A ATTRIBUTION PERFORMANCE WHEN USING OTHER SOTA TRANSFERABLE ATTACKS FOR MODEL PARAMETER EXPLORATION

Table 6: Attribution performance with difference attack methods

	Inception-v3		ResNet-50		VGG-16	
Method	Insertion Score	Deletion Score	Insertion Score	Deletion Score	Insertion Score	Deletion Score
AttEXplore-NoAttack	0.3959	0.0422	0.2584	0.0457	0.2121	0.0312
AttEXplore	0.4644	0.0313	0.4021	0.0308	0.3097	0.0237
AttEXplore-PGD	0.402	0.037	0.2901	0.0294	0.2258	0.0199
AttEXplore-DI-FGSM	0.4007	0.0416	0.2908	0.0427	0.2283	0.031
AttEXplore-TI-FGSM	0.3943	0.0344	0.3229	0.0348	0.2511	0.0268
AttEXplore-MI-FGSM	0.398	0.0414	0.2606	0.0447	0.2137	0.0304
AttEXplore-SINI-FGSM	0.4418	0.0314	0.3134	0.0292	0.2564	0.0227
AttEXplore-NAA	0.436	0.034	0.3058	0.0381	0.2559	0.0248

B DETAILED PROOFS OF TWO AXIOMS

Firstly, during the iterative process, the changes in the gradient along the integration path are captured by the original input information. Furthermore, it is not retroactive since feature values in previous iterations are unchanged in subsequent iterations. Therefore, the attribution result must be non-zero, which meets the definition of sensitivity. Here is the mathematical proof.

Poof of Eq.3 :

We use the first-order Taylor approximation to expand the loss function and combine the information for the path from Δx^0 to Δx^T.

$$\begin{aligned} L\left(x^t\right) &= L\left(x^{t-1}\right) + \frac{\partial L\left(x^{t-1}\right)}{\partial x^{t-1}}\left(x^t - x^{t-1}\right) + \varepsilon \\ \sum_{t=1}^{T} L\left(x^t\right) &= \sum_{t=0}^{T-1} L\left(x^t\right) + \sum_{t=0}^{T-1} \frac{\partial L\left(x^t\right)}{\partial x^t}\left(x^{t+1} - x^t\right) \\ A = L\left(x^T\right) - L\left(x^0\right) &= \sum_{t=0}^{T-1} \frac{\partial L\left(x^t\right)}{\partial x^t}\left(x^{t+1} - x^t\right) \\ &= \sum_{t=0}^{T-1} g(x^t) \odot \triangle x^t = \int \triangle x^t \odot g(x^t) \mathrm{d}t \end{aligned} \tag{7}$$

Here ϵ is omitted due to the principle of higher-order Taylor expansions. And $\triangle x^t = x^{t+1} - x^t$, $g(x^t) = \frac{\partial L(x^t)}{\partial x^t}$.

Secondly, it is clear that the computational processes in AttEXplore follow the chain rule of gradients, which meets the definition of implementation invariance.

C ADDITIONAL EXPERIMENTS ON VIT-B/16

D INFD SCORE TESTS

E FPS DEFINITION

we use Frames Per Second (FPS) as an evaluation metric for our running efficiency. A higher FPS indicates a greater number of images generated per second, signifying a higher operational efficiency of the method.

$$FPS = \frac{Number\ of\ samples}{Running\ time\ of\ these\ samples} \tag{8}$$

Table 7: Attribution performance of AttEXplore and other competitive baselines on ViT-B/16

Model	Method	Insertion Score	Deletion Score
ViT-B/16	Saliency Map	0.373	0.125
ViT-B/16	BIG	0.422	0.093
ViT-B/16	GIG	0.335	0.046
ViT-B/16	DeepLIFT	0.296	0.063
ViT-B/16	EG	0.361	0.329
ViT-B/16	Fast IG	0.216	0.071
ViT-B/16	SG	0.428	0.035
ViT-B/16	AGI	0.425	0.069
ViT-B/16	IG	0.346	0.051
ViT-B/16	AttEXplore (ours)	**0.470**	**0.062**

Table 8: INFD Score

Model	AGI	BIG	DeepLIFT	EG	Fast IG	GIG	IG	Saliency Map	SG	AttEXplore
Inception-v3	3.839	3.928	110.158	111.631	111.44	37.67	66.509	4.078	63.659	3.728
ResNet-50	1.003	0.708	18.828	143.593	135.651	39.659	85.834	0.696	42.504	0.671
VGG16	0.88	0.498	9.746	220.376	211.104	47.988	124.474	0.499	72.912	0.6

F Additional Ablation Studies

F.1 Ablation Studies for the number of approximate features

Table 9: Result for the number of approximate features ($N < 10$)

	Inception-v3		ResNet-50		VGG-16	
N	Insertion Score	Deletion Score	Insertion Score	Deletion Score	Insertion Score	Deletion Score
1	0.4536	0.0282	0.3841	0.0262	0.2915	0.0190
2	0.4568	0.0284	0.3931	0.0275	0.2970	0.0196
3	0.4624	0.0298	0.3957	0.0274	0.3020	0.0210
4	0.4606	0.0292	0.3987	0.0286	0.3058	0.0214
5	0.4588	0.0300	0.3995	0.0282	0.3041	0.0223
6	0.4602	0.0288	0.3989	0.0289	0.3059	0.0221
7	0.4597	0.0301	0.3999	0.0288	0.3071	0.0224
8	0.4619	0.0301	0.4005	0.0289	0.3073	0.0229
9	0.4646	0.0299	0.3995	0.0291	0.3064	0.0225

We observe that when the perturbation rate is set to a larger value such as 48, the trend in model performance across different N does not become more clear. This might be attributed to the fact that a larger perturbation rate represents a larger search space. Although the number of approximate samples increases, it doesn't mean that all these samples are necessarily effective for the attribution result.

Table 10: Result for the number of approximate features N when the perturbation rate is 48

Model	N	1	2	3	4	5	6	7	8	9	10	20	30	40	50	60
Inception-v3	Insertion score	0.459	0.461	0.465	0.465	0.464	0.467	0.468	0.47	0.472	0.466	0.471	0.473	0.473	0.474	0.474
	Deletion score	0.028	0.028	0.028	0.027	0.028	0.028	0.027	0.028	0.029	0.028	0.029	0.029	0.03	0.029	0.03
ResNet-50	Insertion score	0.406	0.414	0.417	0.417	0.419	0.42	0.422	0.422	0.422	0.423	0.425	0.427	0.427	0.428	0.428
	Deletion score	0.027	0.028	0.028	0.029	0.03	0.029	0.03	0.03	0.03	0.03	0.031	0.031	0.032	0.032	0.032
VGG16	Insertion score	0.298	0.304	0.306	0.308	0.308	0.308	0.31	0.31	0.311	0.311	0.312	0.313	0.313	0.313	0.315
	Deletion score	0.019	0.019	0.019	0.02	0.02	0.02	0.02	0.02	0.021	0.021	0.021	0.021	0.022	0.022	0.022

F.2 Ablation Studies for the total attack iterations

Table 11: Result for the total attack iterations num_steps

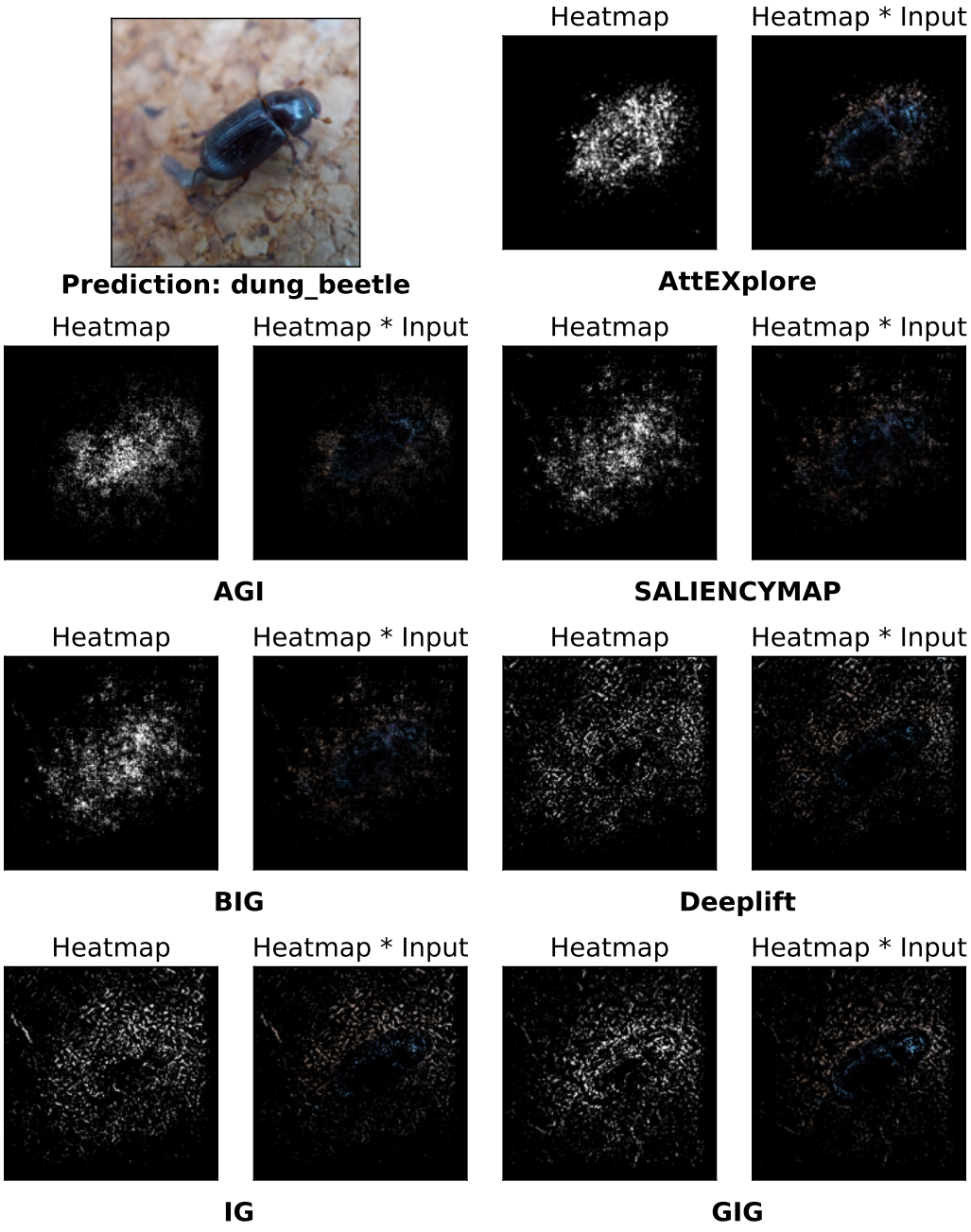
	Inception-v3		ResNet-50		VGG-16	
num_steps	Insertion Score	Deletion Score	Insertion Score	Deletion Score	Insertion Score	Deletion Score
1	0.4236	0.0281	0.3469	0.0249	0.2645	0.0195
2	0.4488	0.0291	0.3835	0.0261	0.2929	0.0208
3	0.4557	0.0297	0.3936	0.0273	0.3029	0.0217
4	0.4607	0.0301	0.3966	0.0283	0.3061	0.0218

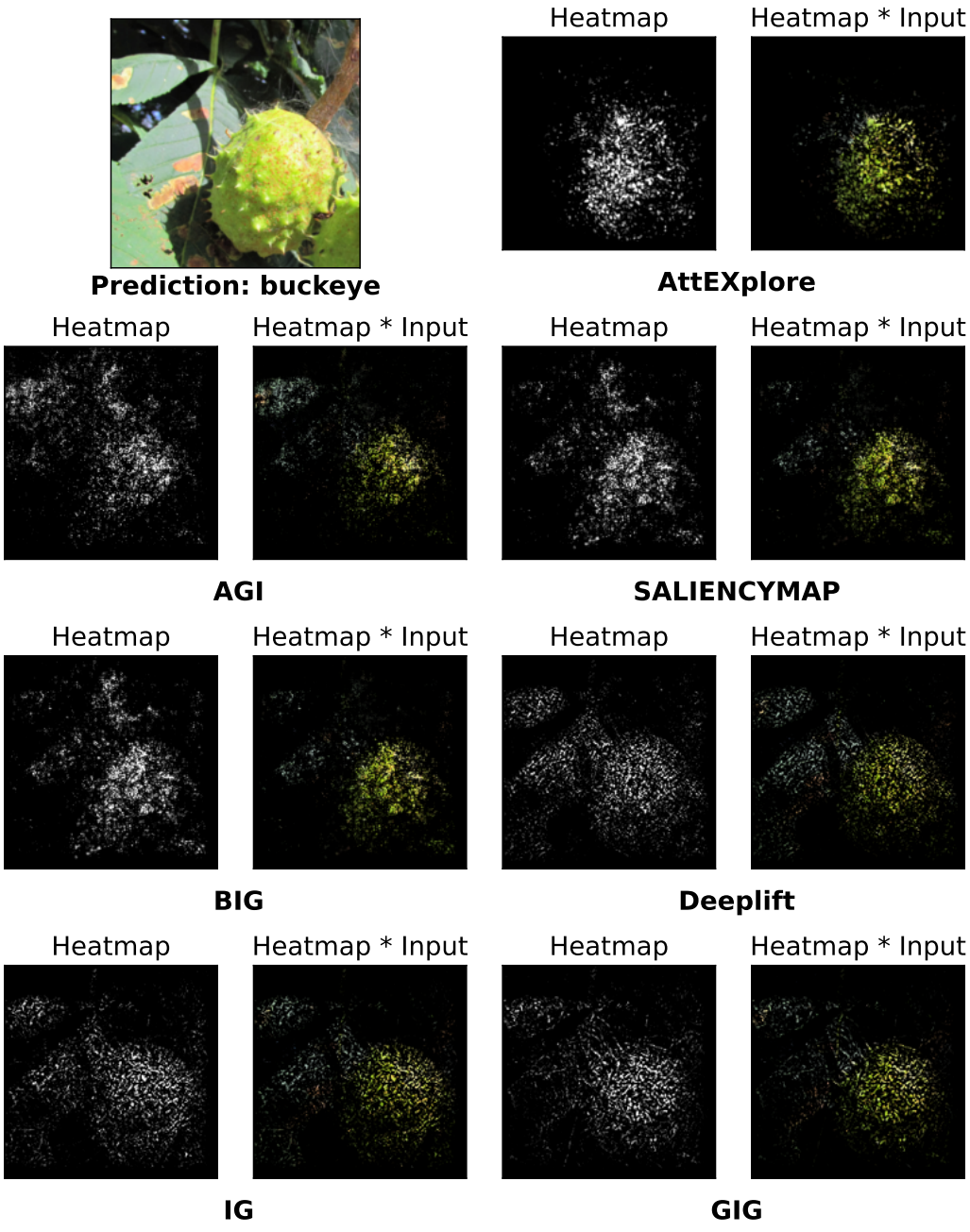
F.3 Ablation Studies for the perturbation rate

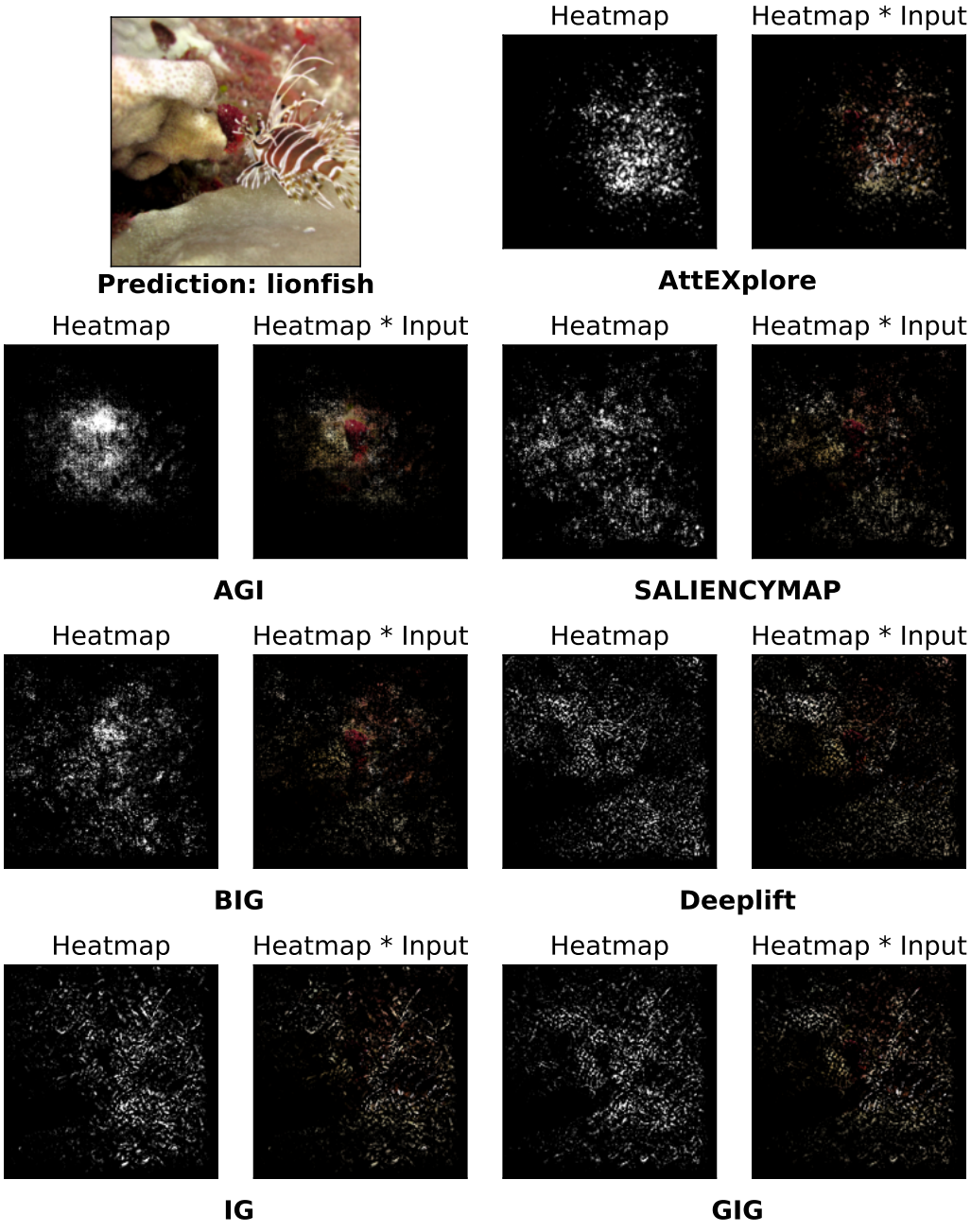
Table 12: Result for the perturbation rate ($\epsilon < 8$)

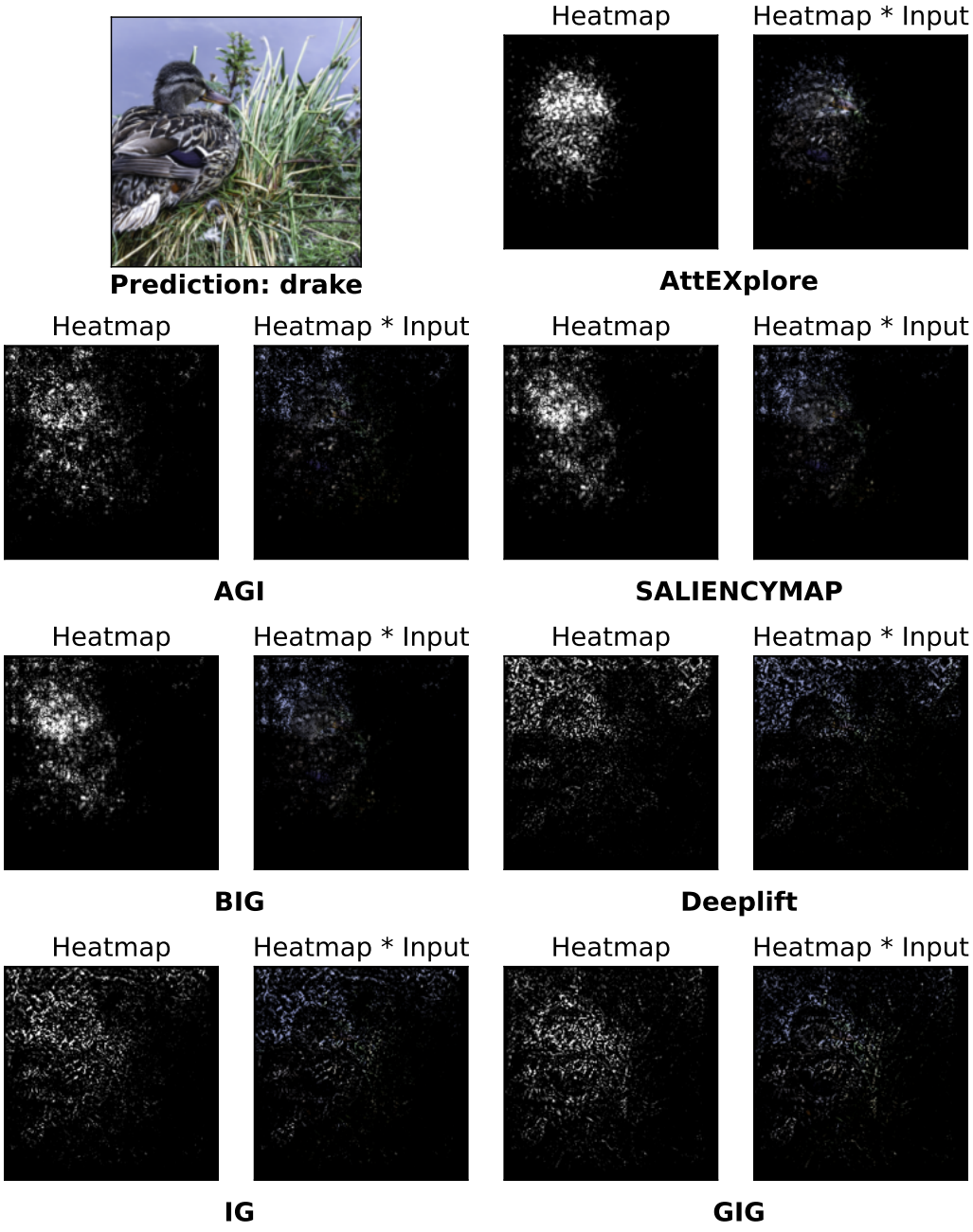
Model	ϵ	1	2	3	4	5	6	7
Inception-v3	Insertion score	0.459	0.459	0.459	0.461	0.46	0.461	0.46
	Deletion score	0.031	0.032	0.032	0.031	0.032	0.031	0.032
ResNet-50	Insertion score	0.397	0.397	0.398	0.401	0.4	0.403	0.403
	Deletion score	0.031	0.031	0.031	0.032	0.032	0.032	0.032
VGG16	Insertion score	0.298	0.299	0.3	0.3	0.3	0.302	0.303
	Deletion score	0.021	0.022	0.022	0.022	0.022	0.022	0.022

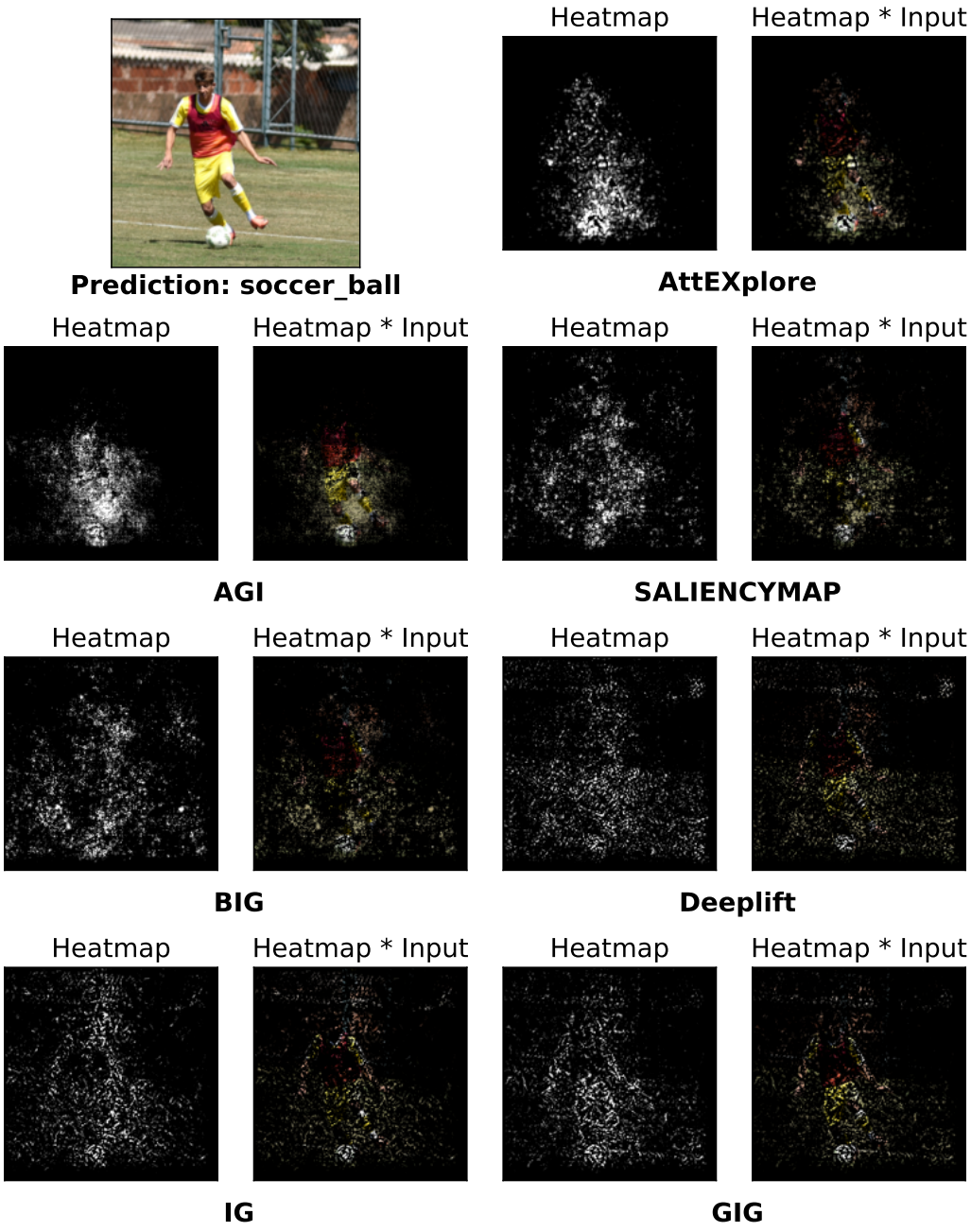
G ADDITIONAL VISUALIZATION RESULTS OF OUR ATTEXPLORE

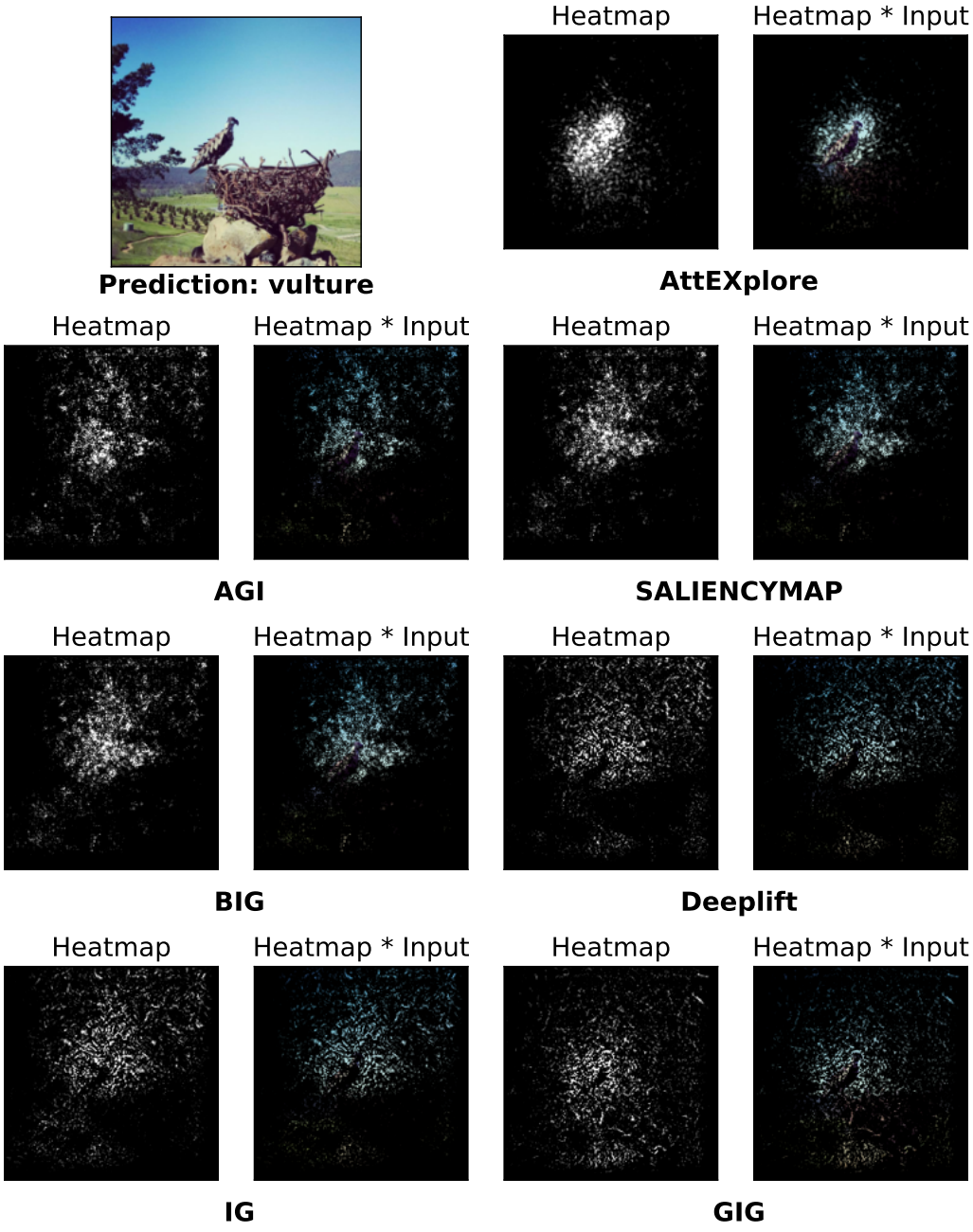


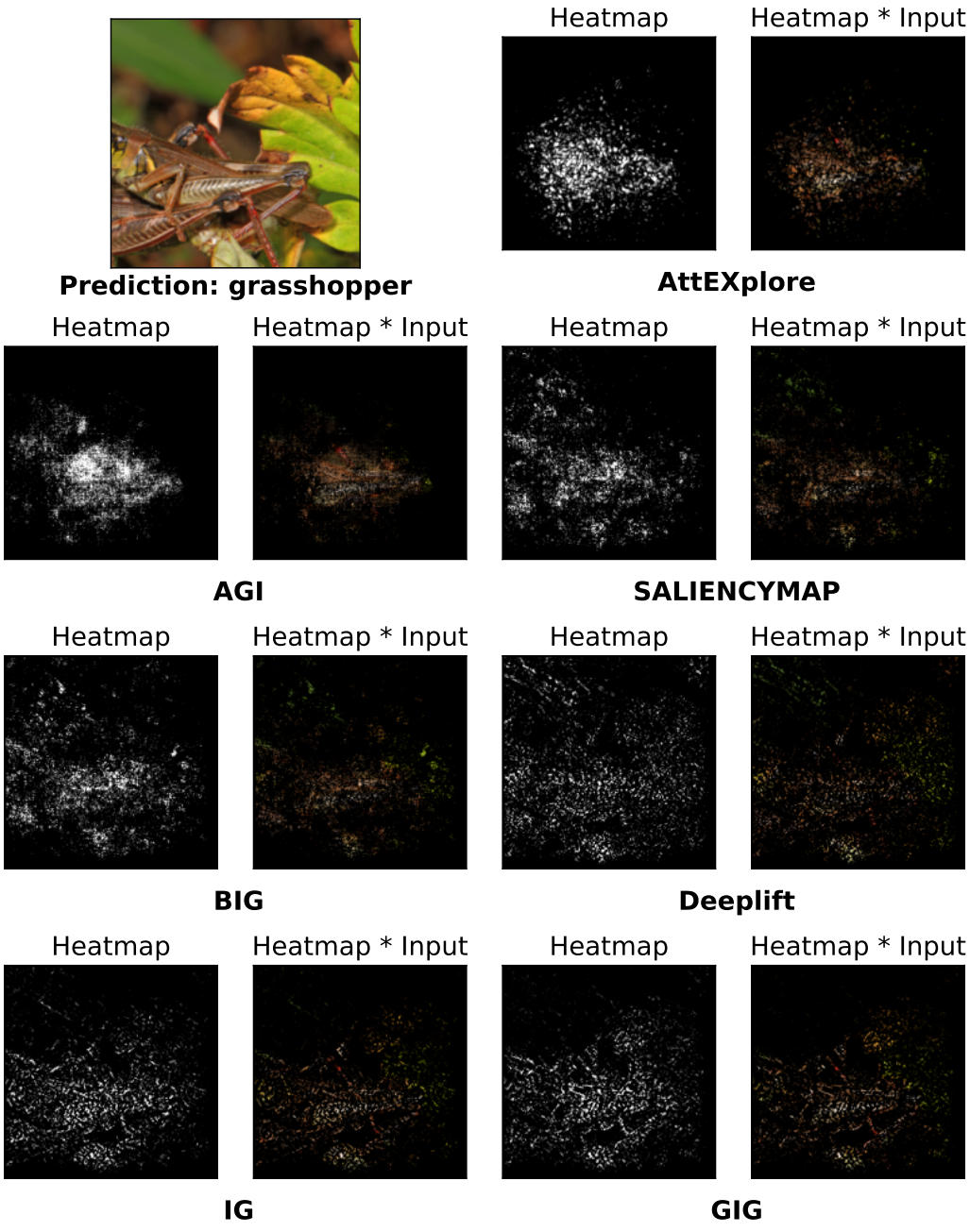

Heatmap
Heatmap * Input
Prediction: buckeye
AttEXplore
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
AGI
SALIENCYMAP
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
BIG
Deeplift
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
IG
GIG

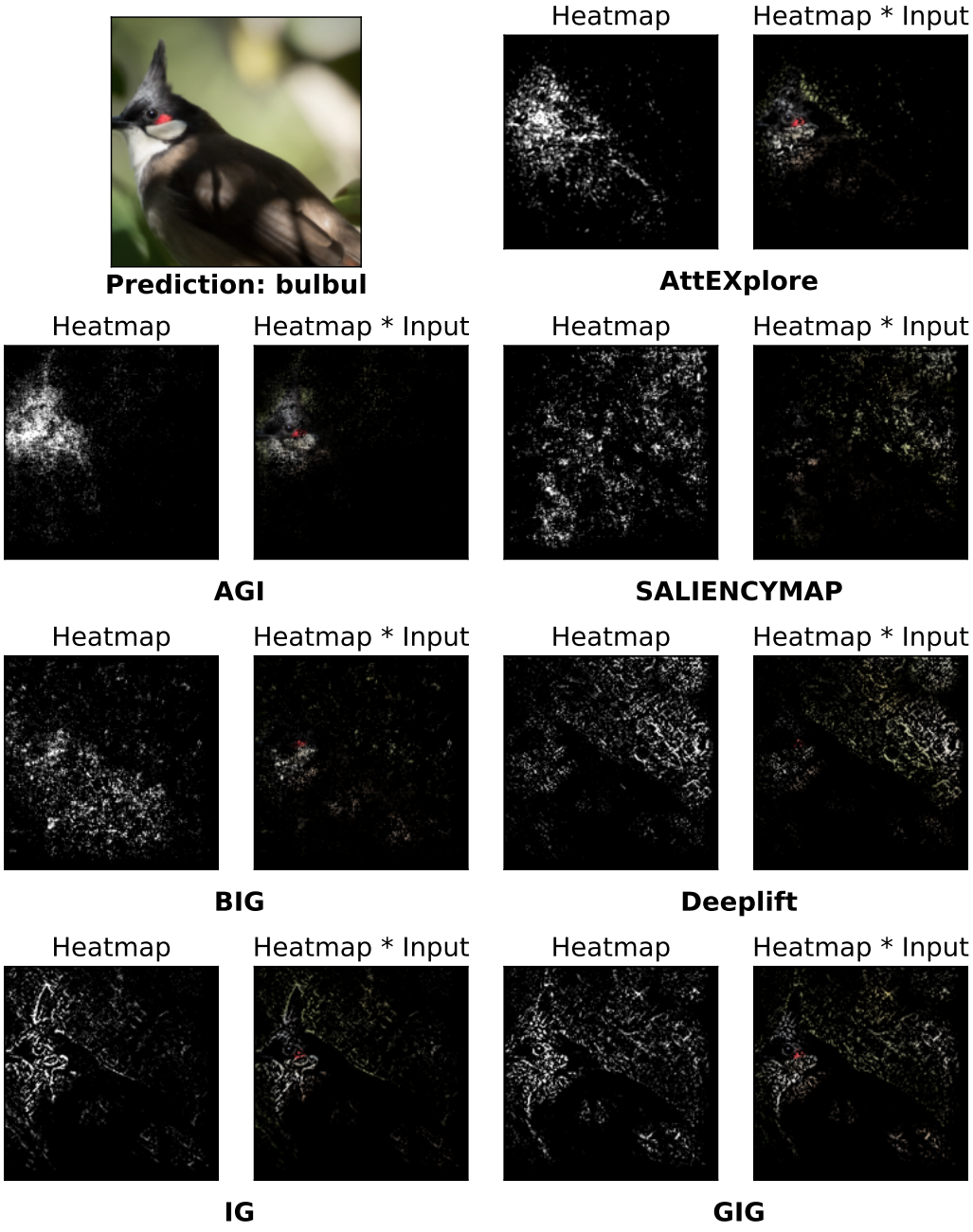

Heatmap
Heatmap * Input
Prediction: lionfish
AttEXplore
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
AGI
SALIENCYMAP
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
BIG
Deeplift
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
IG
GIG

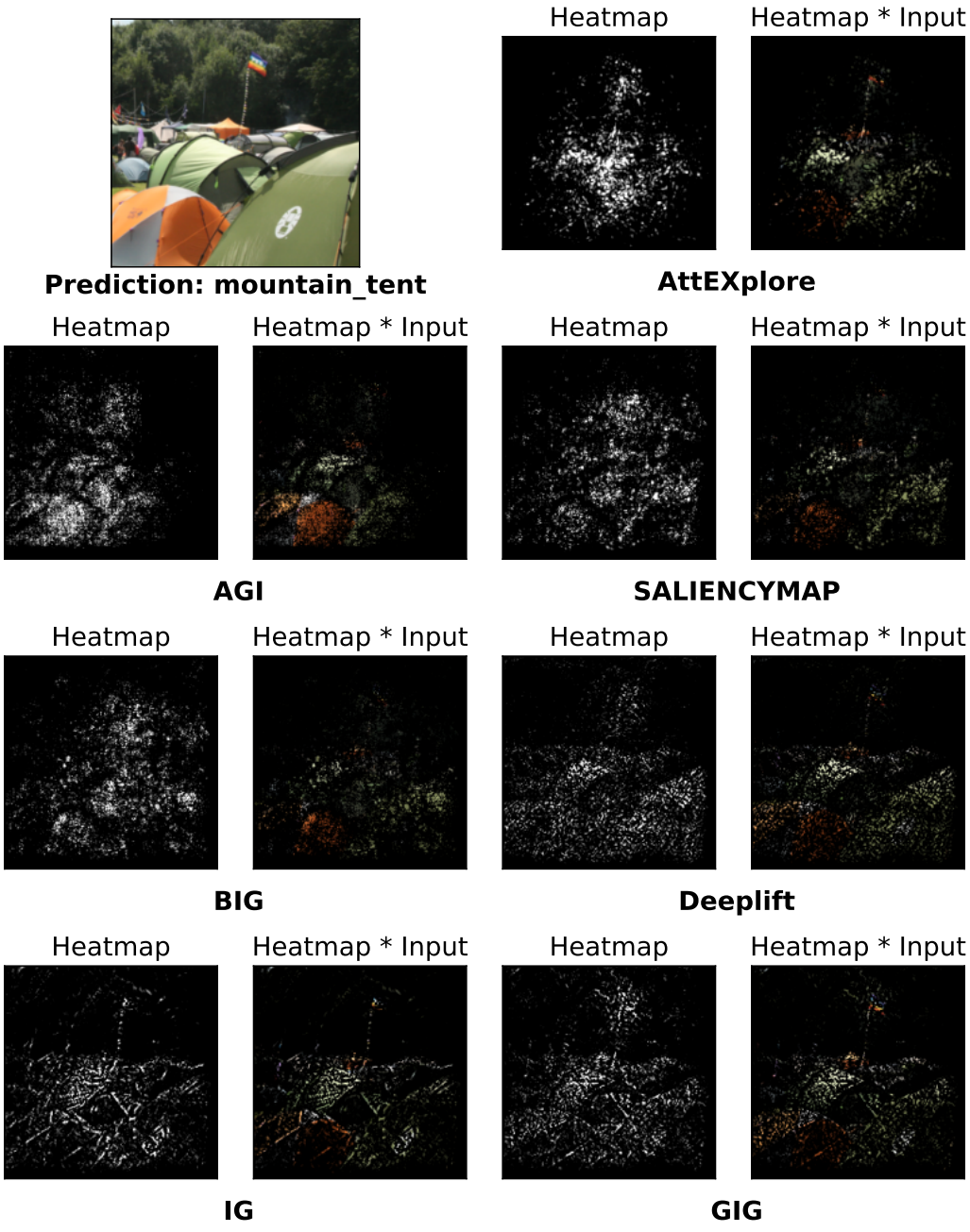

Heatmap
Heatmap * Input
Prediction: drake
AttEXplore
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
AGI
SALIENCYMAP
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
BIG
Deeplift
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
IG
GIG

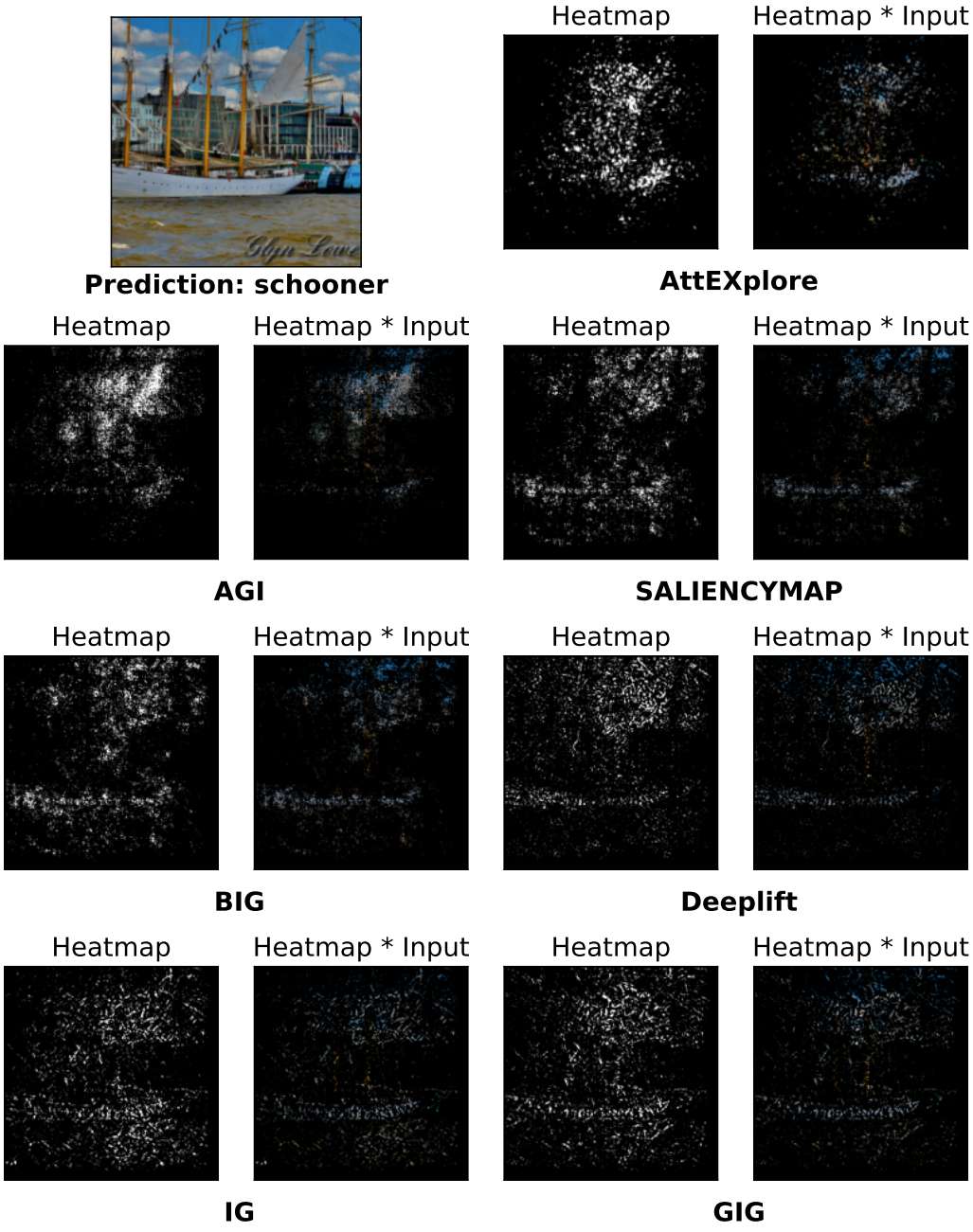

Heatmap
Heatmap * Input
Prediction: soccer_ball
AttEXplore
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
AGI
SALIENCYMAP
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
BIG
Deeplift
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
IG
GIG

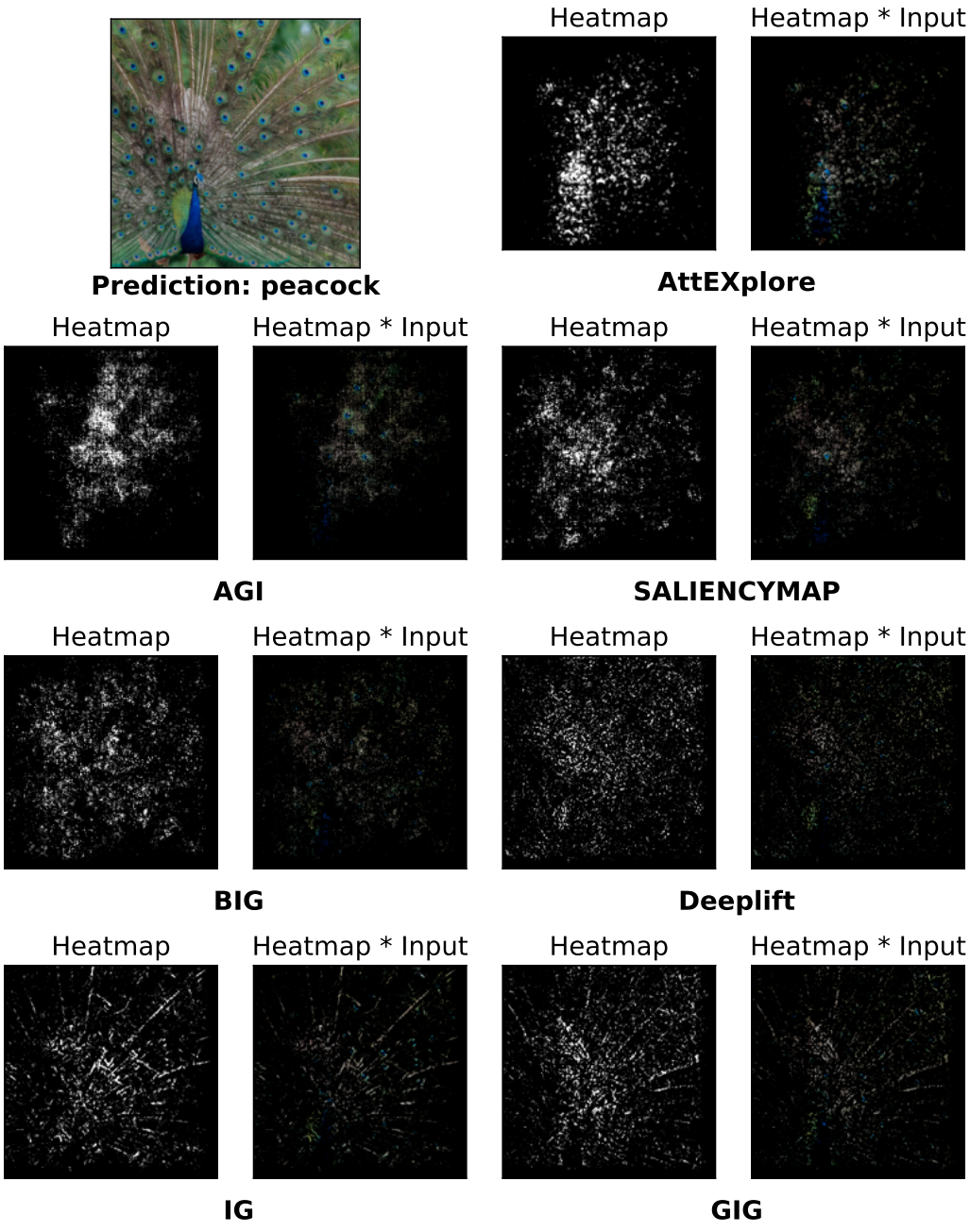

Heatmap
Heatmap * Input
Prediction: vulture
AttEXplore
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
AGI
SALIENCYMAP
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
BIG
Deeplift
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
IG
GIG


Heatmap
Heatmap * Input
Prediction: grasshopper
AttEXplore
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
AGI
SALIENCYMAP
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
BIG
Deeplift
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
IG
GIG


Heatmap
Heatmap * Input
Prediction: bulbul
AttEXplore
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
AGI
SALIENCYMAP
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
BIG
Deeplift
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
IG
GIG


Heatmap
Heatmap * Input
Prediction: mountain_tent
AttEXplore
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
AGI
SALIENCYMAP
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
BIG
Deeplift
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
IG
GIG


Heatmap
Heatmap * Input
Prediction: schooner
AttEXplore
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
AGI
SALIENCYMAP
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
BIG
Deeplift
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
IG
GIG


Heatmap
Heatmap * Input
Prediction: peacock
AttEXplore
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
AGI
SALIENCYMAP
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
BIG
Deeplift
Heatmap
Heatmap * Input
Heatmap
Heatmap * Input
IG
GIG