A Appendix

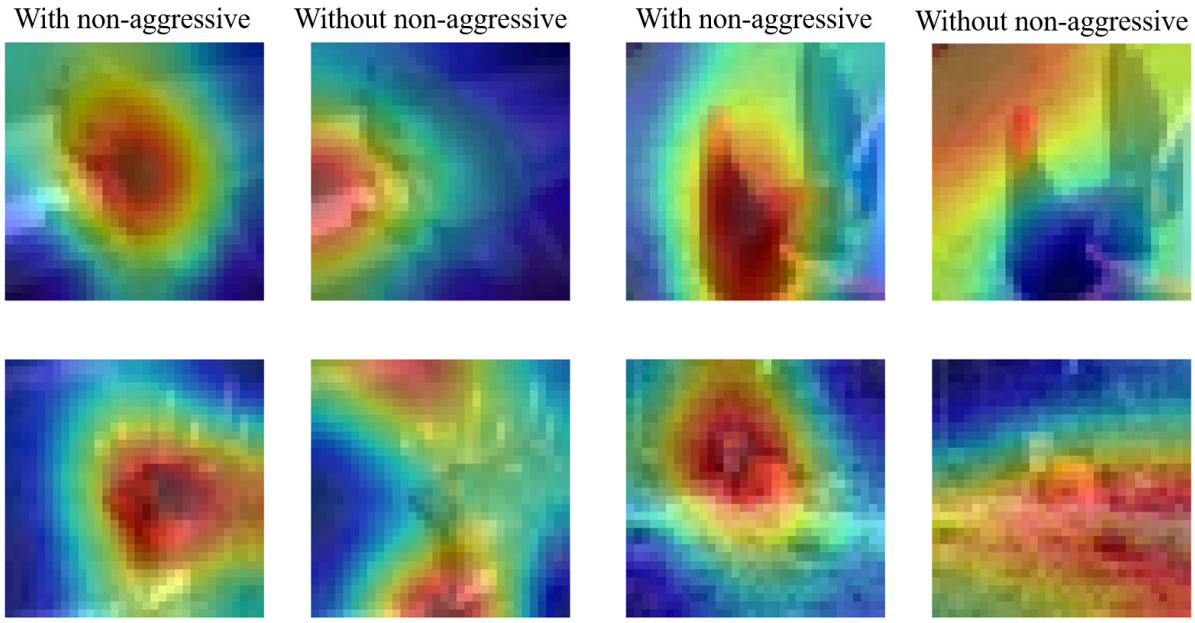
In this supplementary material, we have conducted the experiments to support our analysis in Section. **In-depth analysis** and Section. **Evaluation**. We show the details about the specific performance and relevant schematic diagrams of our MFABA algorithm under different conditions.

A.1 Impact of non-aggressive samples on MFABA algorithm

A.1.1 Heat map analysis of figure 2 in our paper

In Figure. 2 of the main paper, we have visualized the results with and without the non-aggressive sample for the linear gradient ascending path in MFABA. We can see that, the left diagram is better than right one, supporting our analysis in Section. **Definition of Aggressiveness**.

A.1.2 Additional figures for Comparison of heatmap with and without non-aggressive samples



A.2 Comparison of MFABA-cosine and MFABA-norm

A.2.1 Results for BIG Algorithm, and MFABA Approximation Algorithms about Eq. 18 and Eq. 19

In Figure. 1, we compare the results of BIG, MFABA with cos method and MFABA with norm method separately. For the cos and norm methods, they can be referred to Eq. 18 and 19. Figure 1 demonstrates the linear MFABA is able to approximate BIG algorithm and has less noise in the attribution output, supporting our analysis in Section. **Comparision with Other State-fo-the-art Methods**.

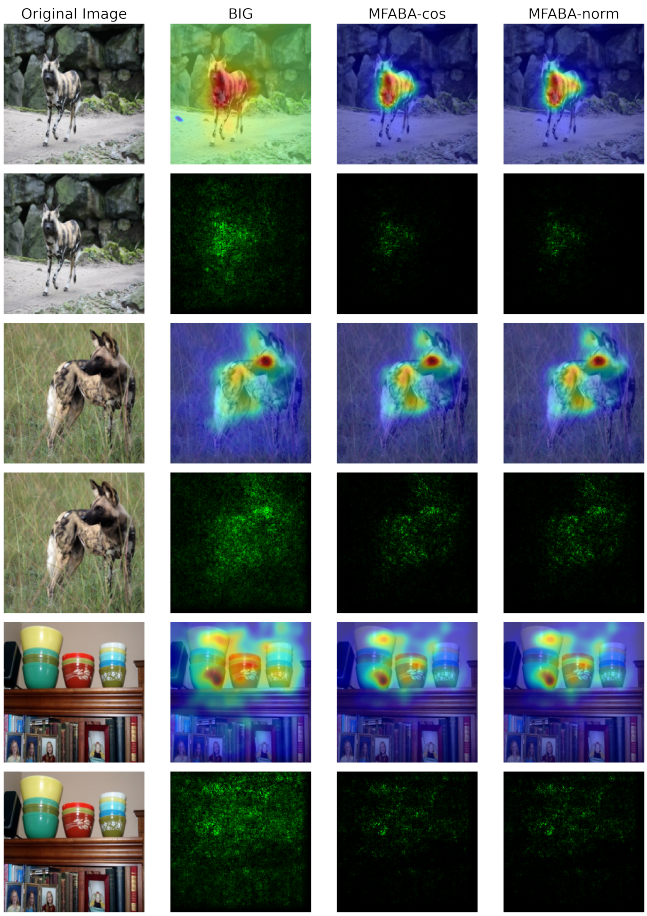
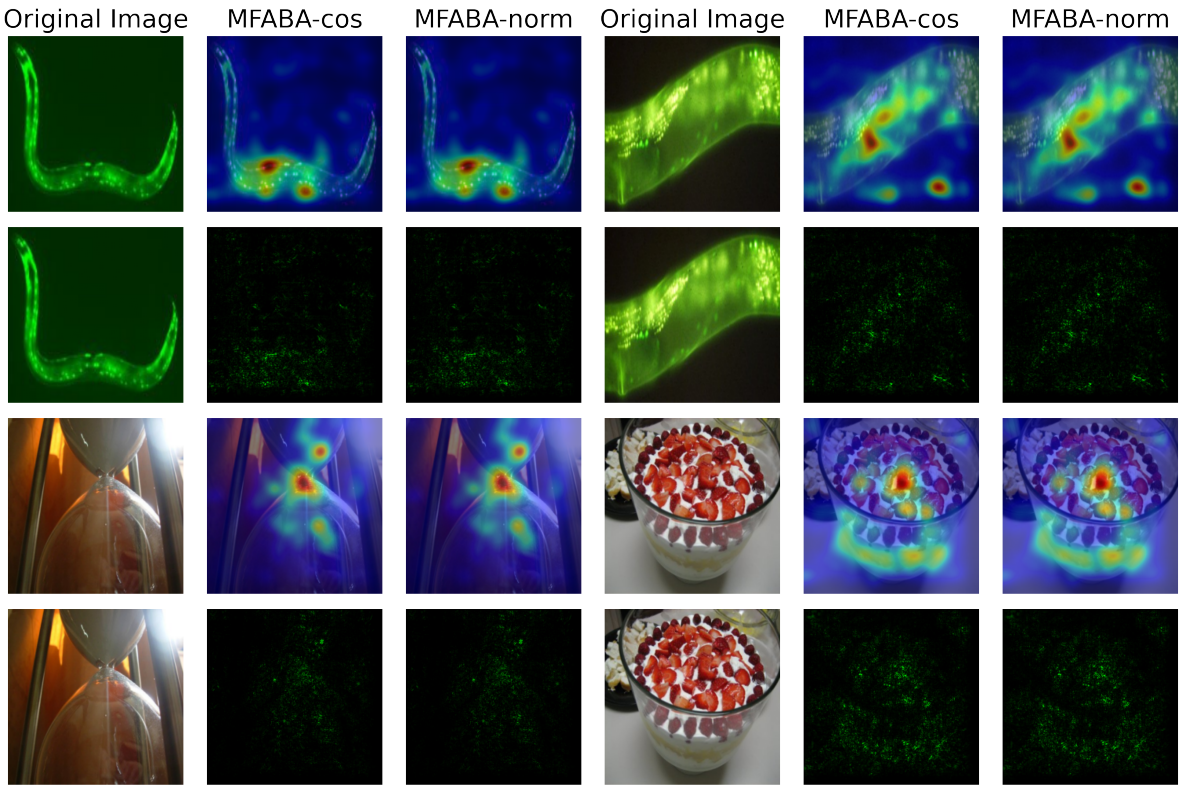


Figure 1: Results for BIG Algorithm, and MFABA Approximation Algorithms about Eq. 18 and Eq. 19

A.2.2 Additional figures for BIG Algorithm, and MFABA Approximation Algorithms about Eq. 18 and Eq. 19



A.3 Evaluation of MFABA on Imagenet Dataset

A.3.1 Evaluation of MFABA with and without the softmax on Imagenet Dataset

In Figure 2, we can see that the fitness of softmax is higher, which provide better visualization results and support our analysis in Section. **Attribution Method in MFABA**.

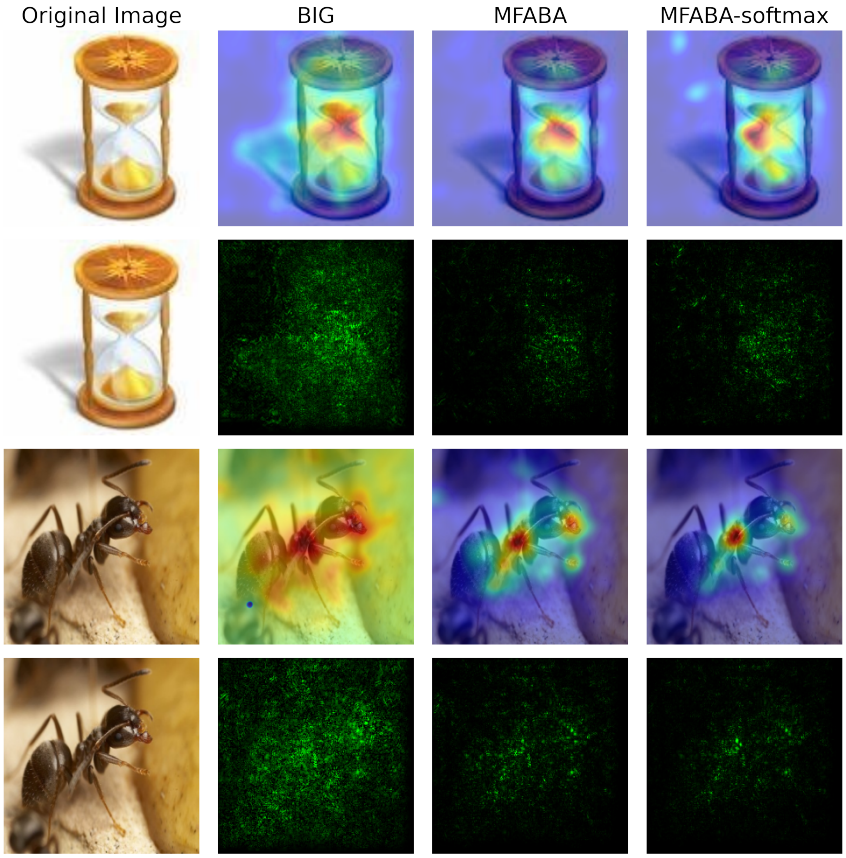
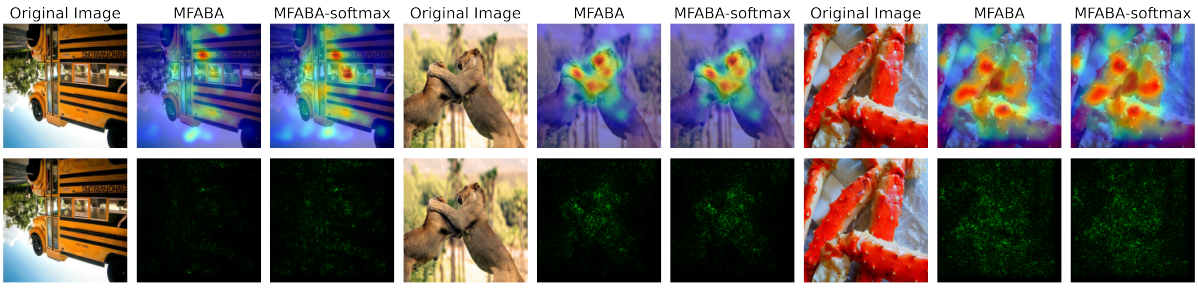


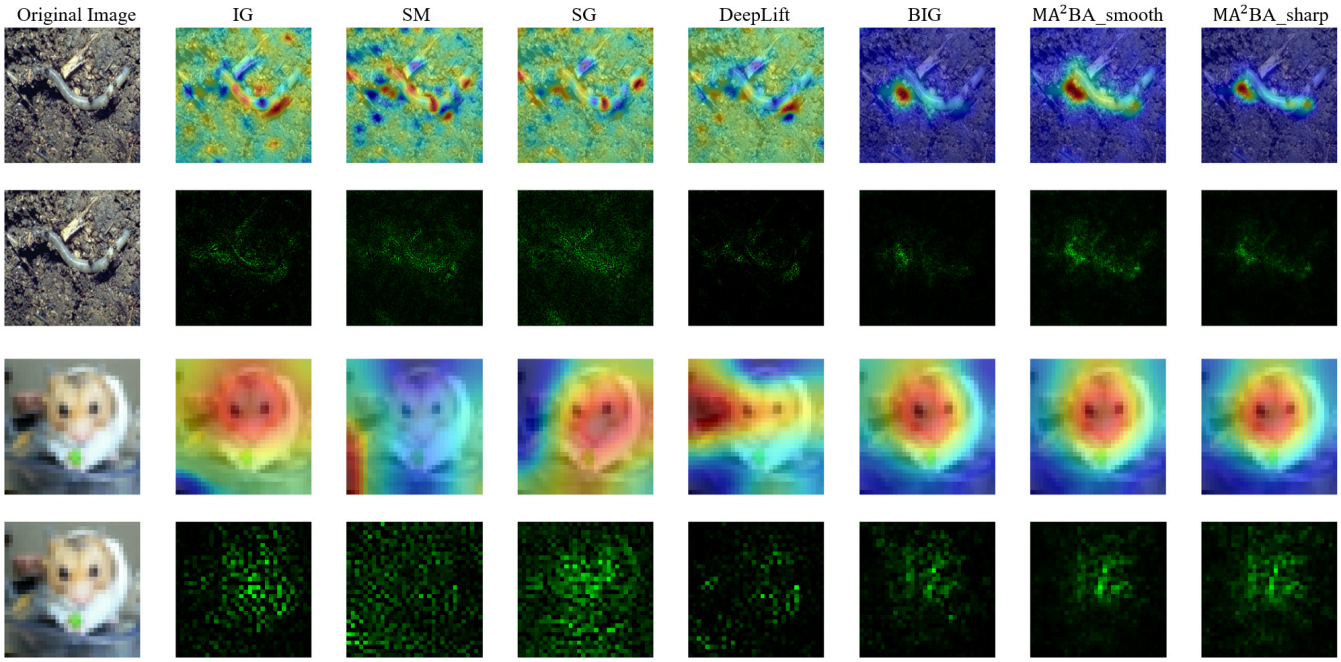
Figure 2: Evaluation of MFABA with and without the softmax for Imagenet Dataset

A.3.2 Additional figures for Evaluation Results of MFABA with and without Softmax for Imagenet Dataset



A.4 Additional figures for attribution results of MFABA in comparison with other state-of-the-art methods

In this section we provide more quantitative visualization results (the figure below and figure. 3), supporting our analysis in Section. **Empirical Evaluation**.



A.5 Algorithm Procedure

As shown in the following pseudo code, the gradient ascent method is used to line up adversarial data pairs, and each gradient ascent will return the corresponding gradient degree $\frac{\partial L(x_j)}{\partial x_j}$, the attack result x_j and the incoming label y'. y' is the new label for the sample to determine whether the gradient ascent will be stopped. The attribution $\sum_{j=0}^{n-1} \frac{\left(\frac{\partial L(x_j)}{\partial x_j}+\frac{\partial L(x_{j+1})}{\partial x_{j+1}}\right)}{2}\left(x_{j+1}^i - x_j^i\right)$ is then given according to Algorithm 1, which optimises the summation rate by computing dot product and finally returns the corresponding attribution result.

Algorithm 1 MFABA(f,m,x_0,y)

Input: model m,target f,input x_0,label y,Iterative number n,learning rate η,method
$X = [x_0]$, $grads = []$, $j = 1$
for $j \leq n$ **do**
    $grad = \frac{\partial L(x_{j-1})}{\partial x_{j-1}}$
    $x_j = x_j + method(grad)$
    $x_j \in R^{w \times h \times 3} grad \in R^{w \times h \times 3}$
    $x.append(x_j)$
    $y' = m(x)$
    $grads.append(grad)$
    $j = j + 1$
    if $y' \neq y$
        break
end for
$grads.append(\frac{\partial f(x_j)}{\partial x_j})$
$addr = -(x[1:] - x[:-1]) \cdot \frac{(\text{grads}[:-1]+\text{grads}[1:])}{2}$
return addr

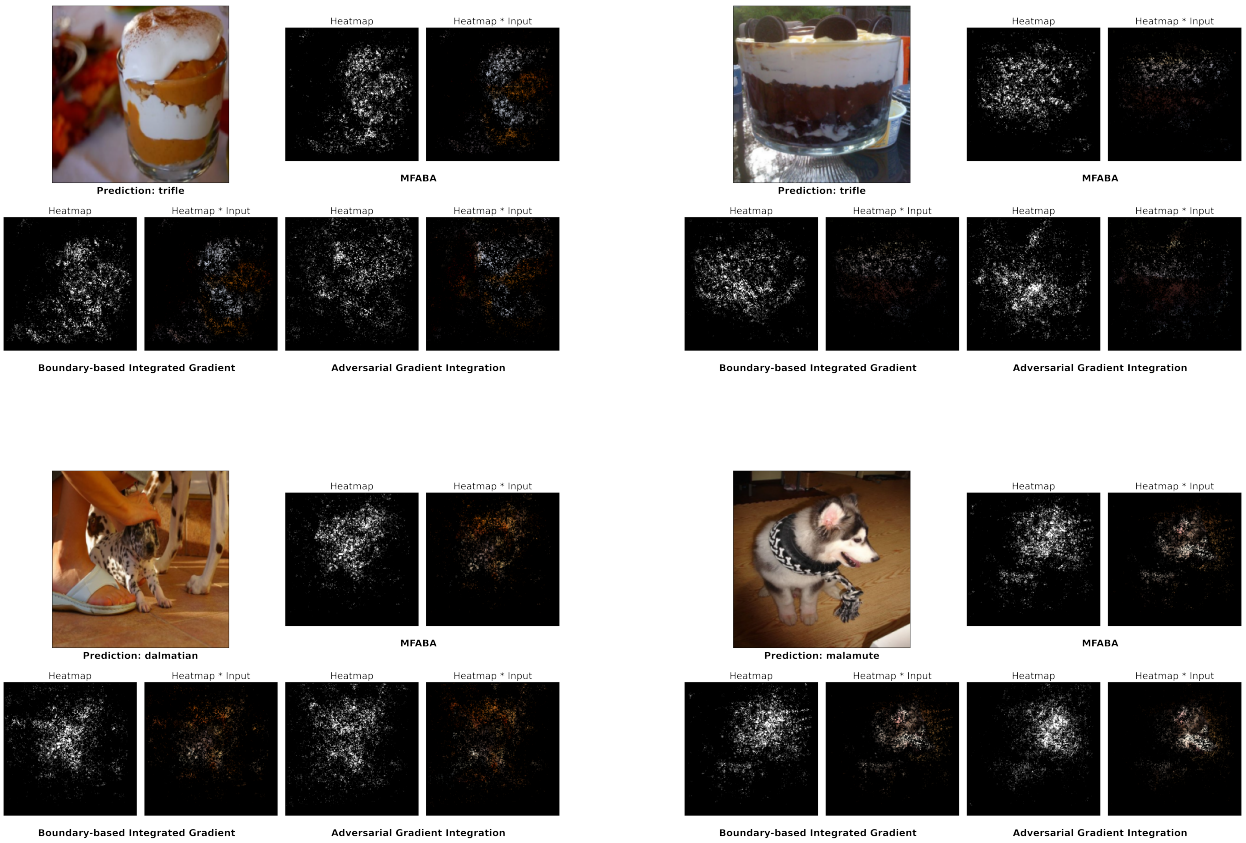


Figure 3: Additional figures for attribution results of MFABA in comparison with other state-of-the-art methods