

A PROOF OF AXIOMS

SENSITIVITY PROOF. Given that:

$$\mathbb{E}_{\hat{x}=u(\tilde{x}), \tilde{x} \sim B_{\frac{\epsilon}{2}}(x)} [L(\hat{x}; y, w) - L(x; y, w)] = \sum_{i=1}^n \mathbb{E}_{\hat{x}=u(\tilde{x}), \tilde{x} \sim B_{\frac{\epsilon}{2}}(x)} \left[(\hat{x}_i - \tilde{x}_i) \cdot \frac{\partial L(\tilde{x}; y, w)}{\partial \tilde{x}_i} \right] \quad (1)$$

the total attribution on the right side equals the expected change in the loss function. A change in the loss function necessarily results in a non-zero attribution on the right, proving sensitivity. \square

B IMPLEMENTATION INVARIANCE PROOF

IMPLEMENTATION INVARIANCE PROOF. The Local Attribution (LA) algorithm adheres to the chain rule. Based on the properties of gradients, the LA algorithm satisfies implementation invariance, ensuring that results are consistent across different valid implementations of the same functional relationship. \square

C PROOF OF ϵ -Local Space

PROOF. Consider the functions u^u and u^t defined for perturbation within the local space:

$$\begin{aligned} u^u(\tilde{x}) &= \tilde{x} + \frac{\epsilon}{2} \cdot \text{sign} \left(\frac{\partial L(\tilde{x}, y, w)}{\partial \tilde{x}} \right) \\ u^t(\tilde{x}) &= \tilde{x} - \frac{\epsilon}{2} \cdot \text{sign} \left(\frac{\partial L(\tilde{x}, y^t, w)}{\partial \tilde{x}} \right) \\ |\tilde{x} \cdot x| &\leq \frac{\epsilon}{2} \\ u^u(\tilde{x}) - \tilde{x} &= \frac{\epsilon}{2} \cdot \text{sign} \left(\frac{\partial L(\tilde{x}, y, w)}{\partial \tilde{x}} \right) \\ -\frac{\epsilon}{2} &\leq \frac{\epsilon}{2} \cdot \text{sign} \left(\frac{\partial L(\tilde{x}, y, w)}{\partial \tilde{x}} \right) \leq \frac{\epsilon}{2} \\ |\hat{x} - x| &= |\hat{x} - \tilde{x} + \tilde{x} - x| \leq \epsilon \\ &\text{thus confirming presence within the } \epsilon\text{-Local Space.} \end{aligned} \quad (2)$$

\square

D PROOF OF SPACE CONSTRAINT

PROOF OF SPACE CONSTRAINT. The iterative process for updating positions in an adversarial example generation context can be described as follows:

$$\begin{aligned} x_u^k &= x + \frac{\epsilon}{2} \cdot \text{sign} \left(\frac{\partial L(x_u^{k-1}; y, w)}{\partial x_u^{k-1}} \right) \\ x_t^k &= x - \frac{\epsilon}{2} \cdot \text{sign} \left(\frac{\partial L(x_t^k, y^t, w)}{\partial x_t^{k-1}} \right) \end{aligned}$$

where x_u denotes the untargeted attack, and x_t denotes the targeted attack. We assert the following constraint on the adversarial perturbation:

$$\left| x_u^k - x \right| = \left| \frac{\epsilon}{2} \cdot \text{sign} \left(\frac{\partial L(x_u^{k-1}; y, w)}{\partial x_u^{k-1}} \right) \right| \leq \frac{\epsilon}{2}, \text{ maintaining presence within the } \epsilon\text{-Local Space.}$$

The same reasoning applies to x_t^k , demonstrating that each iterative step remains within the designated local space. \square

E PSEUDOCODE

Algorithm 1 Local Attribution (LA)

Input: sample x , number of iterations N , loss function L , Spatial Range(number of target) s , model parameters w

Output: A

```

1:  $A = \mathbf{0}$    $\mathbf{0} = [0, 0, \dots, 0] \in \mathbb{R}^n$ 
2:  $x_u^0 = x_t^0 = x$ 
3: for  $k$  in  $\text{range}(N)$  do
4:    $x_u^k = x + \frac{\epsilon}{2} \cdot \text{sign} \left( \frac{\partial L(x_u^{k-1}; y, w)}{\partial x_u^{k-1}} \right)$ 
5:    $A += \frac{\epsilon}{2} \cdot \text{sign} \left( \frac{\partial L(x_u^{k-1}; y, w)}{\partial x_u^{k-1}} \right) \cdot \frac{\partial L(x_u^{k-1}; y, w)}{\partial x_u^{k-1}}$ 
6: end for
7: for  $i$  in  $\text{range}(s)$  do
8:   for  $k$  in  $\text{range}(N)$  do
9:      $x_t^k = x + \frac{\epsilon}{2} \cdot \text{sign} \left( \frac{\partial L(x_t^{k-1}; y^t, w)}{\partial x_t^{k-1}} \right)$ 
10:     $A -= \frac{\epsilon}{2} \cdot \text{sign} \left( \frac{\partial L(x_t^{k-1}; y^t, w)}{\partial x_t^{k-1}} \right) \cdot \frac{\partial L(x_t^{k-1}; y^t, w)}{\partial x_t^{k-1}}$ 
11:   end for
12: end for
13: return  $A$ 
```

F EXPERIMENTAL RESULTS

Table 1: The result of replicated the experiments from previous works

Method	Inception-v3		ResNet-50		VGG16		MaxViT-T	
	Insertion	Deletion	Insertion	Deletion	Insertion	Deletion	Insertion	Deletion
FIG	0.201722	0.045629	0.106266	0.032358	0.07933	0.027029	0.462121	0.182173
EG	0.375499	0.265265	0.350081	0.28264	0.356653	0.337217	0.598585	0.515332
DeepLIFT	0.295152	0.041533	0.124774	0.030503	0.09342	0.023026	0.498488	0.181546
GIG	0.318705	0.034337	0.144963	0.019132	0.10252	0.017318	0.545019	0.138792
IG	0.320825	0.04258	0.145412	0.028333	0.095863	0.023163	0.541594	0.18818
SG	0.38911	0.033278	0.277219	0.022857	0.186208	0.016392	0.641634	0.139467
BIG	0.48401	0.053815	0.290461	0.046713	0.226557	0.037233	0.568201	0.186599
SM	0.533356	0.0631	0.31544	0.056741	0.270308	0.041743	0.489565	0.195568
MFABA	0.538468	0.063881	0.320002	0.055452	0.279122	0.040634	0.440624	0.358368
AGI	0.572294	0.058431	0.500747	0.051438	0.397331	0.042029	0.645392	0.198408
AttEXplore	0.618792	0.044244	0.504209	0.033338	0.442779	0.0282	0.615814	0.15773
LA (our)	0.646301	0.067499	0.549666	0.047156	0.437295	0.03378	0.704771	0.208705

Table 2: The result on 1000 images from the ILSVRC 2012

Method	Inception-v3		ResNet-50		VGG16		MaxViT-T	
	Insertion	Deletion	Insertion	Deletion	Insertion	Deletion	Insertion	Deletion
FIG	0.041396	0.05868	0.028387	0.041294	0.018308	0.028613	0.018308	0.028613
DeepLIFT	0.065667	0.046959	0.039129	0.028892	0.028824	0.016541	0.028824	0.016541
GIG	0.077912	0.027204	0.045593	0.016607	0.033138	0.011634	0.033138	0.011634
IG	0.078717	0.030372	0.045788	0.022333	0.031213	0.015054	0.031213	0.015054
SG	0.152161	0.026479	0.11807	0.015998	0.10324	0.012112	0.10324	0.012112
BIG	0.157666	0.069774	0.103139	0.060935	0.06877	0.039905	0.06877	0.039905
SM	0.070064	0.043584	0.05795	0.03228	0.0426	0.019605	0.0426	0.019605
MFABA	0.228086	0.070928	0.121357	0.068072	0.086451	0.042018	0.086451	0.042018
EG	0.333171	0.376207	0.260612	0.300023	0.176334	0.155098	0.176334	0.155098
AGI	0.29917	0.068373	0.310282	0.057212	0.209355	0.037446	0.209355	0.037446
AttEXplore	0.30662	0.062042	0.237887	0.046649	0.18695	0.037588	0.18695	0.037588
LA	0.416221	0.054769	0.417947	0.044304	0.301272	0.029915	0.301272	0.029915

Table 3: The result of replicated the experiments from previous works on 1000 images from the ILSVRC 2012

Method	Inception-v3		ResNet-50		VGG16		MaxViT-T	
	Insertion	Deletion	Insertion	Deletion	Insertion	Deletion	Insertion	Deletion
FIG	0.144397	0.032023	0.087252	0.029646	0.060985	0.019745	0.434148	0.171334
DeepLIFT	0.218215	0.029595	0.10331	0.026814	0.07006	0.01611	0.485588	0.169753
GIG	0.233052	0.022558	0.123626	0.01724	0.079532	0.012764	0.542341	0.117117
IG	0.225676	0.026889	0.112144	0.022977	0.0688	0.015629	0.530153	0.171547
EG	0.303453	0.300624	0.323657	0.300264	0.195079	0.178762	0.523863	0.457302
SG	0.293349	0.020552	0.232979	0.018048	0.140715	0.013417	0.6432	0.111883
BIG	0.357469	0.036464	0.227733	0.04042	0.175332	0.02891	0.527154	0.176192
SM	0.396799	0.040684	0.254677	0.046277	0.208166	0.030726	0.481034	0.175937
AGI	0.423038	0.042443	0.379195	0.044985	0.255715	0.030216	0.607887	0.165533
MFABA	0.396052	0.040035	0.257576	0.045806	0.214512	0.029935	0.419738	0.307159
AttEXplore	0.463166	0.029513	0.402695	0.029542	0.309154	0.02236	0.58905	0.129836
LA (our)	0.495721	0.046036	0.446708	0.037803	0.320236	0.023418	0.676405	0.177876

G ATTRIBUTION RESULTS

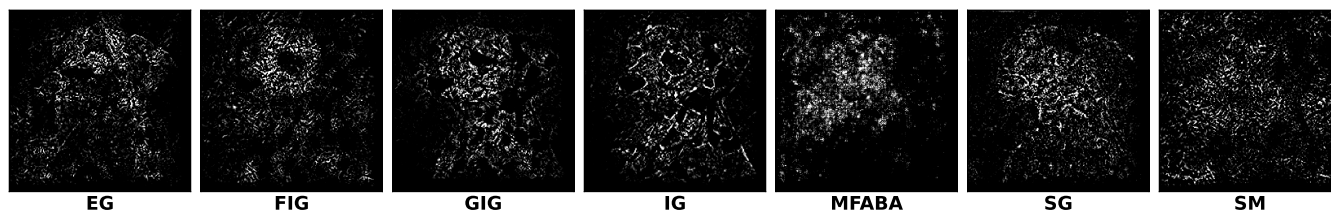
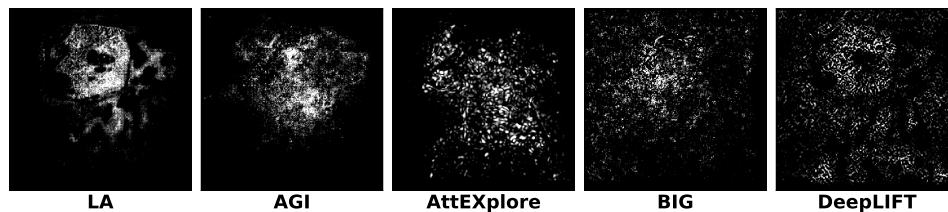


Figure 1: Attribution Results on the Inception-v3

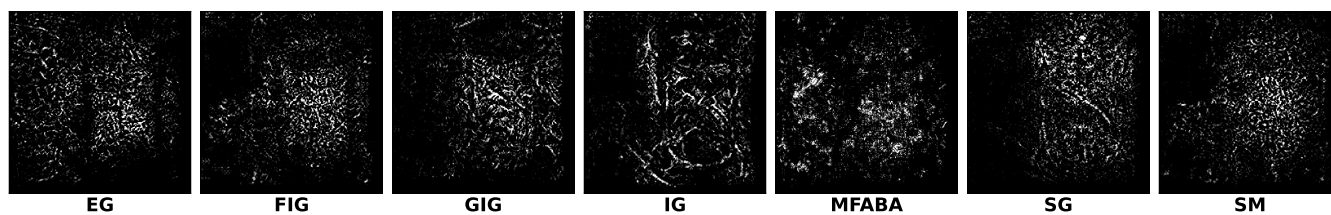
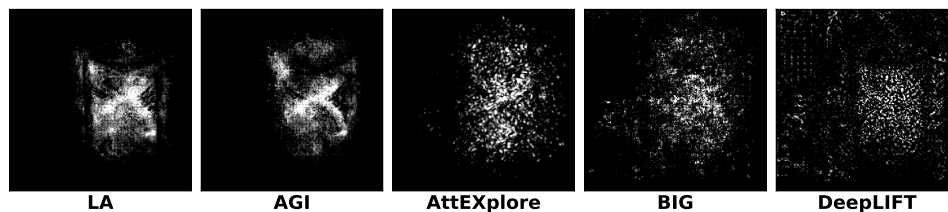


Figure 2: Attribution Results on the Inception-v3

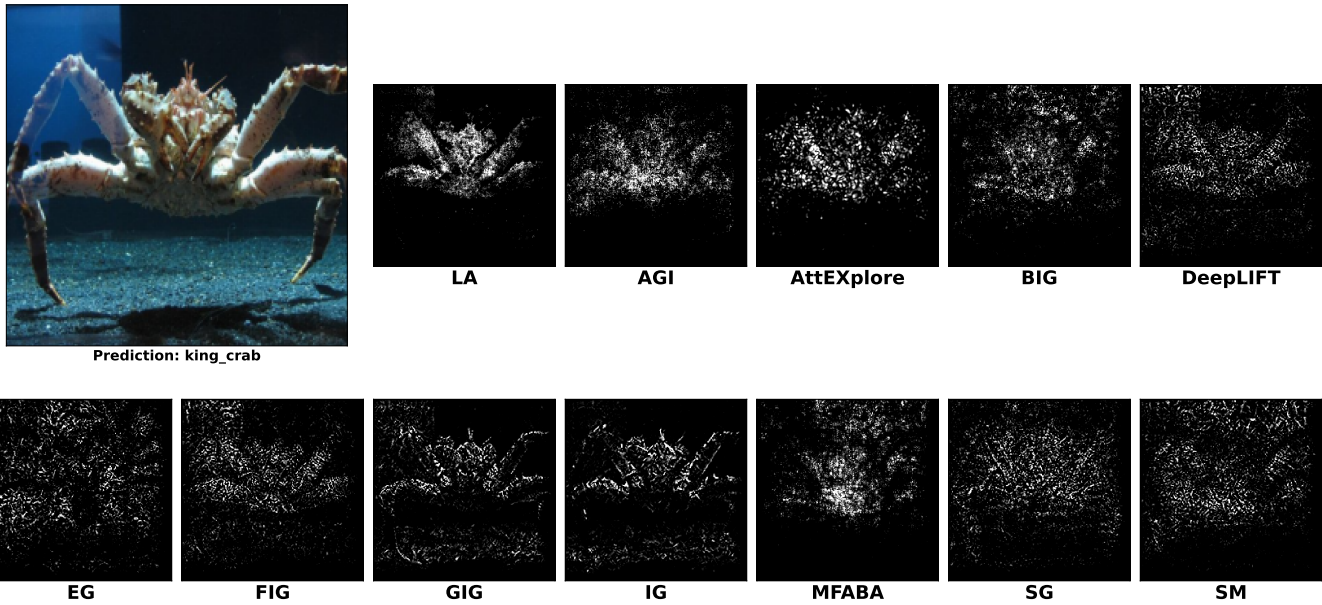


Figure 3: Attribution Results on the Inception-v3

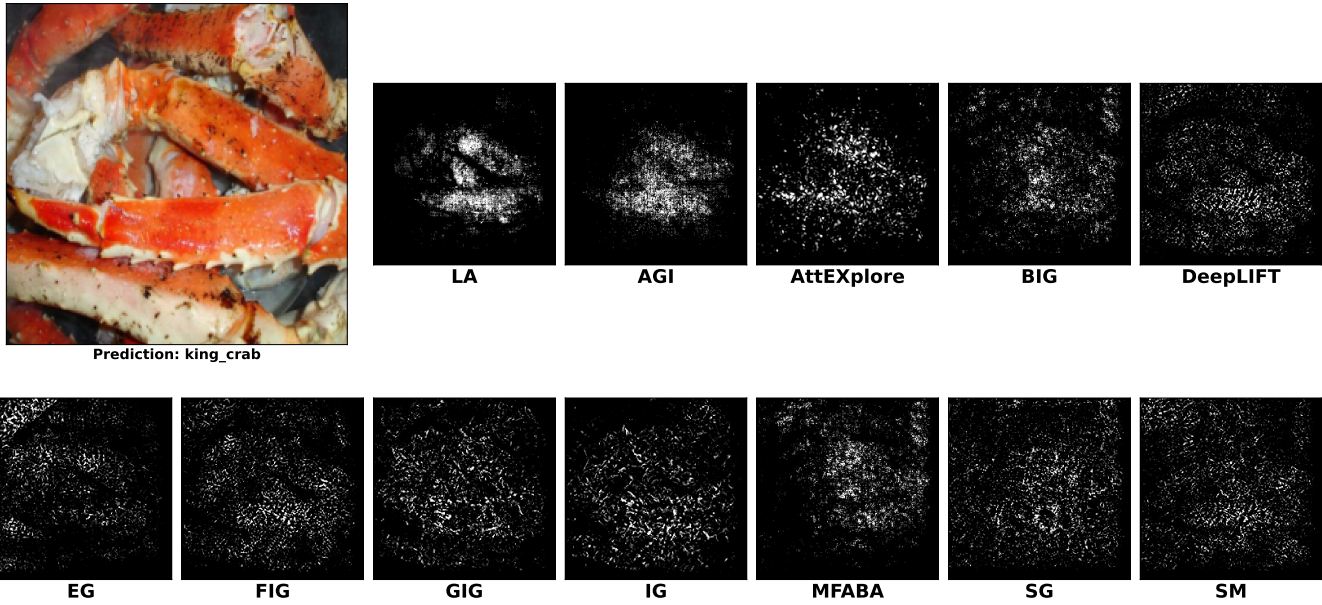


Figure 4: Attribution Results on the Inception-v3

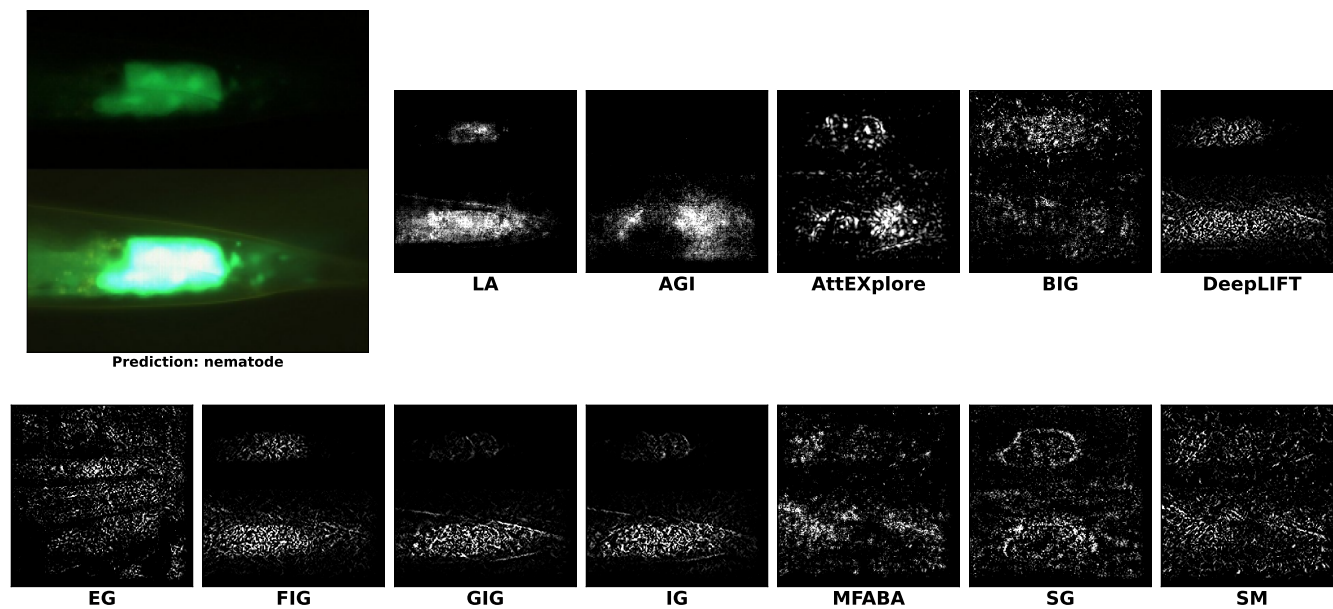


Figure 5: Attribution Results on the Inception-v3

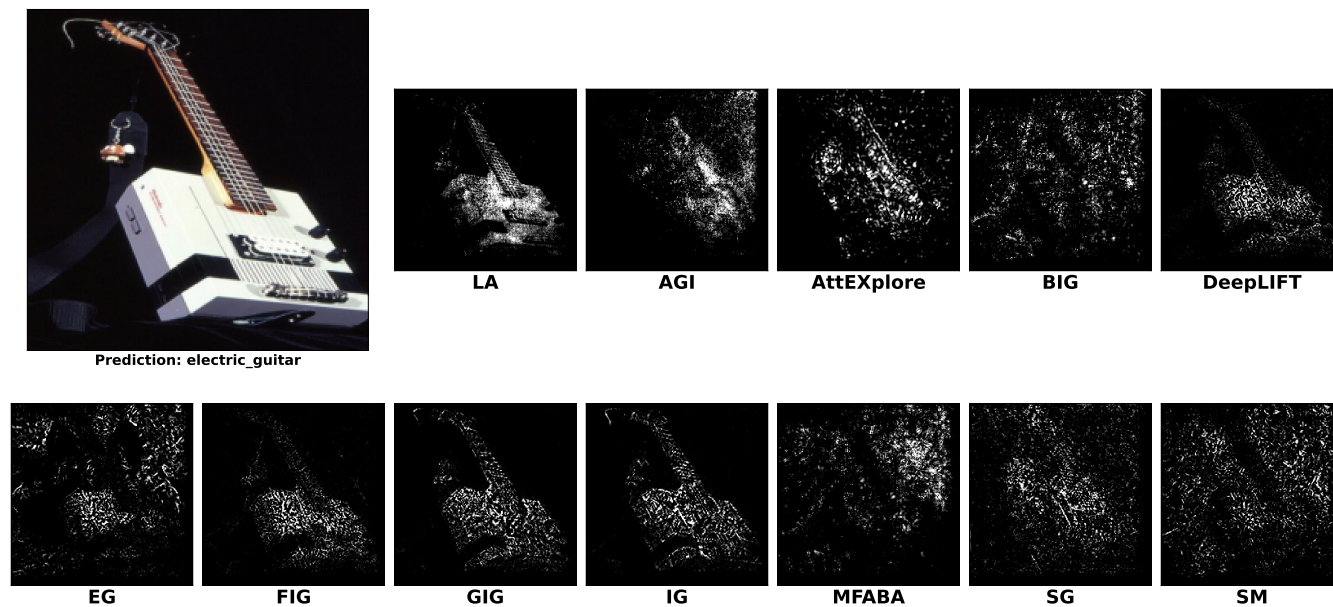


Figure 6: Attribution Results on the Inception-v3

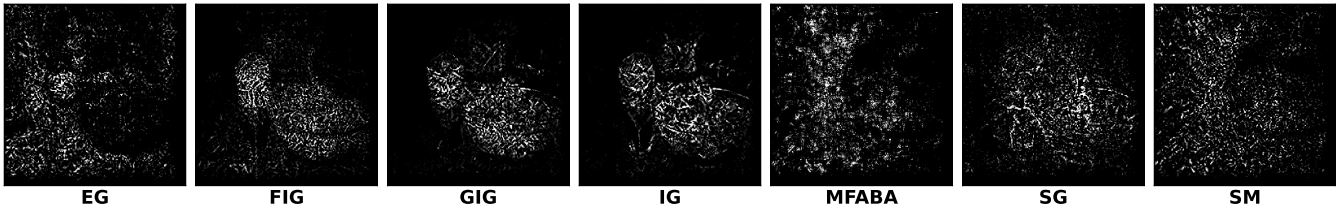
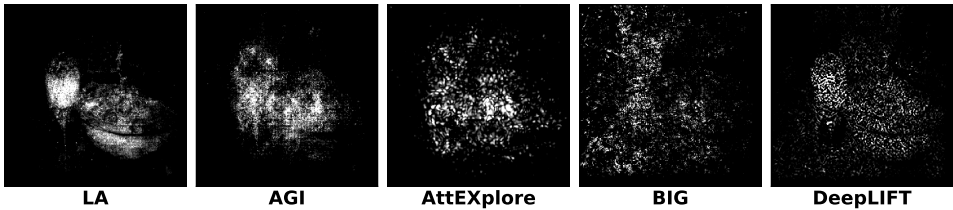


Figure 7: Attribution Results on the Inception-v3

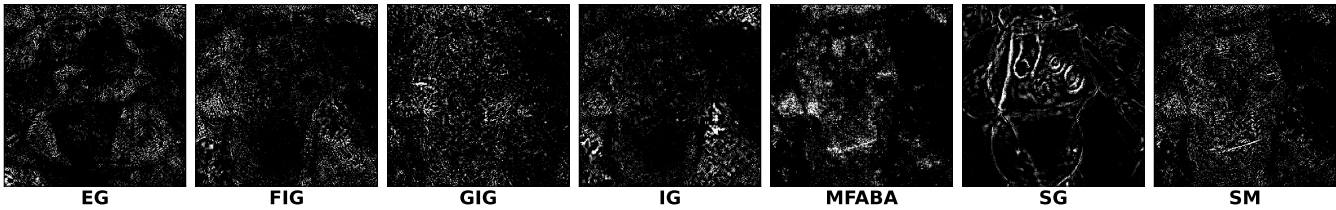
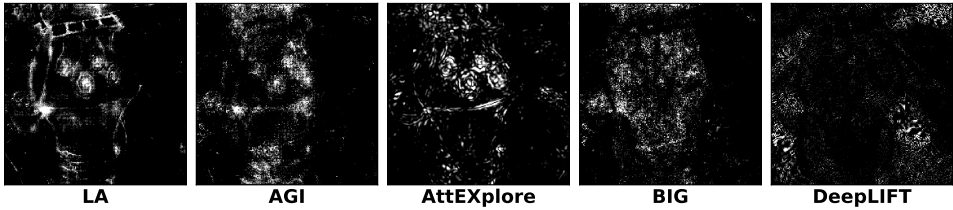


Figure 8: Attribution Results on the MaxViT-T

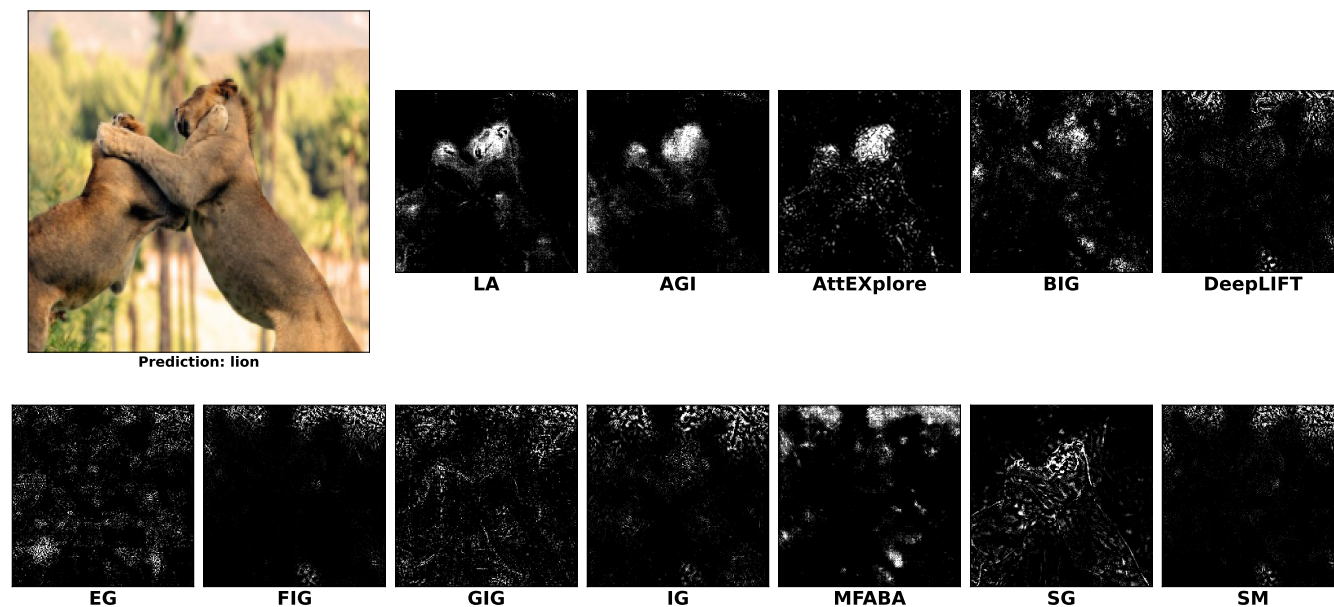


Figure 9: Attribution Results on the MaxViT-T

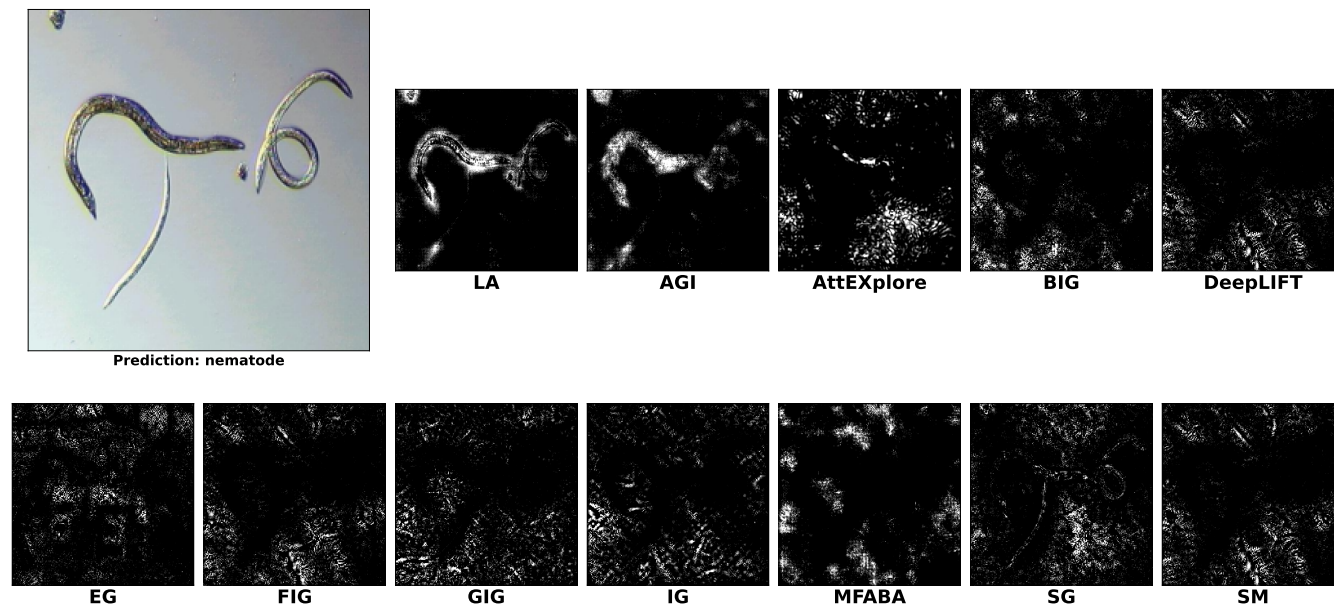
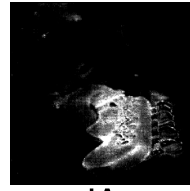


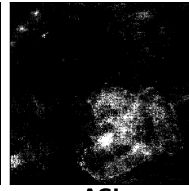
Figure 10: Attribution Results on the MaxViT-T



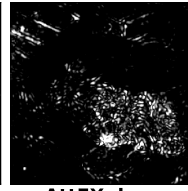
Prediction: electric_guitar



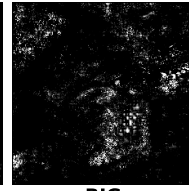
LA



AGI



AttEXplore



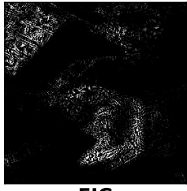
BIG



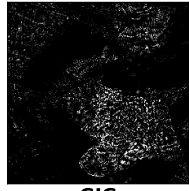
DeepLIFT



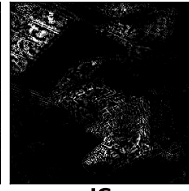
EG



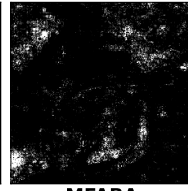
FIG



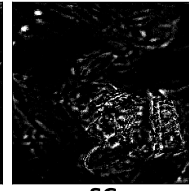
GIG



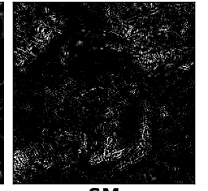
IG



MFABA



SG

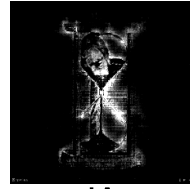


SM

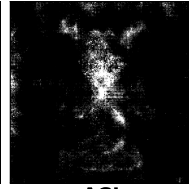
Figure 11: Attribution Results on the MaxViT-T



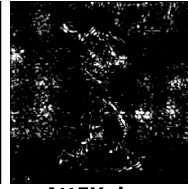
Prediction: hourglass



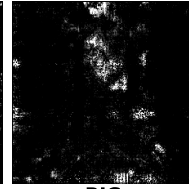
LA



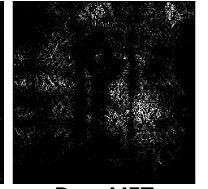
AGI



AttEXplore



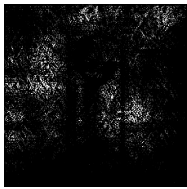
BIG



DeepLIFT



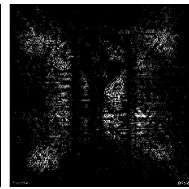
EG



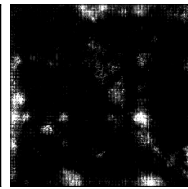
FIG



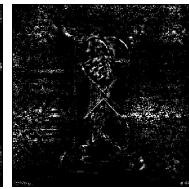
GIG



IG



MFABA



SG



SM

Figure 12: Attribution Results on the MaxViT-T

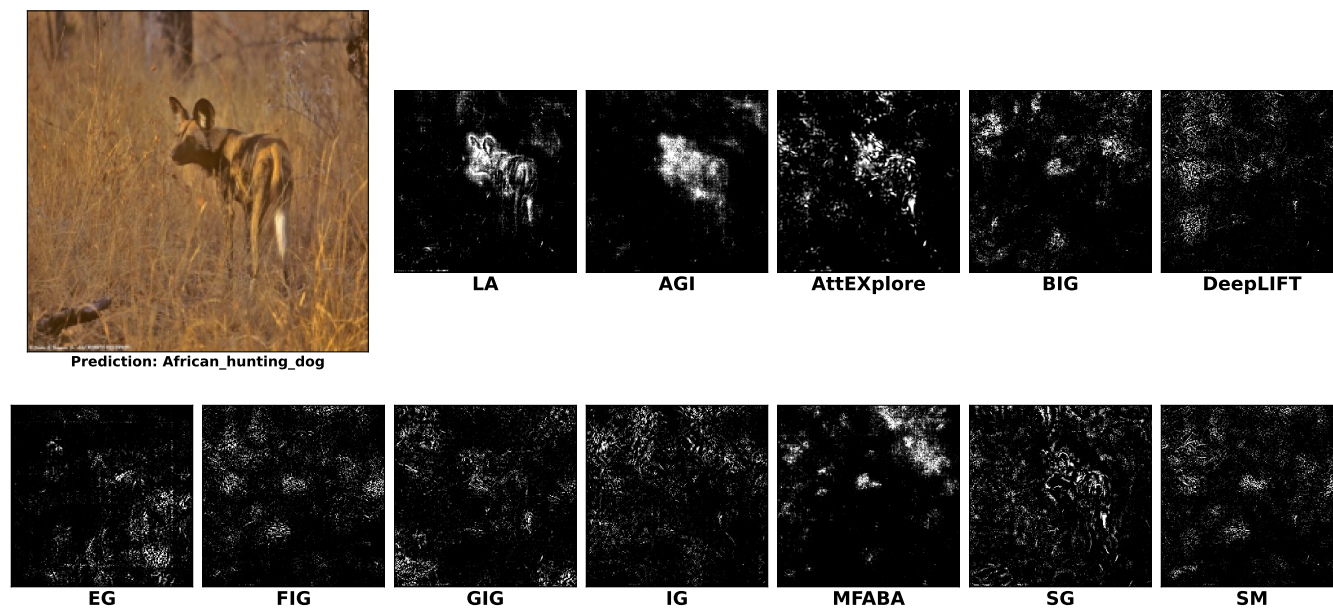


Figure 13: Attribution Results on the MaxViT-T

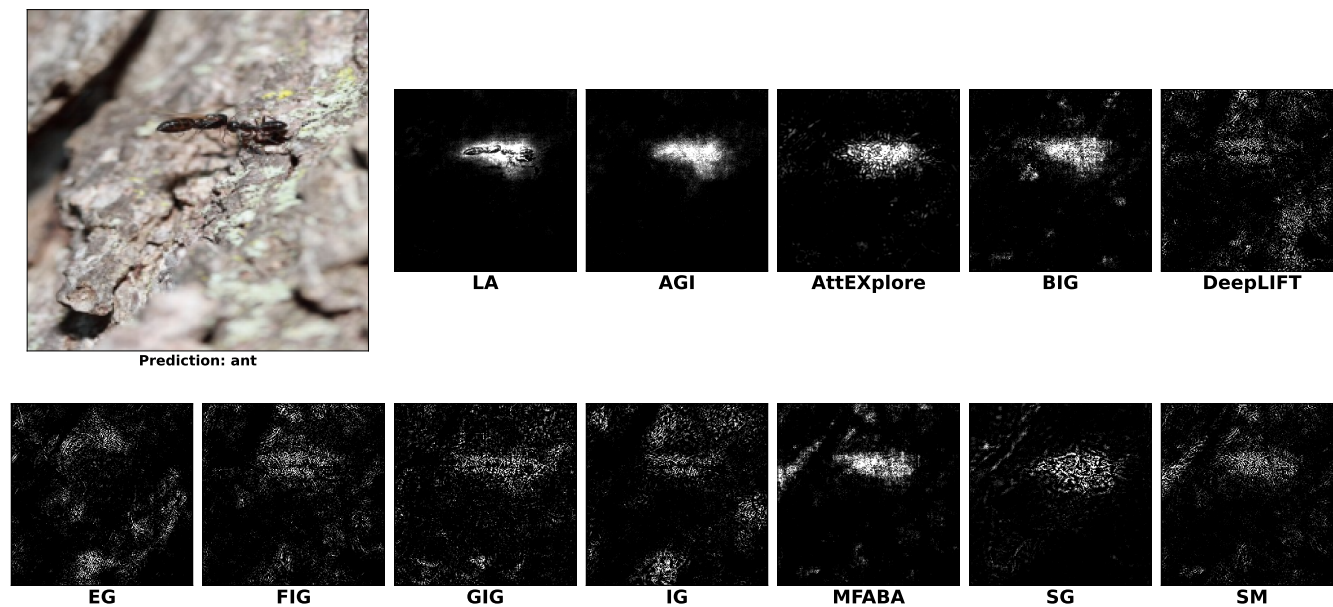


Figure 14: Attribution Results on the MaxViT-T