Supplementary Materials for the XAI Benchmark $\mathcal{M}^4$

1 Detailed Results

Numerical Results. In Table 1, we present numerical results demonstrating different feature attribution approach for image classification tasks. For future work, we aim to expand our analyses by incorporating additional methods and models, and even different data modalities such as graphs, audio, and video. There are several perspectives from which to analyze these benchmark results, making comprehensive exploration challenging. Advanced statistical techniques may yield further insights in future work.

Separate Results of Metrics. In Section 3.2 of the main text, we showed the averaged metric scores with respect to models and attribution methods for image classification task. Here we present the detailed scores of each metric for images in Figure 1 and texts in Figure 2.

Finetuning Details of NLP Models on MovieReview. Language models are all pretrained on large text corpus and yet not specified on downstream tasks. It is thus required to do the finetuning step on MovieReview to adapt the pretrained language model on the sentiment analysis task for better performance. The finetuning details are described as follows:

- Batch Size: 32
- Learning Rate: 3e-5
- Optimizer: AdamW
- Epochs: 10
- Max Sequence Length: 300 for Bert/L, 512 for others.

Training Details on Synthetic Dataset. We briefly recall the background. The idea with the synthetic dataset is that the synthetic patches on the images serve as supervision signals for training models. Labels of the images with synthetic patches will be reversed. These patches constitute explanation ground truths because no other patterns can lead to correct predictions by design. Therefore, models trained on such datasets must attribute their predictions to the synthetic ground truths. A subset of ImageNet training set with 10K images is used for training, and a subset of ImageNet validation set with 5K images is used for evaluation. During training, yellow patches of fixed size 60×60 and random positions in the image are added with probability of 50% to training images. Images with yellow patches are labeled as 1, others are labeled as 0. Without much tuning, we apply almost the same setting for training each model, with only changing the learning rate. Specifically, the hyperparameter details are as follows:

- Batch Size: 64
- Optimizer: AdamW
- Epochs: 3
- Learning Rate: 0.1 for MobileNet, 0.01 for ResNet family, 0.001 for ViTs and VGG16.

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

Table 1: Detailed Results for all pairs of models and attribution methods.

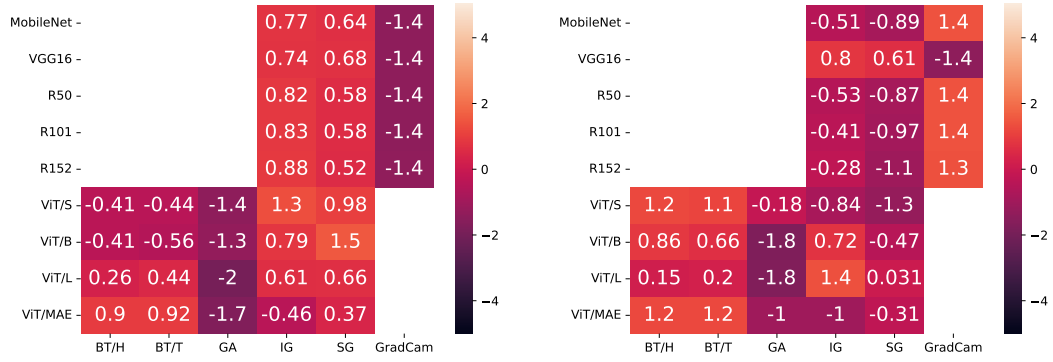
model	it	MoRF	ABPC	Pscore	INFD	SynScore
R50	GradCAM	0.6283	0.4243	0.8354	2.4959	0.9997
R50	IG	0.7093	0.3772	0.8117	2.3734	0.9990
R50	SG	0.7005	0.3687	0.8204	2.3234	0.9983
R101	GradCAM	0.6488	0.4471	0.8381	3.1395	0.9692
R101	IG	0.7291	0.3920	0.8056	3.0128	0.9954
R101	SG	0.7202	0.3746	0.8188	2.9634	0.9651
R152	GradCAM	0.6481	0.4606	0.8380	2.9718	0.9872
R152	IG	0.7302	0.4042	0.8005	2.8786	0.9983
R152	SG	0.7170	0.3767	0.8046	2.8487	0.9848
MobileNet	GradCAM	0.5892	0.3335	0.7828	1.4315	0.9923
MobileNet	IG	0.7214	0.2993	0.7727	1.2611	0.9843
MobileNet	SG	0.7134	0.2924	0.7988	1.2277	0.9598
VGG16	GradCAM	0.5263	0.1942	0.5586	208.3502	0.9777
VGG16	IG	0.8610	0.3731	0.7871	58.2901	0.9935
VGG16	SG	0.8515	0.3579	0.8151	58.8416	0.9235
ViT/B	BT/H	0.6789	0.3643	0.8745	1.0845	0.9238
ViT/B	BT/T	0.6770	0.3610	0.8687	1.0971	0.9247
ViT/B	GA	0.6672	0.3203	0.7242	1.0539	0.9631
ViT/B	IG	0.6946	0.3620	0.7740	1.1642	0.8480
ViT/B	SG	0.7038	0.3419	0.7894	1.1190	0.8537
ViT/L	BT/H	0.6863	0.3299	0.8233	1.2060	0.8622
ViT/L	BT/T	0.6884	0.3318	0.8593	1.2098	0.8535
ViT/L	GA	0.6614	0.2596	0.6829	1.2542	0.9199
ViT/L	IG	0.6902	0.3748	0.7425	1.2421	0.8792
ViT/L	SG	0.6908	0.3254	0.7443	1.2055	0.8509
ViT/S	BT/H	0.6048	0.4115	0.8855	0.4338	0.8658
ViT/S	BT/T	0.6045	0.4095	0.8790	0.4391	0.9147
ViT/S	GA	0.5907	0.3571	0.7567	0.4360	0.9942
ViT/S	IG	0.6291	0.3305	0.7824	0.4650	0.8059
ViT/S	SG	0.6246	0.3131	0.8183	0.4491	0.7160
ViT/MAE	BT/H	0.5865	0.3683	0.9363	0.1570	0.9821
ViT/MAE	BT/T	0.5869	0.3688	0.9340	0.1574	0.9756
ViT/MAE	GA	0.5335	0.2761	0.8884	0.1788	0.7088
ViT/MAE	IG	0.5590	0.2763	0.7543	0.1624	0.8361
ViT/MAE	SG	0.5758	0.3063	0.8028	0.1581	0.7689

2 Gradient-based Attribution Methods on Vision Transformers

As discussed in the main text, gradient-based explanations can be noisy for Vision Transformers [1]. Here we also show the visualization results of IG and SG. The results show the clear boundaries of patches. This may be caused by back-propagation through the patch embedding layer.

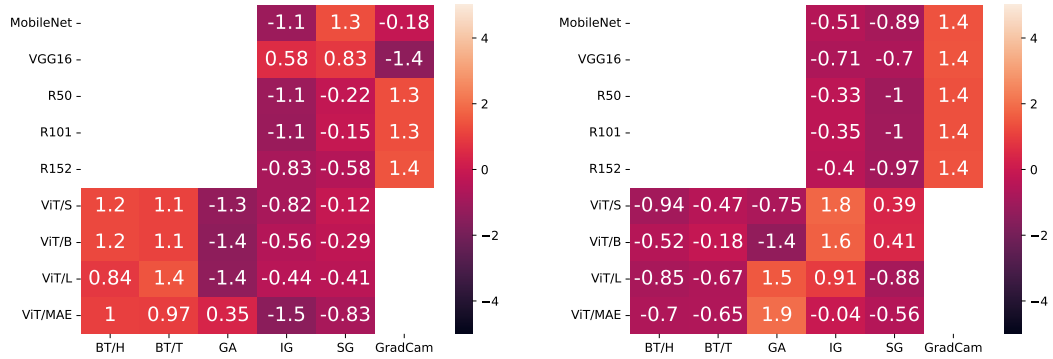
3 Computation Cost

We also report the time cost for performing the computation of explanations. See Table 2 for the results. Note that the implementation for each method is not optimized. Moreover, for modular designs, the core calculation of the attribution algorithm is done with CPU parallelization. Nevertheless, we can still notice that the GA and GradCAM are performed very fast, finishing 100 examples in several seconds, while others can be completed in several to twenty minutes.



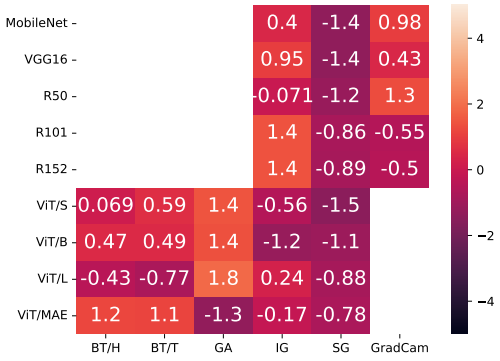
(a) Metric of MoRF.

(b) Metric of ABPC.



(c) Metric of PScore.

(d) Metric of INFD.



(e) Metric of SynScore.

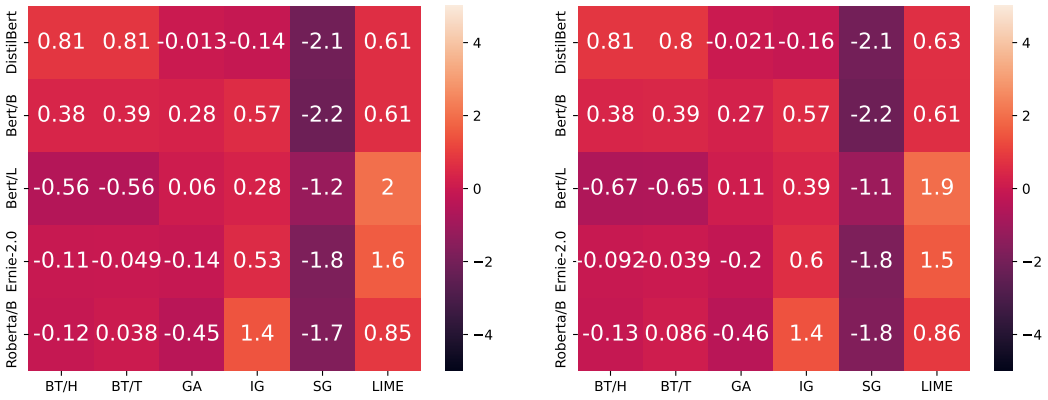
Figure 1: Averaged metric scores. A higher value indicates better faithfulness. The blanks indicate that the algorithm in a vanilla style is not suitable for the model.

4 Visualization

Some visualization examples can be found in `https://github.com/PaddlePaddle/InterpretDL/tree/master/examples`, and more examples can be easily computed.

References

[1] Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. Xai for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning*, pages 435–451. PMLR, 2022.



(a) Metric of MoRF.

(b) Metric of ABPC.

(c) Metric of PScore.

(d) Metric of INFD.

Figure 2: Averaged metric scores. A higher value indicates better faithfulness. The blanks indicate that the algorithm in a vanilla style is not suitable for the model.

Table 2: Time cost for computing the explanation result given all pairs of models and attribution methods. Note that the time cost is recorded after computing the explanations of 100 examples.

	BT/H	BT/T	GA	SG	IG	GradCAM
R50				7m50s	6m32s	5s
R101				12m06s	10m49s	10s
R152				16m58s	15m42s	14s
MobileNet				7m01s	5m49s	6s
VGG16				5m27s	3m49s	5s
ViT/S	1m52s	1m53s	5s	5m46s	4m28s	
ViT/B	2m59s	2m39s	7s	6m31s	7m16s	
ViT/L	10m30s	10m01s	14s	13m08s	15m07s	
ViT/MAE	3m40s	3m39s	8s	6m59s	7m20s	

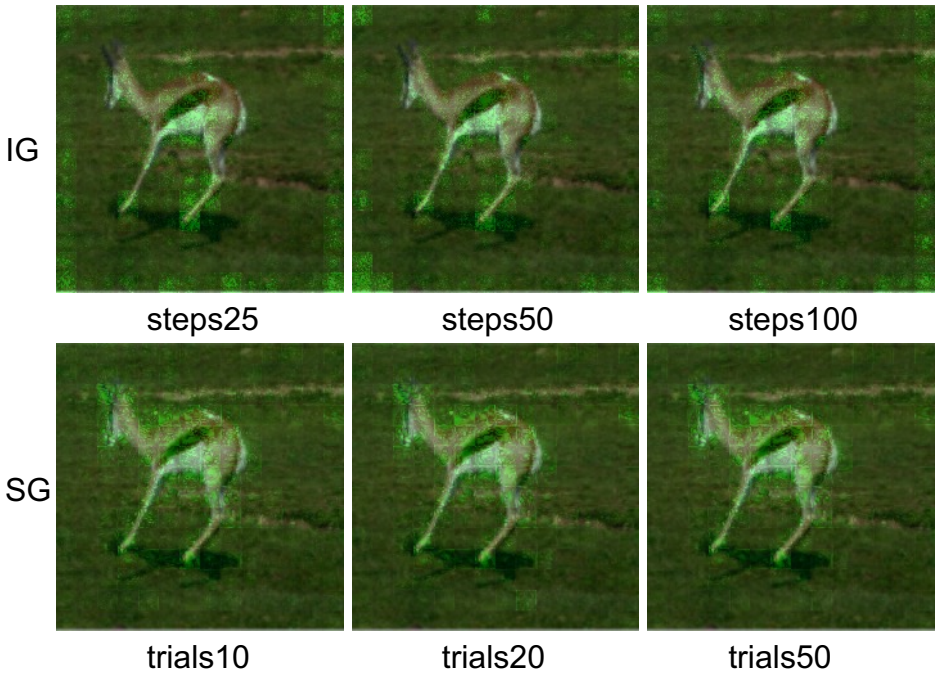


Figure 3: Visualization results of IG and SG.