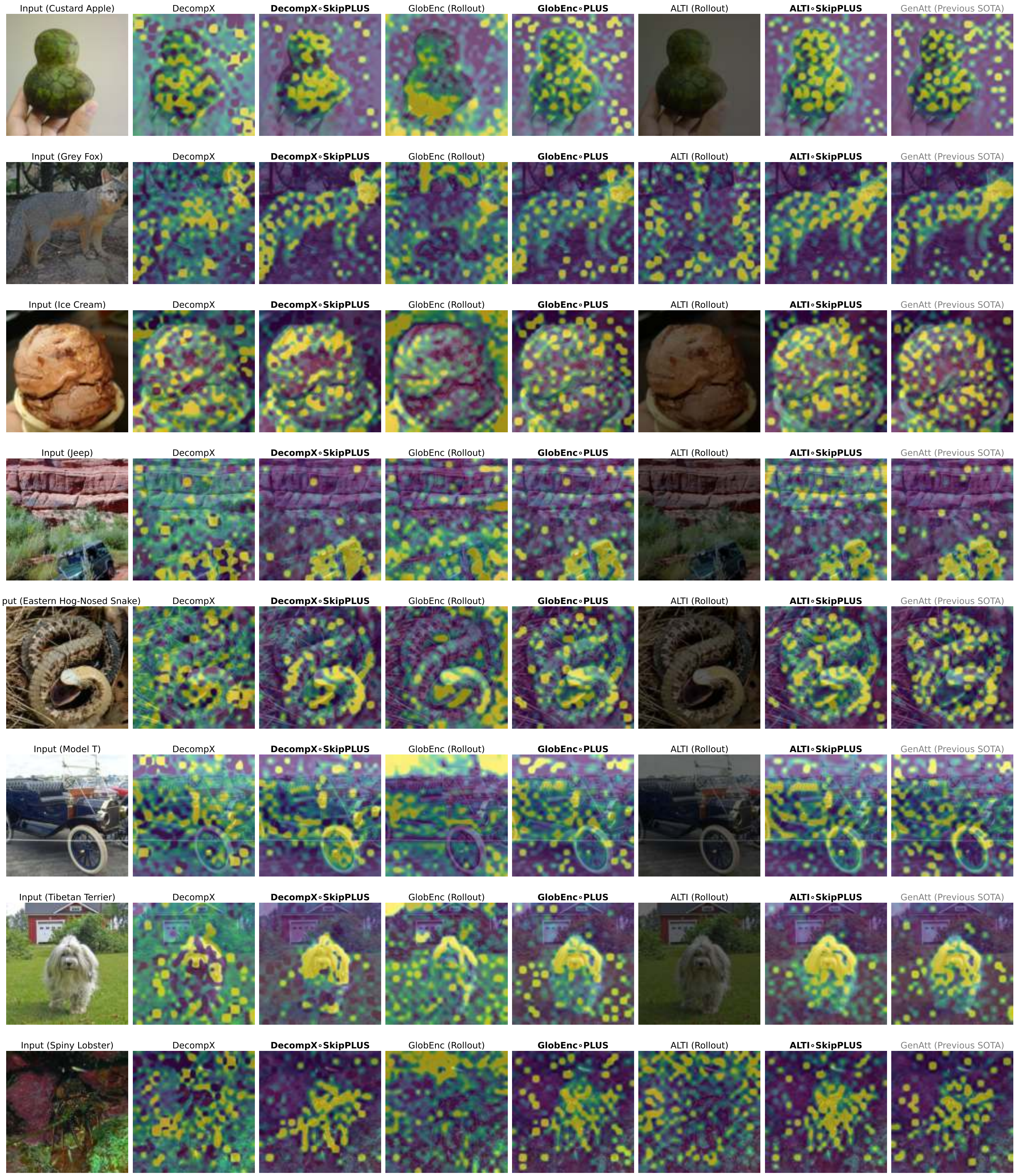

Figure 16. Additional qualitative examples demonstrating the application of SkipPLUS on ViT Base (Patch Size 8). The images presented were selected at random.

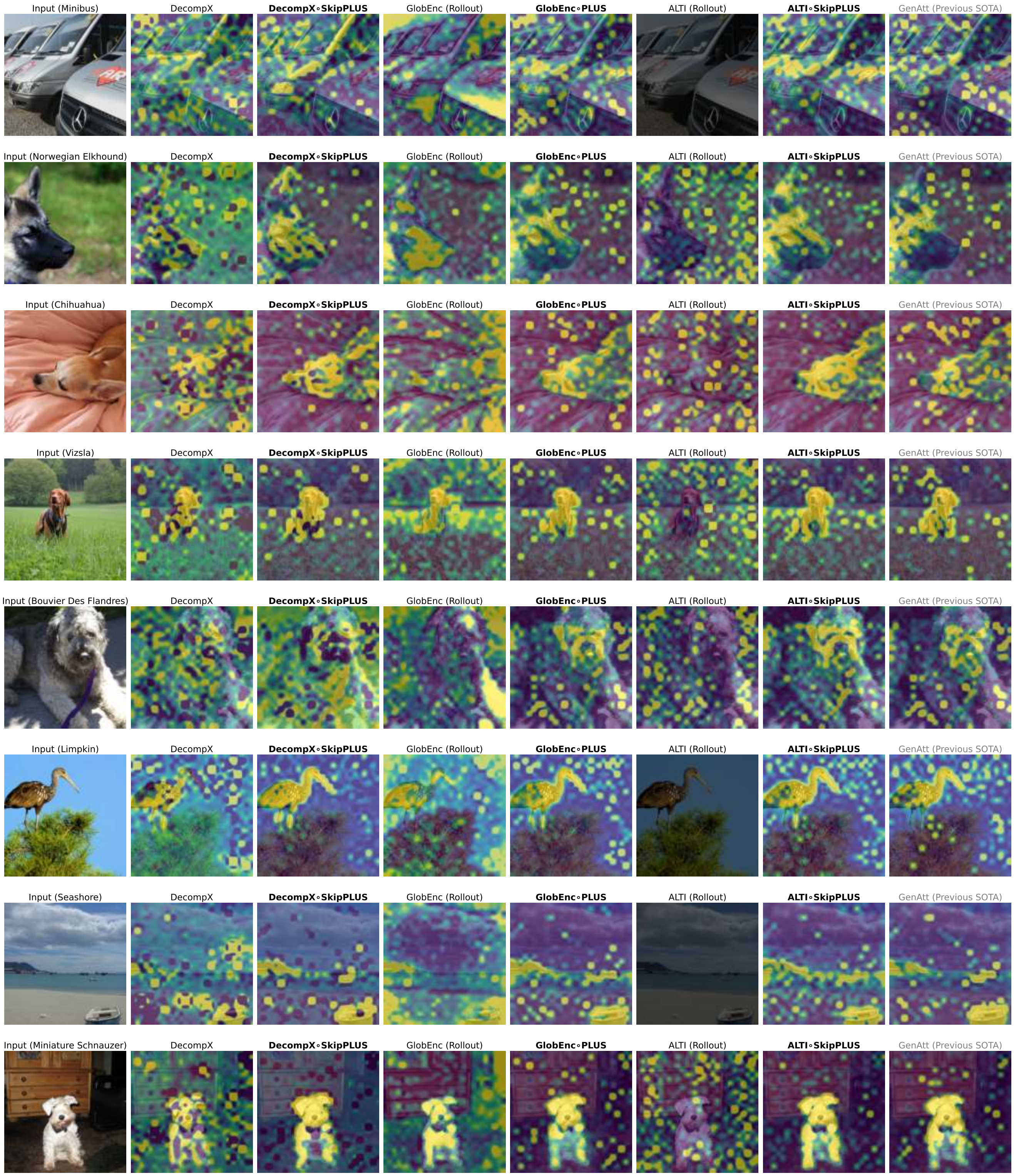

Figure 17. Additional qualitative examples demonstrating the application of SkipPLUS on ViT Base (Patch Size 8). The images presented were selected at random.

D.2. EVA Large (Patch Size 14)

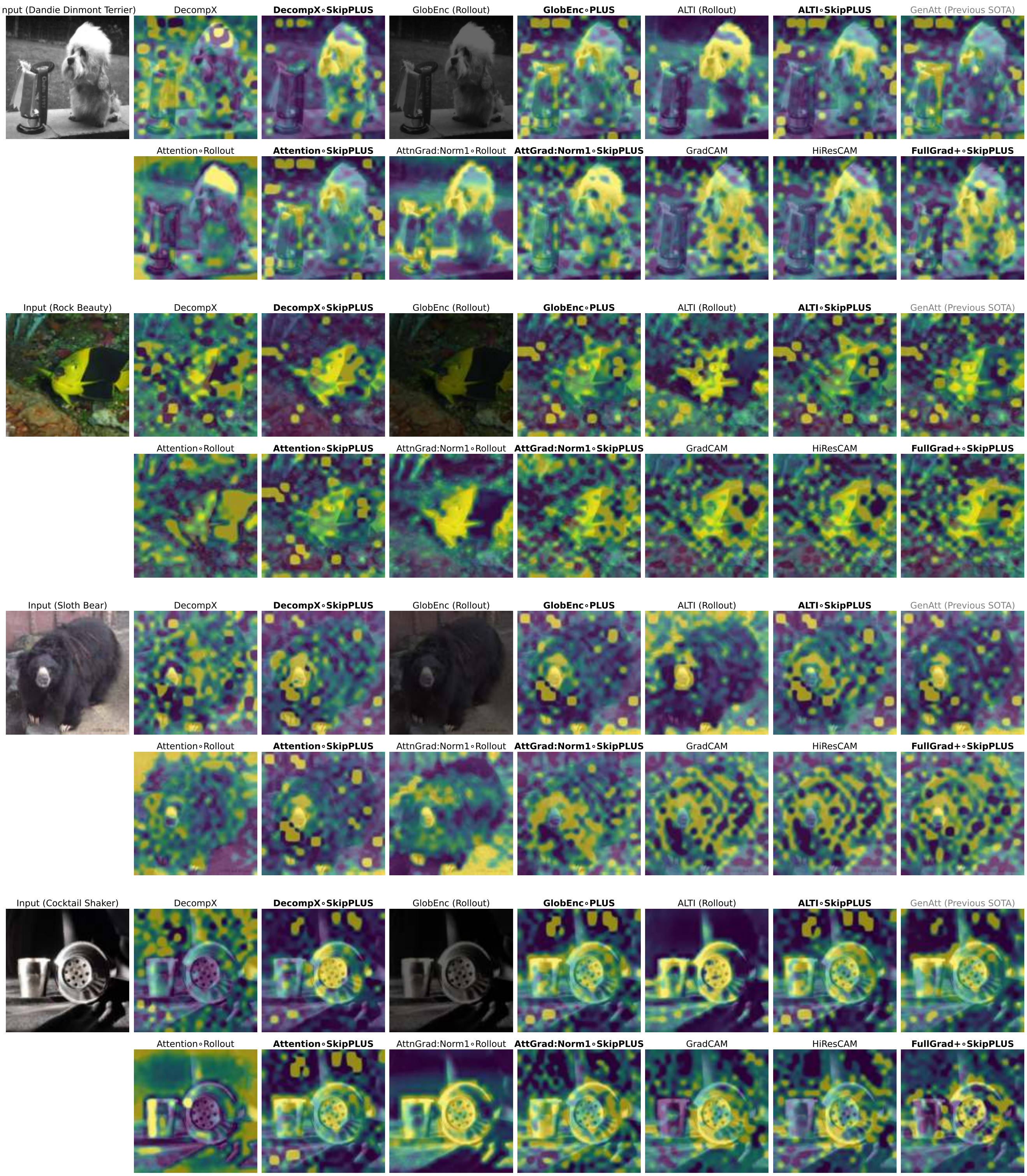

Figure 18. Additional qualitative examples demonstrating the application of SkipPLUS on EVA Large (Patch Size 14). The images presented were selected at random.

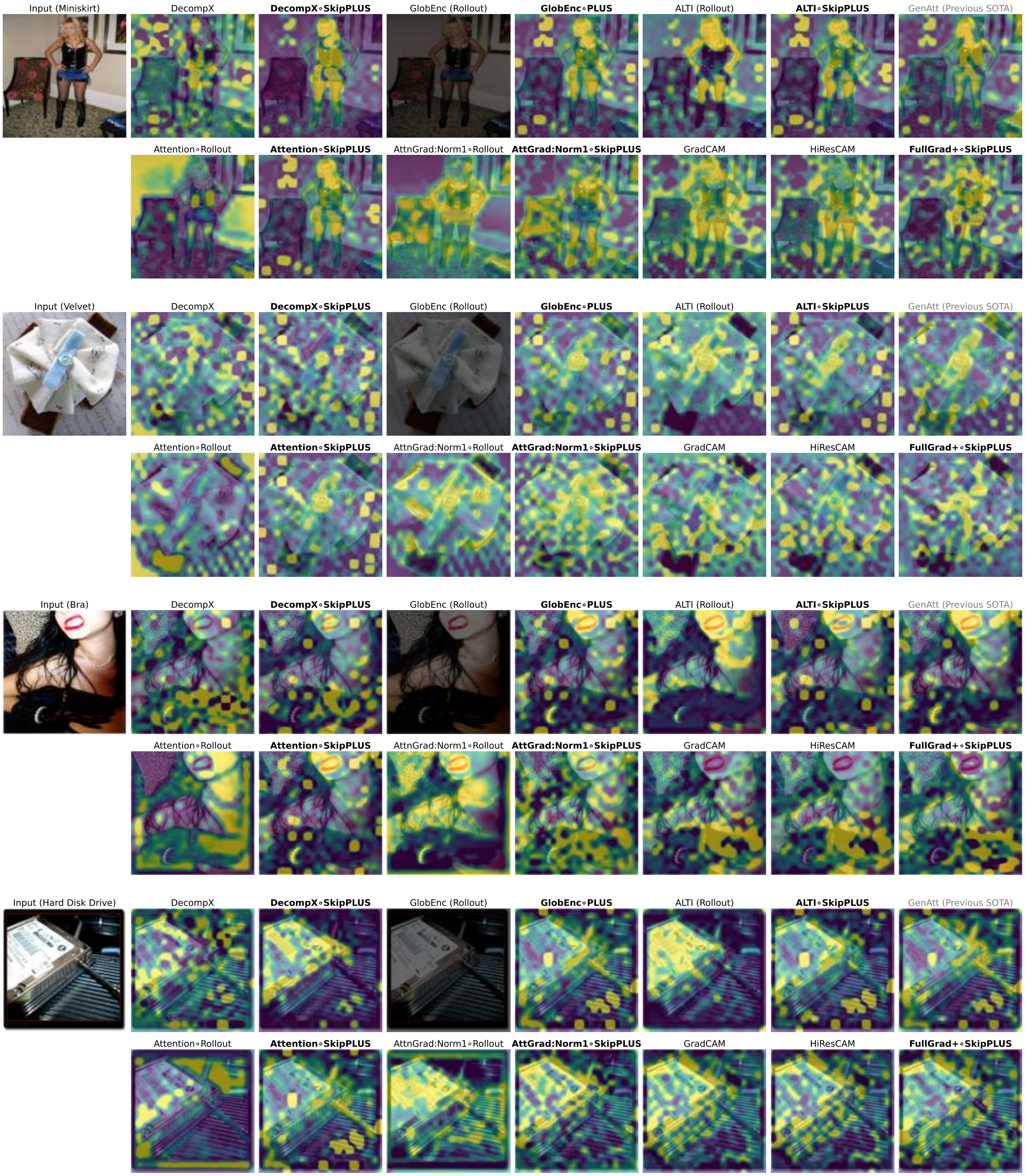

Figure 19. Additional qualitative examples demonstrating the application of SkipPLUS on EVA Large (Patch Size 14). The images presented were selected at random.

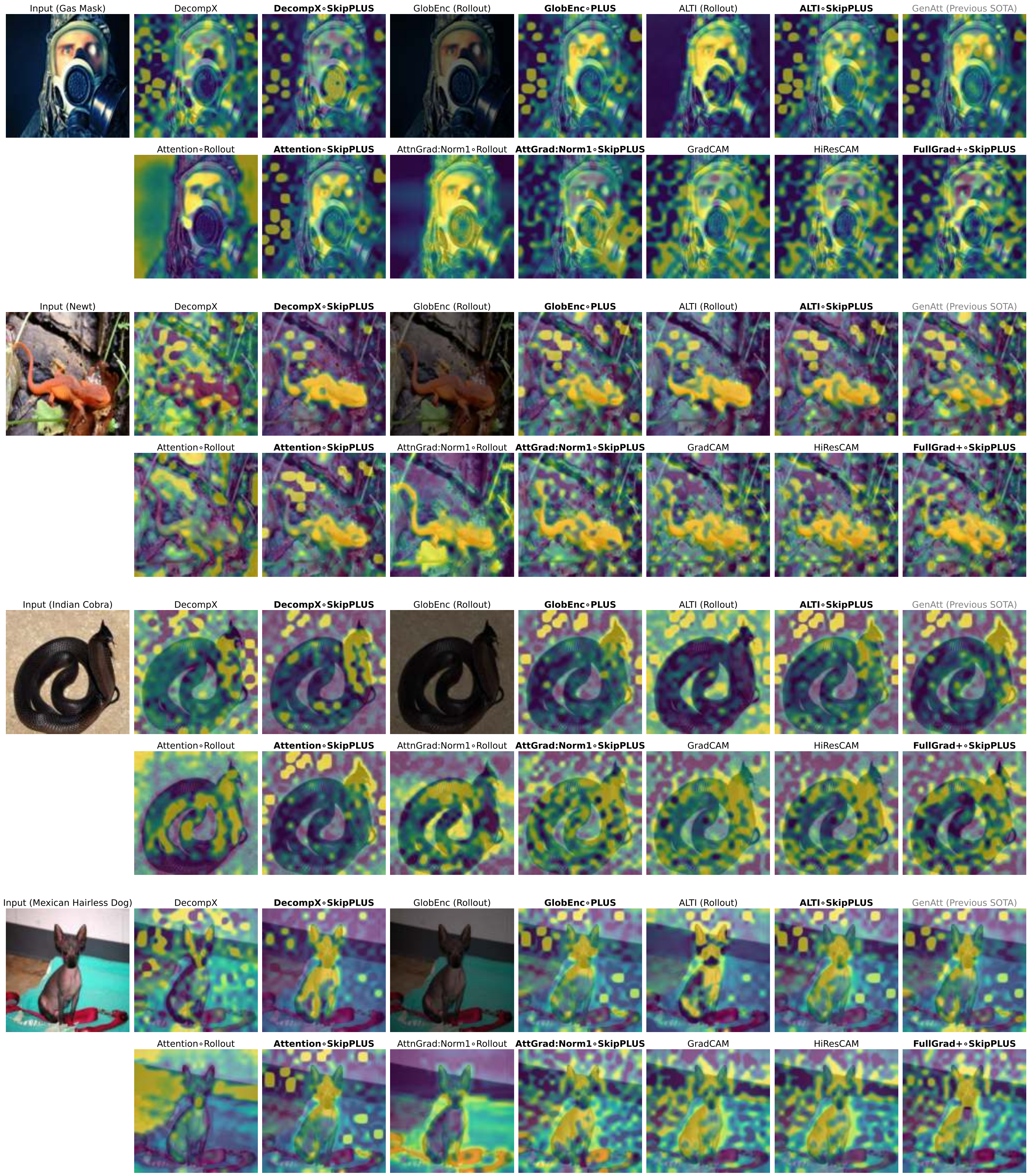

Figure 20. Additional qualitative examples demonstrating the application of SkipPLUS on EVA Large (Patch Size 14). The images presented were selected at random.

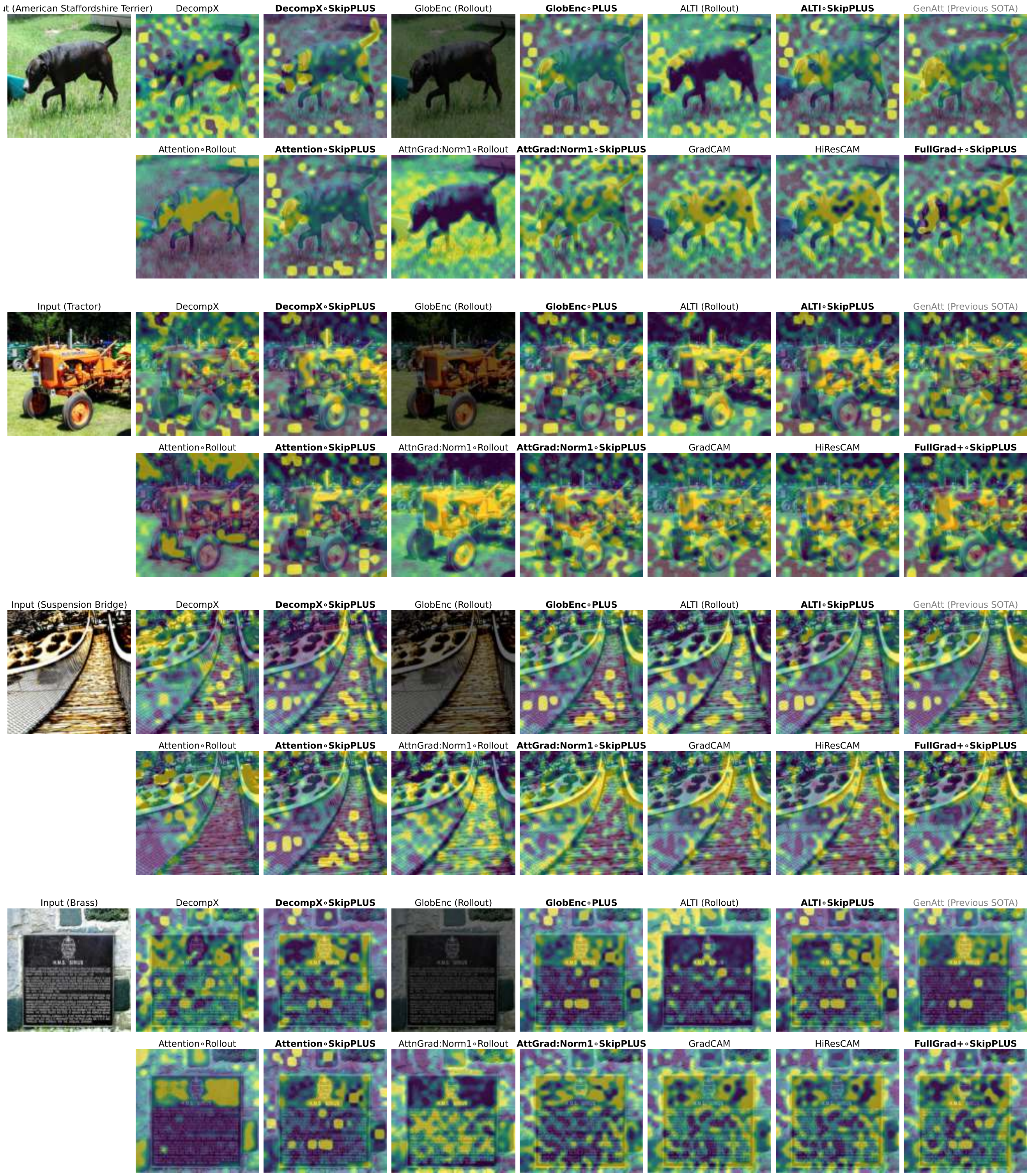

Figure 21. Additional qualitative examples demonstrating the application of SkipPLUS on EVA Large (Patch Size 14). The images presented were selected at random.

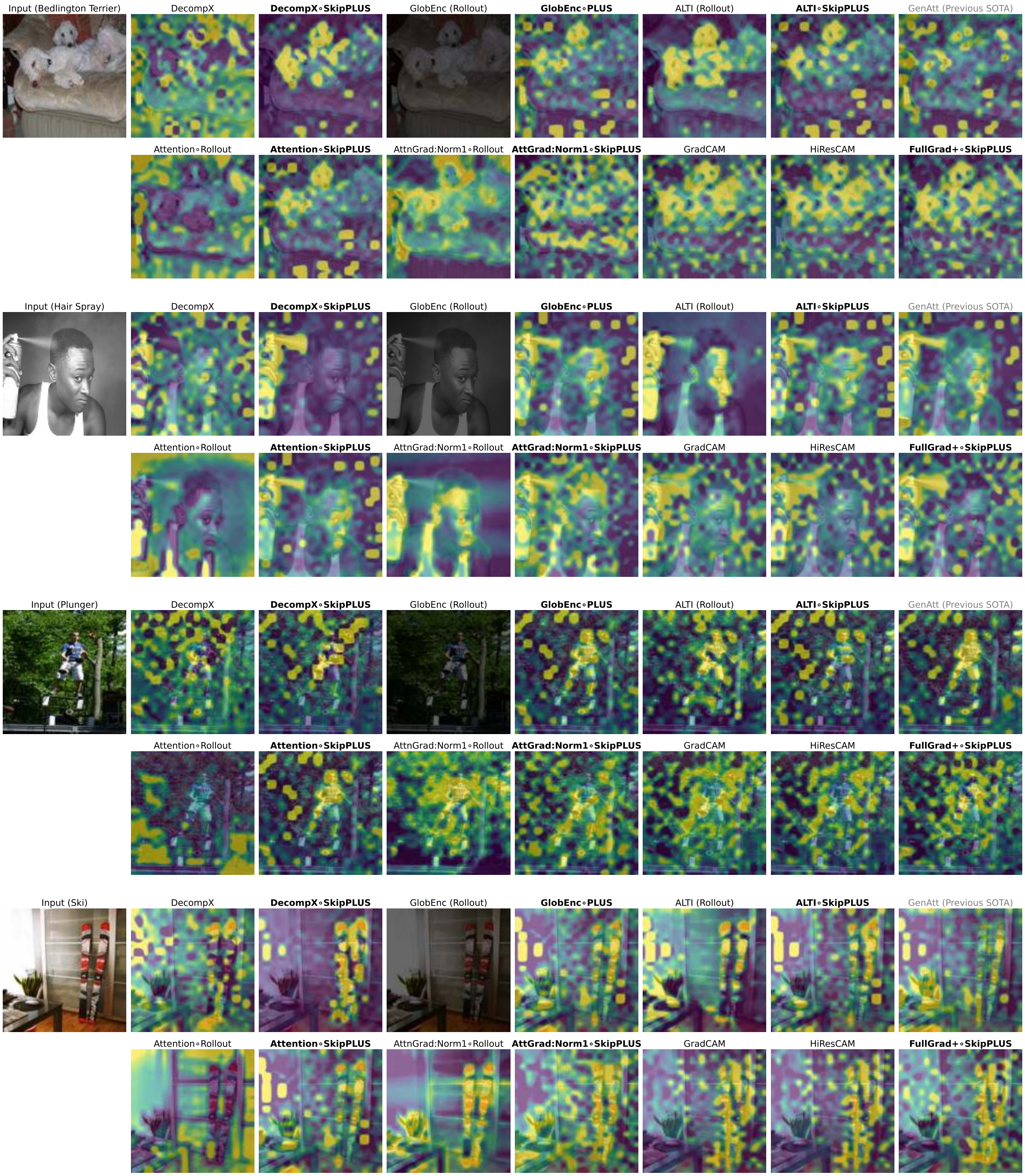

Figure 22. Additional qualitative examples demonstrating the application of SkipPLUS on EVA Large (Patch Size 14). The images presented were selected at random.

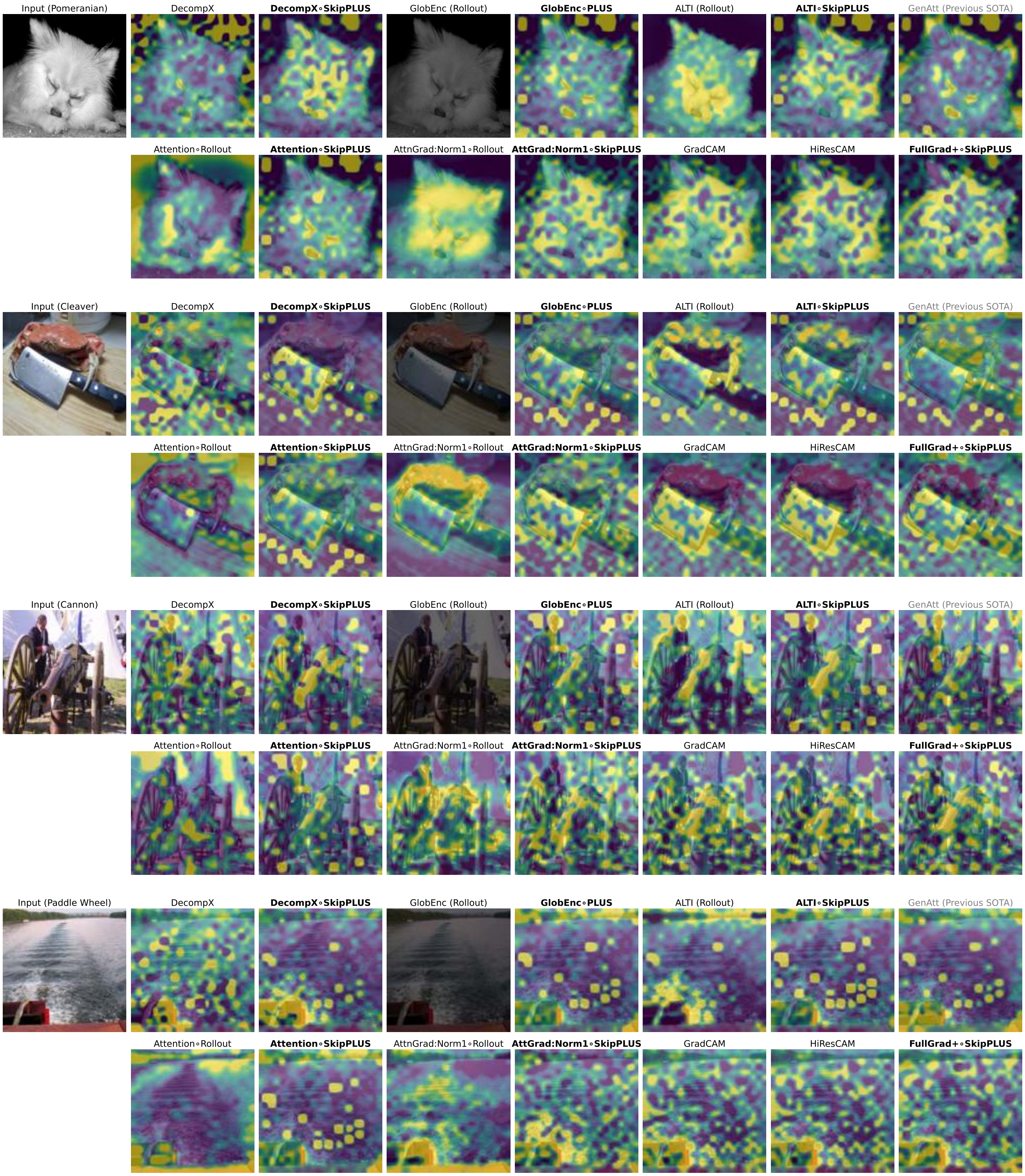

Figure 23. Additional qualitative examples demonstrating the application of SkipPLUS on EVA Large (Patch Size 14). The images presented were selected at random.

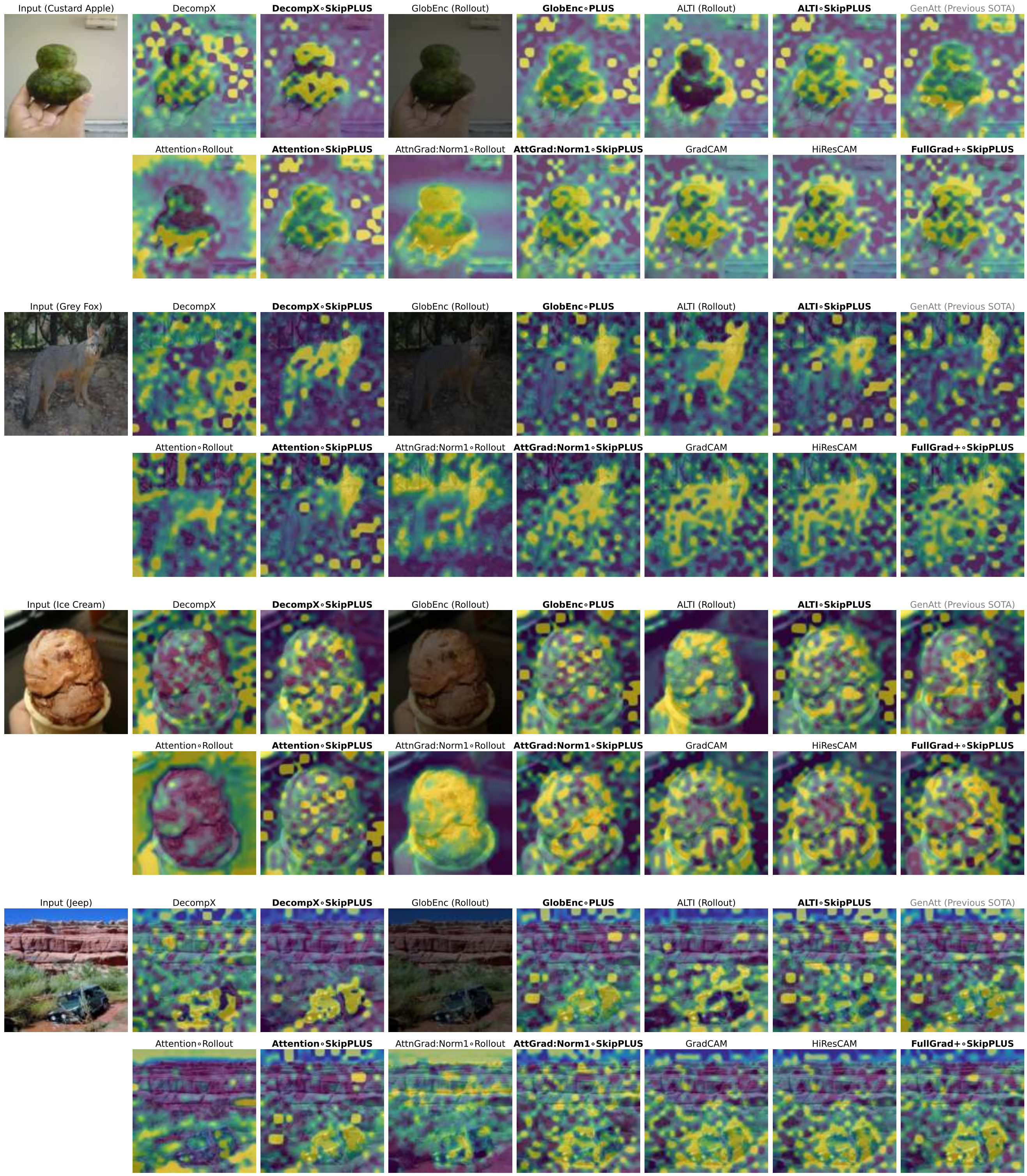

Figure 24. Additional qualitative examples demonstrating the application of SkipPLUS on EVA Large (Patch Size 14). The images presented were selected at random.

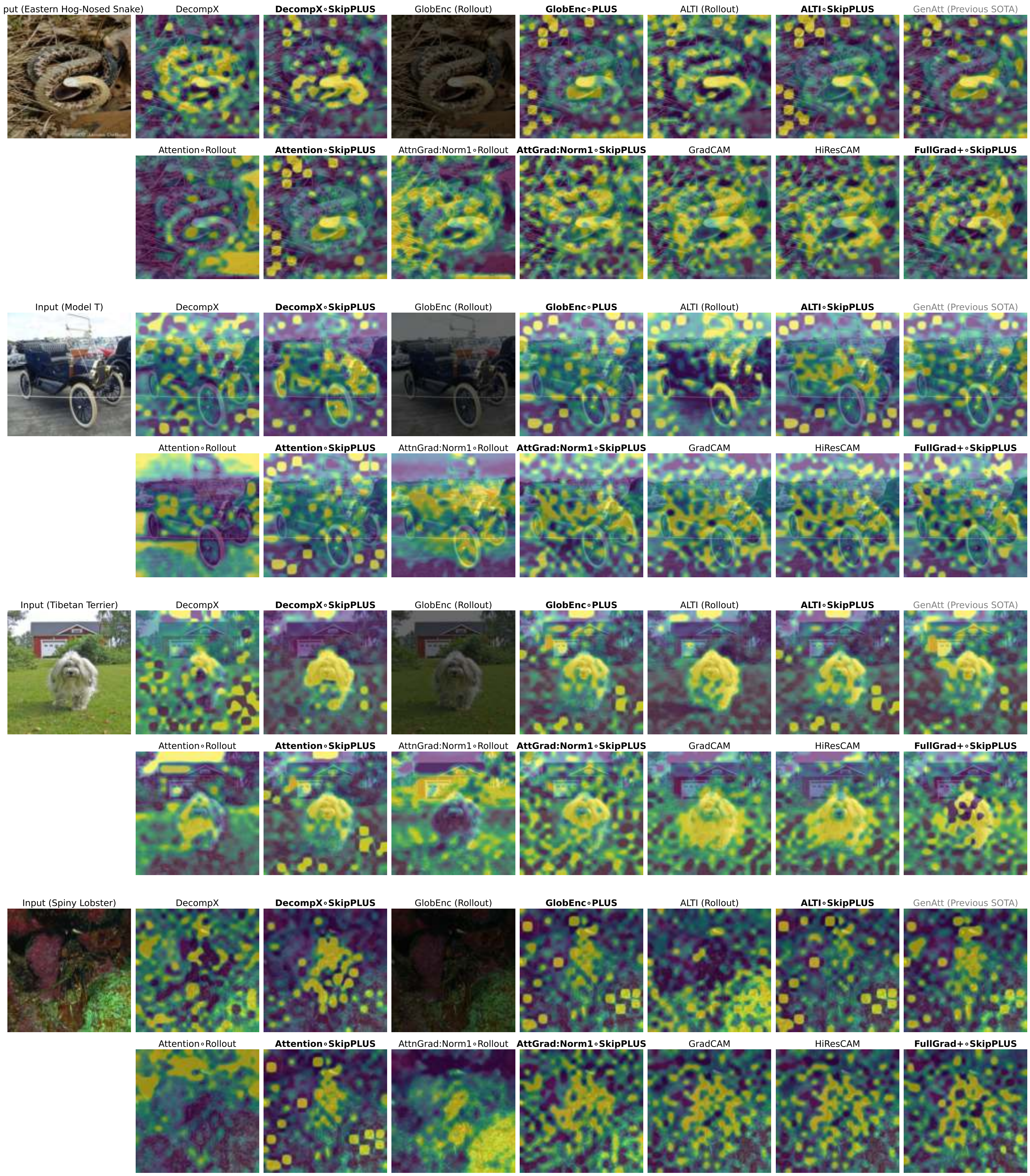

Figure 25. Additional qualitative examples demonstrating the application of SkipPLUS on EVA Large (Patch Size 14). The images presented were selected at random.

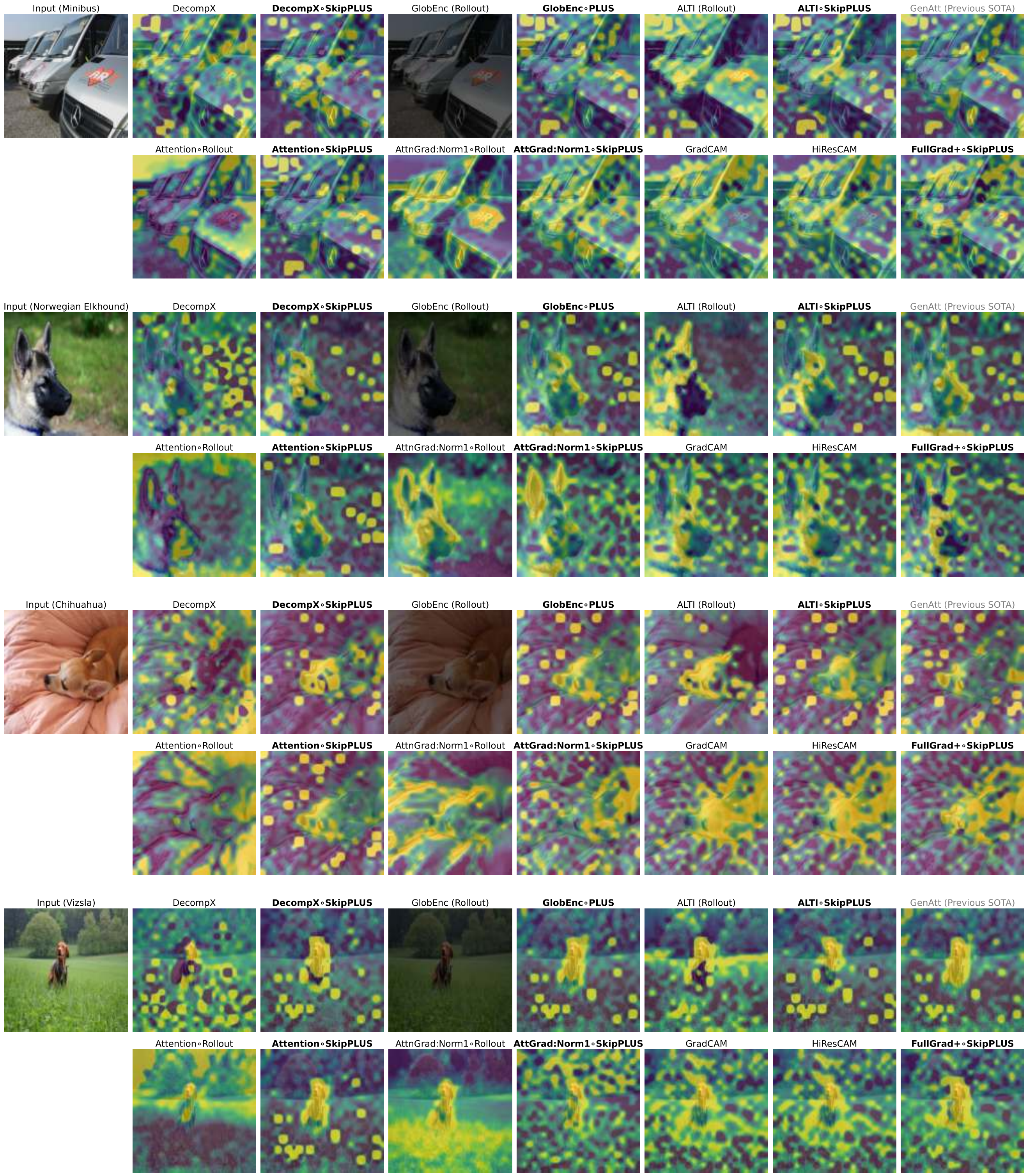

Figure 26. Additional qualitative examples demonstrating the application of SkipPLUS on EVA Large (Patch Size 14). The images presented were selected at random.

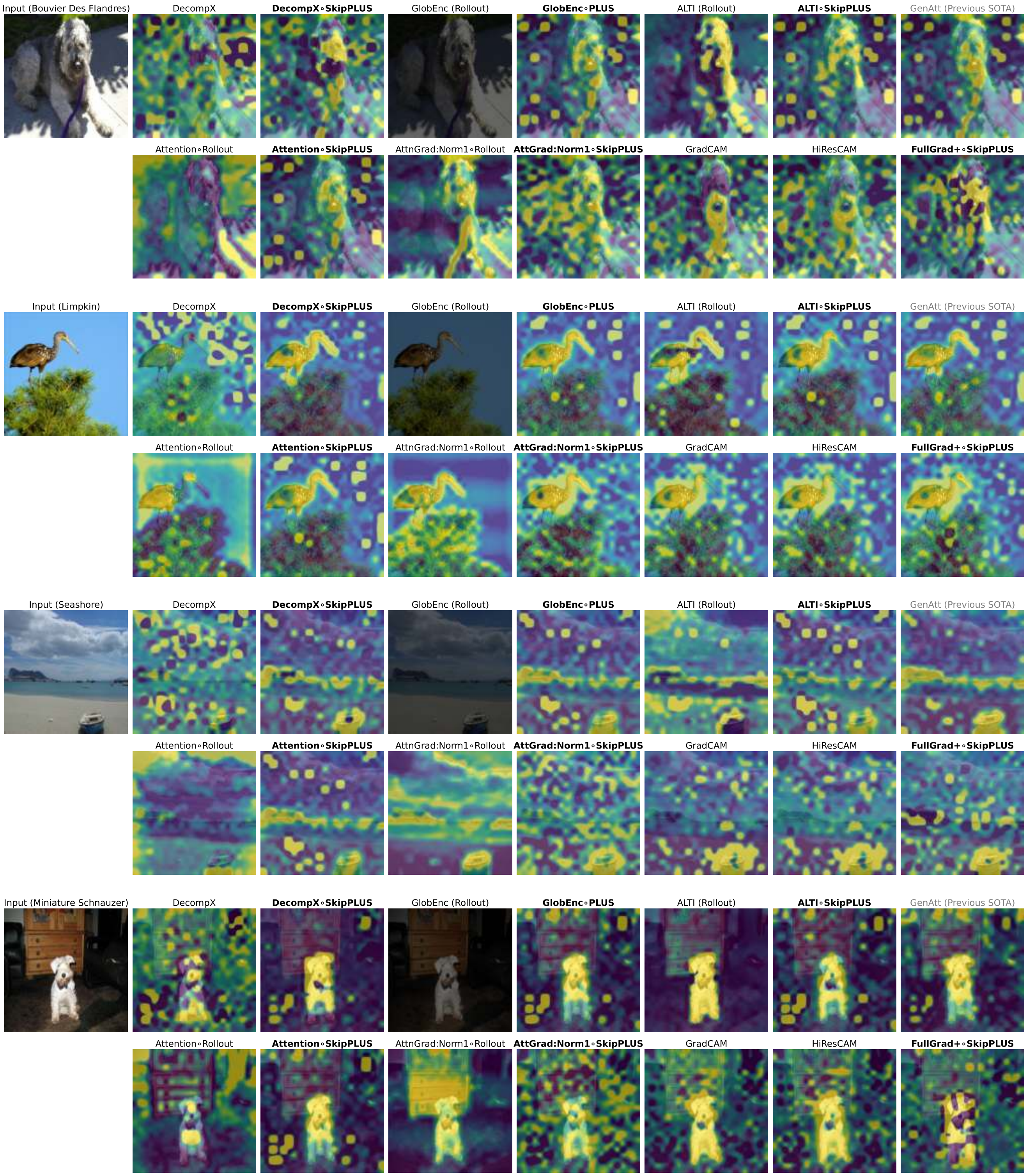

Figure 27. Additional qualitative examples demonstrating the application of SkipPLUS on EVA Large (Patch Size 14). The images presented were selected at random.

E. Related Work

Input attribution methods are techniques designed to quantify the influence of individual input features, or groups of them, on a model's output [6, 38, 43, 44, 61, 68, 69, 80]. Input attribution methods can assist in understanding a model's decision locally for a single input considered in isolation. They also act as foundational elements for more advanced explanation techniques. For instance, in concept-based explanation methods like CRAFT [25], attribution methods are employed for two main purposes: to quantify the impact of each activated concept and to identify the specific input features responsible for activating these concepts.

Attribution methods have a wide array of applications beyond merely explaining model outputs to humans [22, 63, 72, 74]. They are useful for enhancing the robustness of models against out-of-distribution data, spurious correlations, and adversarial inputs [3, 12, 51, 79]. Additionally, attribution methods have been employed to improve the performance of text-to-image models [13, 37, 54]. Furthermore, adapting forward-mode attribution methods has been explored for on-the-fly feature pruning [24, 46] and model quantization [5]. Attribution methods have been utilized to construct more effective adversarial attacks against models [33, 77, 81].

E.1. Gradient-Based Methods

Gradient-based methods compute the gradient of the model's output y_c w.r.t. the input features x_i which can be pixels, regions, or tokens of the whole input x. The general idea is that larger gradients indicate higher importance of the input features x_i on the prediction y_c.

Vanilla Gradients. The most straightforward approach to these methods is to use the gradient as the exact importance score [68].

$$\text{VanillaGrad}_i = \sum_j \frac{\partial y_c}{\partial x_{i,j}}.$$

$$\text{VanillaGrad:Norm2}_i = \left\| \frac{\partial y_c}{\partial x_i} \right\|_2.$$

Input×Gradient (IxG). IxG [2, 66, 67] multiplies the input values by their corresponding gradients. Let x_i be a spatial feature of the input, where $x_{i,j}$ represents the j-th channel of x_i. The Input×Gradient attribution for the spatial feature x_i with respect to the target class c is computed as follows, where y_c is the output of the model for the target class c:

$$\text{Input}\times\text{Gradient}_i = \sum_j \frac{\partial y_c}{\partial x_{i,j}} \cdot x_{i,j}, \qquad (7)$$

E.1.1 CAM Methods

CAM methods, popularized by GradCAM [62], usually start from the very last layer of the network. In this, their intuition has certain similarities with SkipPLUS; they both recognize the noisy nature of the first layers of the network.

GradCAM.

- A^k: the k-th channel of the feature map in the final layer
- c: the class w.r.t. which the attribution map is computed
- y^c: the class score (logit)
- Gradients are averaged over the width and height dimensions (indexed by i and j respectively) to obtain the neuron (channel) importance weights α_k^c:

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

AttCAM. Chefer et al. [11] introduce a Transformer-specific adaptation of GradCAM [62], and reuse the name GradCAM for it. We term this modified method AttCAM.

XGradCAM. It weights the gradients by their corresponding activation value when computing the spatial average [29]. XGradCAM was proposed on ReLU CNNs where the activations were always positive, hence they did not specify using the absolute value of the activations in the above computation, as is more intuitive. We name the variant with absolute activations XGradCAM+, and test both of them.

Other CAM methods include GradCAM++ [9], HiResCAM [21], and GradCAMElementWise [31].

E.1.2 Gradient-Based Rollout Methods

TransAtt. TransAtt [11] employs the Deep Taylor Decomposition technique [49] to attribute local relevance and subsequently propagates these relevance scores through the entire architecture of a Transformer model. This process effectively enables the backward propagation of information across all layers, starting from the output and extending back to the input. Additionally, this method incorporates gradients of attention weights. The method's functioning can be summarized as follows:

$$\textit{Rollout}\left(\mathbb{E}_{H:=\text{Heads}}\left[(\mathtt{R} \odot \mathtt{AttnGrad})^{+}\right]\right),$$

where R stands for the relevancy scores of attention weights. The Rollout technique is a method to aggregate the layer-wise attribution maps. We refer the reader to [1] for a detailed overview.

GenAtt. The dependence of TransAtt on specific rules for the propagation of relevance scores imposes limitations on its capacity to furnish explanations for various types of Transformer architectures. To cope with this issue, GenAtt [10] attempts to explain predictions for any Transformer-based architecture by using the attention weights in each block to update the relevancy maps, as demonstrated by the following expression:

$$\mathit{Rollout}\left(\mathbb{E}_{H:=\text{Heads}}\left[(\texttt{Attn} \odot \texttt{AttnGrad})^{+}\right]\right).$$

The notation $()^{+}$ denotes a filtering through the ReLU function. [10] show that GenAtt is at least as effective as TransAtt, if not better.

E.1.3 Special Cases of PLUS

The methods described below can be viewed as special cases of PLUS, where PLUS is composed with a previously existing method.

GradSAM. GradSAM [4] is equivalent to composing GenAtt [10] with PLUS, instead of using Rollout. (cf. Fig. 9 in the appendices)

CAT. Class Activation Tokens [56] is equivalent to IxG∘PLUS. (cf. Fig. 11 in the appendices)

AttCAT. We can define an attention-enhanced variant of IxG, AttIxG, by multiplying IxG with AttnFrom:

$$\texttt{AttnFrom}_j = \frac{1}{H \times N} \sum_{h=1}^{H:=\text{Heads}} \sum_{i=1}^{N:=\text{Tokens}} \text{RawAttn}_{h,i,j}$$

Note that attention weights have three dimensions: heads, to, from.

Attentive Class Activation Tokens [56, AttCAT] would then be equivalent to AttIxG∘PLUS. (cf. Fig. 11 in the appendices)

LayerCAM. LayerCAM [36] was introduced for ReLU CNN networks, where it is equivalent to applying a normalization process on the layer-wise attribution maps obtained from GradCAMElementWise [31], followed by the PLUS aggregation method. The normalization step is proposed because earlier layers tend to have smaller attribution maps compared to later layers. By normalizing the maps, LayerCAM ensures that each layer contributes more equally to the final attribution map. However, this approach is not suitable for ViTs, as we explicitly want to avoid giving earlier layers the same impact on the final attribution map as later layers (cf. Fig. 2, also supported by our preliminary quantitative evaluations).

E.2. Forward Attention-Based Token Attribution Methods

Although we have mathematically detailed most of the previous methods, the complexity of the subsequent approaches surpasses the scope of this paper. Therefore, we will provide a succinct overview of their core concepts. For a more thorough understanding, we recommend readers refer to the original papers.

Attention×Input_Norm (AttIN). Kobayashi et al. [39] multiply the attention weights by the norms of the vectors corresponding to each attention weight. Kobayashi et al. [40] extends AttIN to also incorporate the residual connections.

GlobEnc & ALTI. AttIN assumes that tokens retain their original identity. As each self-attention module mixes all the tokens, this assumption might not necessarily hold. Using gradient-based techniques, Brunner et al. [7] studies contextual information aggregation across the model. Following Brunner et al. [7] work, the global token attribution analysis method [45, GlobEnc] further extends AttIN by including the Transformer block's second normalization layer in its analysis. In parallel with GlobEnc, the Aggregation of Layer-Wise Token-to-Token Interactions method [27, ALTI] was introduced. ALTI shares core concepts with GlobEnc, but the two differ in certain mathematical specifics.

DecompX. DecompX [47] enhances GlobEnc by integrating the one element previously overlooked by GlobEnc: the MLP module in the Encoder Transformer layer. This inclusion enables DecompX to generate a set of decomposed vectors that collectively sum up to the actual output vector. Unlike GlobEnc and ALTI, which require computing and aggregating layer-wise attribution maps using techniques like Rollout, DecompX facilitates the direct propagation of these decomposed vectors across layers. This capability allows for the direct computation of attribution maps from any layer to any other layer.

In this paper, unless specified otherwise, we utilize the DecompX variation that omits biases, referred to as DecompX W/O Bias in [47]. Our decision stems from our preliminary tests where no significant differences were observed across various methods of handling biases. To sidestep the complexities and hyperparameters introduced by distributing bias attributions among tokens, we opted for the simplest approach. A detailed evaluation of different methods for bias attribution distribution is reserved for future research.

E.3. Black-Box Methods

Black-box attribution methods treat the model as an opaque entity, (partially) disregarding its internal structure and gradients. These methods typically involve perturbing the input and observing the corresponding changes in the model's output to infer the importance of each input feature. However, this approach often comes with significant computational costs due to the need for multiple model evaluations. In contrast, white-box methods leverage the internal structure and gradients of the model, providing a more efficient and fine-grained understanding of the model's behavior.

In this paper, we focus on white-box methods for several reasons. Firstly, they offer a more computationally efficient approach compared to black-box methods. Secondly, and more importantly, black-box methods can be seen as directly optimizing the faithfulness metrics on which we evaluate the attribution methods. This raises concerns related to Goodhart's law, which states that when a measure becomes a target, it ceases to be a good measure. In other words, the faithfulness metrics we use are merely proxies for the ultimate desirable properties we seek in attribution methods. By directly optimizing these metrics, black-box methods may inadvertently introduce biases or artifacts that undermine the true faithfulness of the attributions. Therefore, to avoid this potential pitfall and maintain a more objective evaluation, we refrain from including comparisons with black-box methods in this study, acknowledging that they have different trade-offs and use cases.

LIME [59] explains the predictions of any classifier by learning a local interpretable model around the prediction.

RISE [55] is a black-box approach that generates an importance map indicating the saliency of each pixel for the model's prediction by probing the model with randomly masked versions of the input image and obtaining the corresponding outputs.

PAMI [65] masks the majority of the input and uses the corresponding model output as the relative contribution of the preserved input part to the original model prediction.

ScoreCAM [73] is a post-hoc visual explanation method based on class activation mapping that eliminates the dependence on gradients by obtaining the weight of each activation map through its forward passing score on the target class.

ViT-CX [78] adapts ScoreCAM for ViTs.

AtMan [17] is a perturbation method that manipulates the attention mechanisms of transformers to produce relevance maps for the input with respect to the output prediction.

HSIC [52] is a black-box attribution method based on the Hilbert-Schmidt Independence Criterion, measuring the dependence between regions of an input image and the model's output using kernel embeddings of distributions.