METRIC-DRIVEN ATTRIBUTIONS FOR VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attribution algorithms explain computer vision models by attributing the model response to pixels within the input. Existing attribution methods generate explanations by combining transformations of internal model representations such as class activation maps, gradients, attention, or relevance scores. The effectiveness of an attribution map is measured using attribution quality metrics. This leads us to pose the following question: if attribution methods are assessed using attribution quality metrics, why are the metrics not used to generate the attributions? In response to this question, we propose a Metric-Driven Attribution for explaining Vision Transformers (ViT) called MDA. Guided by attribution quality metrics, the method creates attribution maps by performing patch order and patch magnitude optimization across all patch tokens. The first step orders the patches in terms of importance and the second step assigns the magnitude to each patch while preserving the patch order. Moreover, MDA can provide a smooth trade-off between sparse and dense attributions by modifying the optimization objective. Experimental evaluation demonstrates the proposed MDA method outperforms 7 existing ViT attribution methods by an average of 12% across 12 attribution metrics on the ImageNet dataset for the ViT-base 16×16, ViT-tiny 16×16, and ViT-base 32×32 models.

1 INTRODUCTION

Computer vision models have been broadly adopted with applications in domains such as self-driving cars (Ando et al., 2023), surveillance (Ho et al., 2019), and disease detection (Esteva et al., 2021). However, due to the black-box nature of these models, proper explanations must be available to incite trust in their safety-critical system deployments. Effective explanation methods have been thoroughly explored for Convolutional Neural Networks (CNNs), but are still in their infancy for the more prominent Vision Transformers (ViTs). The most popular attribution methods create human-readable explanations using a model's internal information such as class activation maps (Selvaraju et al., 2017), gradients (Simonyan et al., 2014; Sundararajan et al., 2017), attention (Vaswani et al., 2017), or relevance scores (Binder et al., 2016). A less abstract method of creating explanations is through feature perturbation. Here, the model internals are ignored and input features are masked to discover which are the most relevant to the models prediction. However, this class of method has been shown to be prohibitively slow for creating the per-pixel explanations required for CNNs (Fisher et al., 2019; Zeiler & Fergus, 2014). Perturbation-based methods have only recently seen a resurgence for ViTs (Xie et al., 2023; Englebert et al., 2023) whose reasoning over coarse-grained patches allows the runtime complexity to be controlled. The broad diversity of attribution algorithms necessitates quantitative comparison using metrics for accurate quality assessment.

The most common attribution quality metrics are perturbation-based tests (Petsiuk et al., 2018; Kapishnikov et al., 2019; Walker et al., 2024) which do not need ground-truth labels. The most commonly used perturbation-based tests are insertion and deletion (Petsiuk et al., 2018). These metrics ablate the model input in order of highest value attribution features to determine if the features are as important for the model's decision as indicated. These tests were extended into magnitude aligned scoring (MAS) which also assess the magnitude of the attribution assigned to each feature (Walker et al., 2024). A feature's magnitude is evaluated by measuring the alignment with the feature's model importance. This raises the question: can we avoid abstract explanation approaches and use these perturbation metrics directly to create better explanations?

In this paper, we propose a method for creating Metric-Driven Attributions for the ViT called (MDA). The approach creates high-quality explanations for ViTs using a perturbation-based approach that mimics the methodology of the attribution quality metrics used to assess the specified attributions. The four main contributions of MDA can be summarized, as follows:

1. We provide the first attribution algorithm that is directly driven by the metrics used to assess the attributions produced by the algorithm. This is a stark departure from existing approaches that mainly leverage internal model representations and methods.
2. The MDA method creates attributions using patch ordering optimization and patch magnitude optimization. The first step orders the patches in terms of importance while the second step assigns an attribution magnitude to each patch while preserving the patch order.
3. Existing metrics encourage sparse explanations that focus on a few important features within an image. We demonstrate that MDA is capable of creating dense attributions that indicate all features of importance by slightly modifying the optimization objective.
4. MDA achieves improved quantitative and qualitative performance compared to 7 state-of-the-art ViT attribution methods. On ImageNet, MDA outperforms the SOTA not only for the perturbation metrics it was optimized for, but also six additional attribution metrics across three ViT models with an average 12% improvement across all metrics.

The remainder of the paper is organized as follows: the related work is in Section 2, the methodology is in Section 3, the experiments are in Section 4, and the conclusion is in Section 5.

2 RELATED WORK

In this section, we first present a detailed explanation of the scoring objectives of the perturbation quality metrics of interest. We then discuss the existing approaches to ViT explanation.

2.1 PERTURBATION-BASED ATTRIBUTION QUALITY METRICS

Attribution metrics fit into two groups: comparison against ground-truth masks (Borji et al., 2013) or perturbation of the model (Adebayo et al., 2018; Hooker et al., 2019) or the input (Petsiuk et al., 2018; Walker et al., 2024) without ground-truth. Model perturbation randomizes or retrains the entire model to evaluate at scale, which is a slow process. Input perturbation is therefore the most popular because it can evaluate single images without ground-truth or model changes. The insertion and deletion perturbation metrics expect the modification of an important feature in the model input to cause a change in the output (Petsiuk et al., 2018). They measure how well an attribution indicates important features by performing a linear perturbation process between a baseline and an input. We explain insertion with the following text and Figure 1. Given an attribution A, the input X of size D^2 belonging to class c, and the model F, the test evaluates A over N steps. Let P_k denote the perturbed image at step k. For insertion, P_0 denotes the blurred input baseline and P_N is the original input. In each step, the set of $\frac{D^2}{N}$ i-th highest magnitude features $A_i \notin P_k$ are selected, seen in Figure 1 (a). These features correspond to the ith most important pixels $X_i \notin P_k$, and we update P_k as $P_{k+1} = P_k + X_i$, removing the blurring as seen in Figure 1 (b). The model response is defined as:

$$MR_k = \text{softmax}(F(P_k))_c. \tag{1}$$

Over the N steps of this insertion (ins) perturbation test, the monotonically non-decreasing MR^{ins} curve will be formed on the range $[0, 1]$ as seen in Figure 1 (c). The score of an attribution is then calculated as the area under the MR curve (AUC):

$$score = \frac{1}{N+1}\sum_{k=0}^{N} MR_k. \tag{2}$$

A higher MR^{ins} AUC indicates a better attribution. The AUC in Figure 1(c) is 0.896. For deletion (del), P_0 denotes the original image baseline and P_N is a black image. The MR^{del} curve formed from the test is monotonically non-increasing and has range $[1, 0]$. A lower MR^{del} AUC is better.

The MAS insertion and deletion tests calculate the MR in the same manner (Walker et al., 2024). These tests additionally introduce the density response (DR) curve which measures the magnitude

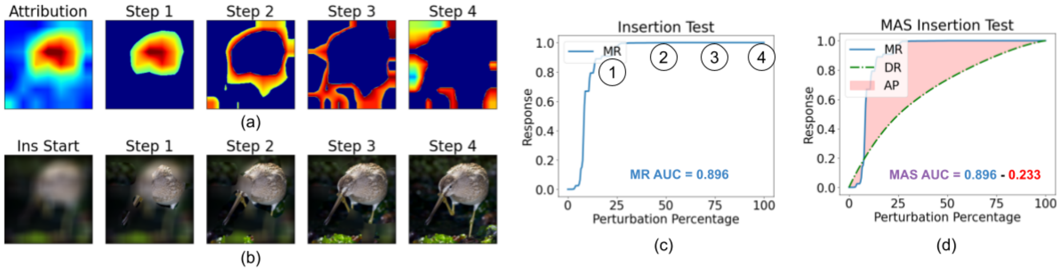


Figure 1: An illustration of the insertion metric where an attribution is evaluated over 4 steps. In each step, the highest attribution values are chosen (a) and those pixels are unblurred (b). The graph (c) illustrates the MR generated by this process as described in Eq (1) and the AUC score from Eq (2). The graph in (d) shows the MR, DR, and AP as described in Eq (3) and the MAS AUC: $MR - AP$.

of each feature. The DR and MR are compared to determine if high and low value attributions are assigned to important and unimportant features, respectively. In insertion, DR tracks what percentage of the total attribution of A has been inserted at step k, and is defined as:

$$DR_k^{ins} = \frac{\sum_{i=0}^{k} |A_i|}{\sum_{i=0}^{N} |A_i|}, \tag{3}$$

where $|.|$ denotes absolute value. DR^{ins} is monotonically increasing and on the range $[0, 1]$, seen as the dashed line in Figure 1 (d). The difference $|MR - DR|$ is called the alignment penalty (AP), seen in Figure 1 (d). The MAS score is then ($MR^{ins} - AP$) AUC, seen as $0.896 - 0.223$ in Figure 1 (d). The AP indicates important (unimportant) features were given too low (high) of a magnitude in the attribution. In deletion, pixels are removed each step, so $DR_k^{del} = (1 - DR_k^{ins})$, the DR^{del} curve is monotonically decreasing and ($MR^{del} + AP$) AUC is the score.

2.2 VISION TRANSFORMER EXPLANATIONS

The ViT model (Dosovitskiy et al., 2020) embeds the input image as a linear array of patch tokens prepended with a classification token [CLS]. These are passed through a decoder-only transformer (Vaswani et al., 2017) which extracts important relationships between all tokens via self-attention, and a final linear layer classifies the output embedding of the [CLS] token. A ViT explanation assigns one value to each of the M^2 patches. For ViT 16×16, the 224×224px input is broken into 14^2 patches which are each 16×16px. The resulting 14^2 grid of attribution values is upscaled to 224×224px.

ViT explanations can be derived from the self-attention values of the last layer's [CLS] token (Vaswani et al., 2017), or the matrix multiplication of the self-attention through all layers (Rollout) (Abnar & Zuidema, 2020). However, attention does not contain class information and is a poor explanation alone (Bastings & Filippova, 2020). GradCAM (GC) multiplies the last layer self-attention by its gradients (Selvaraju et al., 2017). Integrated Gradients (IG) generates an explanation from the mean attention gradients of interpolated inputs between a baseline and input. Current state-of-the-art methods use a mixture of these techniques. Transformer Attribution (T-Attr) (Chefer et al., 2021a) gathers the attention of all layers with layer-wise relevance propagation (Binder et al., 2016) and multiplies by the gradients. Transition Attention (T-Attn) (Yuan et al., 2021) multiplies Rollout by IG. Bidirectional Attention (Bi-Attn) (Chen et al., 2023) introduces head importance to Rollout and multiplies by IG. Attributions generated by internal model representations are valuable, but most methods sacrifice their theoretical foundations by applying ad-hoc post-processing in this manner.

Due to the ViT patch structure, two perturbation methods show promise without using internal model representations. ViT-CX (Xie et al., 2023) creates a set of masked inputs from the model's output embeddings. These are scored by the model's softmax output and the weighted sum of all masks is the explanation. Transformer Input Sampling (TIS) (Englebert et al., 2023) performs the same process, but masking occurs inside the model at the token embedding level instead of the input level.

None of these existing approaches use the attribution quality metrics to generate attributions. By leveraging the metrics' ability to measure attribution quality, we propose to generate attributions that optimize the metric scores, creating high-quality attributions from the source of their evaluation.

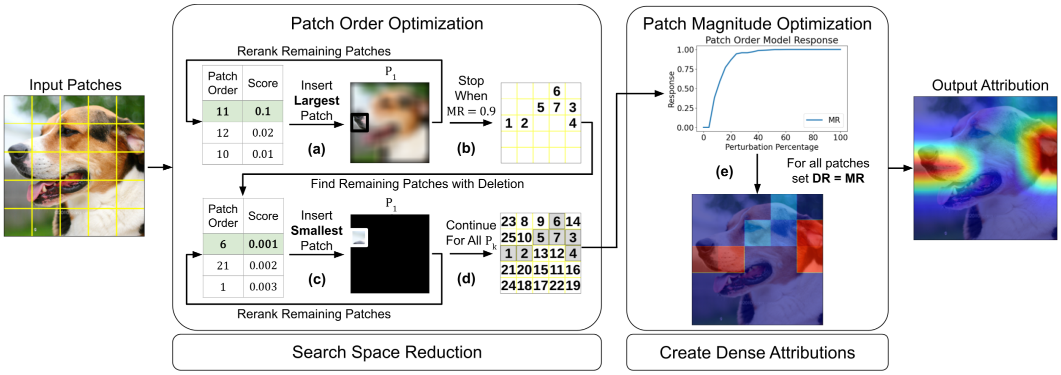


Figure 2: An overview of the MDA method. First, the insertion process finds the patch (X_i) which induces the highest $MR(P_{k+1})$ at each step k (a) until the MR reaches $\tau = 0.9$ (b). This produces the minimum ordering needed for good insertion performance. The remaining patches are found through deletion, where the patch which induces the lowest MR is selected each step (c). This yields a single, optimized ordering of the patches (d). Next, the produced MR informs the magnitude of each patch to be set such that $MR = DR$ (e) (see Figure 3 for details). This yields a sparse, high-scoring attribution. The final attribution is upscaled to the image size.

3 METHODOLOGY

We propose MDA, the first attribution method which generates explanations by directly maximizing attribution quality metric scores. The framework consists of four components: two core optimization steps which generate an attribution and two add-on steps for calibrating and speeding up the process. First, we order the patches to maximize insertion and deletion scores. Second, we set the magnitude of each patch to maximize the MAS insertion and deletion scores. The initial two steps produce sparse attributions with high qualitative scores. Third, we introduce a sliding parameter to allow the creation of dense attributions at the expense of a lower score. Fourth, we present techniques to control the runtime complexity. An overview of the framework is illustrated in Figure 2.

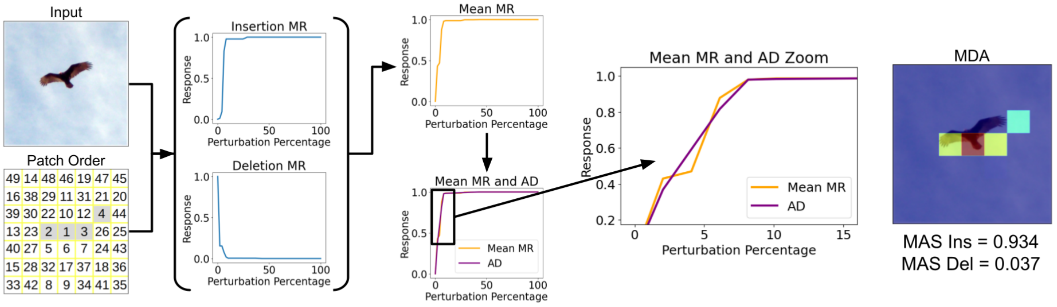
3.1 PATCH ORDER OPTIMIZATION

In this section, we describe our patch order optimization technique that orders the M^2 patches of the ViT in terms of model importance, which is illustrated in (a) through (d) of Figure 2. The objective of the optimization is to determine the patch order $\mathcal{I}$ which maximizes the insertion and deletion scores. Recall from Eq (2) that the insertion and deletion scores are only a function of the patch order. A potential challenge is that the ideal patch order for insertion may be conflicting with the ideal patch order for deletion. Therefore, our approach attempts to determine a patch order which maximizes the joint insertion and deletion scores. We observe insertion is only sensitive to the order of the first few patches whereas deletion is sensitive to the order of all patches. We utilize this by first determining the patches needed for a strong insertion score, and then find the remaining patches using deletion. We now describe the method for optimizing the insertion and deletion patch orders.

Insertion-Based Patch Ordering: Starting with the blurred input image P_0, the MDA framework finds the patch X_i that produces the strongest model response and inserts it into the patch order $\mathcal{I}$, which is illustrated in Figure 2(a). The process of iteratively inserting a patch mimics the method for computing the insertion score metric and is therefore expected to score well in terms of insertion. Formally, the patch X_i is found directly from the model response as follows:

$$\arg\max_{i \notin \mathcal{I}} MR(P_k + X_i) \tag{4}$$

where i is the index of any patch X_i not already inserted in P_k. Next, P_{k+1} is set to $P_k + X_i$. This process of inserting patches X_i and updating P_k is repeated until C patches have been inserted, where $MR(P_C) \geq \tau \cdot MR(P_N)$ with $N = M^2$. In practice, we set the parameter τ to 0.90. The patches determined by the insertion-based ordering process are shown in Figure 2(b). Next, the remaining $M^2 - C$ patches are determined using the deletion-based patch ordering process.



Measure MR Ins and Del
(a)

Find the Mean MR and solve for AD
(b)

Set DR = AD
(c)

Figure 3: Given the original image and the patch order, we find MR^{ins} and MR^{del} (a). MR^{del} is transformed and averaged with MR^{ins} to find MR^{mean} and a quadratic solver finds the new curve AD from MR^{mean} (b). Note in the zoom, it can be observed that AD is monotonically increasing and has a strictly monotonically decreasing derivative. The AD curve is used to assign patch magnitudes, producing a high MAS score, attribution (c).

Deletion-Based Patch Ordering: Now, we address the deletion test which transitions from the original image with a high model response to a black image with a low model response. Ideally, each deletion generates a decrease in the model response and patches which contain the most important class information will evoke the smallest model responses (a large decrease in model response). Intuitively, it may seem obvious to find the patch which creates the largest decrease in the model response at every step. However, in deletion, *all important features must be removed* to have a significant decrease in the model response. One ablated patch on an important feature may not remove enough information to change the model response. We therefore reverse the deletion problem to be a modified insertion problem and aim to *minimize* the insertion score.

In this new problem, we begin with a black image of low model response and move towards the original image of high model response. At every step of the perturbation process, we find the patch that has the lowest model response:

$$\arg\min_{i \notin \mathcal{I}} MR(P_k + X_i). \tag{5}$$

The resulting order sorts patches from least to most important for the model's decision, therefore we reverse this to have the ideal deletion order. As seen in Figure 2, the C patches found from insertion are an input to deletion and locked as the C most important patches. Figure 2 (c) shows that the patches of least importance are modified first, and the final patch order which creates the best possible deletion score is found in Figure 2 (d).

3.2 PATCH MAGNITUDE OPTIMIZATION

The patch order $\mathcal{I}$ from the previous optimization step maximizes the insertion and deletion scores. Therefore, the objective of the patch magnitude optimization is to minimize the MAS alignment penalty while preserving the patch order $\mathcal{I}$, thus maximizing the MAS score. Let $\mathcal{I}_i$ and A_i denote the ordering index and attribution magnitude of a patch i, respectively. We wish to assign a magnitude to each patch such that the order $\mathcal{I}$ is respected as follows:

$$A_i > A_j \quad \textit{iff} \quad \mathcal{I}_i < \mathcal{I}_j, \quad \forall i \neq j. \tag{6}$$

The patch order $\mathcal{I}$ inherently defines the curves MR^{ins} and MR^{del}, seen in Figure 3 (a). To minimize the alignment penalty AP, the density response DR should be set to be equal to both the model response MR^{ins} and MR^{del}, which is impossible if MR^{ins} and MR^{del} are not identical. Therefore, we first combine MR^{ins} and MR^{del} into MR^{mean} as:

$$MR^{mean} = \frac{(MR^{ins} + (1 - MR^{del}))}{2}, \tag{7}$$

where $(1 - MR^{del})$ transforms MR^{del} to increase like MR^{ins}, as seen in Figure 3 (b). Now, we can perform $DR = MR^{mean}$ by setting the attribution value of each patch A_i as follows:

$$A_i = MR_i^{mean} - MR_{i-1}^{mean}, \tag{8}$$

where $A_1 = MR_1^{mean} - MR_0^{mean}$ and $MR_0^{mean} = 0$. However, the assignment of the attributions using Eq (8) does not ensure that the initial patch order would be preserved, i.e., the constraint on the relative ordering of the patches in Eq (6) may be violated, which would in turn change the model response and insertion and deletion scores. For example, if the patch ordering $\mathcal{I} = \{1, 2, 3\}$ generated the model response $MR^{mean} = [0, 3, 3, 4]$, then Eq (8) yields $A_1 = 3$, $A_2 = 0$, and $A_3 = 1$. These assigned attributions result in a revised patch order $\bar{\mathcal{I}} = \{1, 3, 2\}$ which is inconsistent with the initial ordering of the patches. We explain this behavior in the following paragraph.

To determine the attributions using Eq (8), we propose to specify a new attribution definition curve, AD. The curve should be similar to MR^{mean} to minimize the alignment penalty. However, we require the curve to satisfy two additional properties which ensure the patch order is preserved, i.e., Eq (6) is satisfied.

1. **Property 1:** The curve AD must be monotonically increasing.
2. **Property 2:** The derivatives of the curve AD must be strictly monotonically decreasing.

The properties are from two observations. 1) All attributions A_i are positive, so AD is, by definition, monotonically increasing. 2) From Eq (8), it can be understood that AD'_i = A_i; therefore, to preserve the order for all A_i, it is implied that AD' must be strictly monotonically decreasing.

We formulate the problem of specifying the curve AD, as follows:

$$\begin{aligned} &\text{Minimize } \frac{1}{N+1}\sum_{i=0}^{N}(AD_i - MR_i^{mean})^2 \\ &\text{subject to } \begin{cases} AD'_{i+1} \leq AD'_i, & \forall i \in [0, N] \\ AD_i \in [0, 1], & \forall i \in [0, N] \\ AD_0 = 0 \text{ and } AD_N = 1 \end{cases}. \end{aligned} \tag{9}$$

The objective is to minimize the quadratic distance between AD and MR^{mean} under the linear derivative constraint. This is clearly a quadratic programming problem and a quadratic solver is used to find the solution. Two additional constraints are placed on the range of the function and the endpoints to ensure the final result does not escape the original model response definition.

One final problem remains after finding AD. Due to the $\leq$ limitation of the quadratic programming problem, AD has a monotonically non-increasing derivative, so Eq (6) is not satisfied. For example, $AD = [0, 2, 4]$ gives $A_1 = 2$ and $A_2 = 2$, implying $\mathcal{I}_1 = \mathcal{I}_2$ which is impossible. We introduce a patch order term to Eq (8) to induce the monotonically decreasing derivative property:

$$A_i = \Delta AD_i + (\Delta AD_i * \frac{N-i}{N}), \tag{10}$$

where $\Delta AD_i = AD_i - AD_{i-1}$. The additional term ensures $A_i > A_{i+1}$ where $AD'' = 0$. For large N, this second term does not increase the AP significantly. We visualize the magnitude optimization in Figure 3 (a) to (c). The final $M \times M$ attribution is upscaled to the input image size.

3.3 MAGNITUDE OPTIMIZATION FOR DENSE ATTRIBUTIONS

The MAS metric for evaluating attributions encourages the creation of sparse attributions. We call a set of attributions sparse if they indicate the minimum amount of information needed for accurate model prediction. On the other side of the spectrum, dense attributions indicate all features of an object which are relevant to accurate prediction. We provide an *optional modification* to MDA which allows a user to specify if the generated attributions are sparse, dense, or a mixture of the two.

Given A^{sparse} as defined from Eq (10), we find the dense attribution A^{dense} by assigning all important patches a magnitude which represents their order:

$$A_i^{dense} = \begin{cases} \frac{N-i}{N} & \text{if } A_i^{sparse} \geq \kappa \\ A_i^{sparse} & \text{if } A_i^{sparse} < \kappa \end{cases}. \tag{11}$$

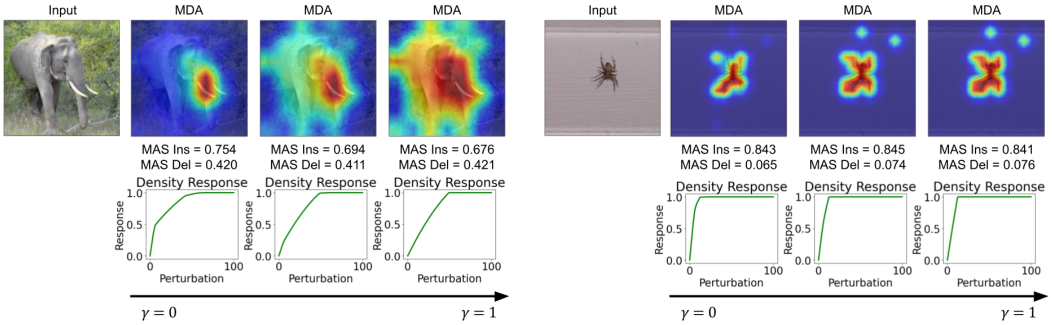


Figure 4: We illustrate MDA for the transition of $\gamma = 0$ to $\gamma = 1$. For the elephant, the attribution grows as the magnitude is more evenly spread among features (seen by the more linear DR), indicating only the tusks are needed for classification. For the spider, the DR and attribution do not greatly change, indicating all features of the spider are important for classification.

The κ parameter controls the model importance cutoff since $A_i \approx \Delta MR_i$. In practice, we employ $\kappa = 0.005$ to only strongly attribute patches with more than 0.5% model importance.

We now introduce a user-controlled parameter γ to blend between A^{sparse} and A^{dense} where $\gamma = 0$ and $\gamma = 1$ yield sparse and dense attributions, respectively:

$$A^{\gamma} = (1 - \gamma)A^{sparse} + \gamma A^{dense}. \tag{12}$$

In Figure 4 we illustrate this for a large and small subject. We see for the large subject, the attribution expands to cover all features, resulting in a more linear density response. For the small subject, the attribution does not expand, indicating all important features were already used for classification. A user can tune γ to their choosing, but, quantitatively, the best explanation is created with $\gamma = 0$.

3.4 SEARCH SPACE REDUCTION

In this section we provide solutions to improving the runtime of MDA. A naive implementation of the optimization search over the M^2 coarse-grained ViT patches results in a $\mathcal{O}(M^4)$ runtime. This optimization problem resembles the knapsack problem, but we cannot reduce the search space through common techniques such as memoization or tabulation. In this problem, unlike knapsack, the next optimal value $MR(P_{k+1})$ is unknown and cannot be assumed to be additive, i.e. $MR(P_1 + X_i) \neq MR(P_1) + MR(P_0 + X_i)$ where each X_i was previously enumerated in the search for P_1.

As typical techniques will not work, we take heuristic approaches to reduce the search space, approaching the absolute minimum $\mathcal{O}(M^2)$ runtime. First, the insertion optimization phase is improved. As we terminate the insertion optimization when we find the C patches needed to produce an $MR = 0.9$ we reduce the search space to $\mathcal{O}(CM^2)$. Across 100 images, only $48/196$ and $6.5/49$ patches were needed to reach this model response for the ViT-base 16×16 and ViT-base 32×32 models, respectively, resulting in $\mathcal{O}(M^3)$ insertion search space. For deletion, the order of all patches is needed, so an early stopping technique cannot be employed, yielding a runtime of $\mathcal{O}(M^4)$. However, we can prune the search space of both insertion and deletion.

We can take advantage of the existing ViT attribution method space, and use a SOTA explanation to provide an initial ordering of the patches in the image. We choose Bi-Attn (Chen et al., 2023) due to its low computational cost and strong performance among other ViT methods. For insertion, the input attribution provides an ordering of all patches from high to low magnitude. We use this ordering to inform a smaller search space, i.e. instead of searching through all $M^2 - k$ patches not inserted in each step k, we search through the first $2M$ patches in the ordering which we can assume are valuable to the model. For deletion, the input attribution provides an ordering from lowest to highest value and we search through the first $2M$ patches in the ordering which we can assume are not valuable to the model. These two steps combined reduce the runtime from $\mathcal{O}(M^4)$ to $\mathcal{O}(M^2) + \mathcal{O}(M^3) = \mathcal{O}(M^3)$ and finally, we can approach $\mathcal{O}(M^2)$ with batched execution. The runtime could be reduced to the minimum $\mathcal{O}(M^2)$ by selecting M patches at every insertion step instead of 1, but this requires better attributions than what exists to avoid a significant quality penalty. We now evaluate the proposed MDA method with $\gamma = 0$ and Bi-Attn (Chen et al., 2023) as input.

4 EXPERIMENTAL EVALUATION

In this section we present qualitative and quantitative evaluations of the proposed MDA ViT attribution method. We perform all experiments using PyTorch (Paszke et al., 2019) and use the ImageNet 2012 validation dataset (Russakovsky et al., 2015) and the ImageNet Segmentation dataset (Guillaumin et al., 2014). We employ three ViT models: ViT-base 16×16, ViT-tiny 16×16, and ViT-base 32×32 as defined in the ViT paper (Dosovitskiy et al., 2020). These models are trained on ImageNet and are from the PyTorch Image Models repository (Wightman, 2019). The experiments were run on one server with four NVIDIA A40 GPUs. We compare our method against GC (Selvaraju et al., 2017), IG (Sundararajan et al., 2017), T-Attn (Yuan et al., 2021), T-Attr (Chefer et al., 2021a), ViT-CX (Xie et al., 2023), TIS (Englebert et al., 2023), and Bi-Attn (Chen et al., 2023).

4.1 QUANTITATIVE EVALUATION

In this section, we provide quantitative evaluation of our MDA method's performance, and evaluate the parameters used in MDA. We will first evaluate how MDA performs in the perturbation tests that we have have optimized the framework for (Petsiuk et al., 2018; Walker et al., 2024). Next, we evaluate how these optimized attributions perform for three other metrics: pointing game (Chefer et al., 2021a), positive and negative perturbation (Chefer et al., 2021a; DeYoung et al., 2020), and monotonicity (Arya et al., 2019). Following these, we evaluate how MDA is affected by the selection of hyperparameters: τ, γ, and κ. Lastly, we evaluate runtime. Due to space constraints and their lower importance, we defer evaluations of monotonicity, τ, γ, and κ to the Appendix.

Optimized Perturbation Metric Evaluation: We have thoroughly described insertion, deletion, and the MAS variants in previous sections. For these tests, all input images are 224×224px and we use a step size of 224px, for a total of 224 perturbation steps as in the original implementations. Additionally, for insertion, we use a blurring kernel which results in a softmax score of less than 1%, to improve test accuracy. We use the code provided by the repositories for insertion/deletion and MAS (Petsiuk et al., 2018; Walker et al., 2024). The results in Table 1 compare all 8 attribution methods over 5000 ImageNet images with 5 images per class for the perturbation metrics. We see significant improvements by MDA in all tests, except for deletion, where it faces a loss. We see a minimum improvement of 9% in insertion - deletion, and a maximum improvement of 51% in MAS insertion - deletion. The deletion performance of MDA is easily explained by the definition of our patch order optimization. As we optimize the first few patches for a high insertion score, deletion can be penalized as a result (we explore this in A.6). However, as shown by insertion - deletion scores, joint optimization nets a positive improvement over all attribution methods. We show the results of this test on ViT-base 32×32 and ViT-tiny 16×16 models in Tables 4 and 5 in the Appendix.

Pointing Game Evaluation: In the pointing game, we measure the intersection-over-union (IoU), mean average precision (MAP), and F1 scores of each method's pointing accuracy when compared to ground-truth segmentation maps of ImageNet Segmentation data. This test uses the default code available in the original author's repository (Chefer et al., 2021b). For each attribution method, a threshold is applied, binarizing the attribution to make pointing game accuracy measurements. The metric authors use an attribution's mean as the threshold. However, since this threshold is equivalent to γ in MDA we select the best γ for MDA given the three output scores. The results are shown in Table 2 for the ViT-base 32×32 model. First, we see that choosing the ideal γ for MDA to perform

Table 1: Evaluation of MDA on the optimized metrics for the ViT-base 16×16 model.

	Metric From Petsiuk et al. (2018)			Metric From Walker et al. (2024)		
Test Type	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**
GC (Selvaraju et al., 2017)	0.737	0.241	0.496	0.622	0.312	0.311
IG (Sundararajan et al., 2017)	0.741	0.222	0.518	0.623	0.349	0.274
ViT-CX (Xie et al., 2023)	0.722	0.236	0.486	0.582	0.408	0.174
T-Attn (Yuan et al., 2021)	0.748	0.228	0.520	0.631	0.347	0.284
T-Attr (Chefer et al., 2021a)	0.741	0.232	0.508	0.638	0.331	0.307
Bi-Attn (Chen et al., 2023)	0.760	0.218	0.542	0.649	0.320	0.329
TIS (Englebert et al., 2023)	0.761	**0.196**	0.565	0.615	0.385	0.230
MDA (ours)	**0.856**	0.232	**0.624**	**0.775**	**0.279**	**0.497**

Table 2: ImageNet segmentation results for the ViT-base 32 × 32 model.

Metric	MAP	IoU	F1
GC (Selvaraju et al., 2017)	0.715	0.570	0.366
IG (Sundararajan et al., 2017)	0.687	0.573	0.462
ViT-CX (Xie et al., 2023)	0.678	0.535	0.428
T-Attn (Yuan et al., 2021)	0.705	0.594	0.465
T-Attr (Chefer et al., 2021a)	0.758	0.648	0.459
Bi-Attn (Chen et al., 2023)	0.753	0.653	0.491
TIS (Englebert et al., 2023)	0.705	0.586	0.460
MDA (ours)	**0.796**	**0.702**	0.487
MDA F1 Focused (ours)	0.760	0.661	**0.504**

well on all three tests leads to a win for MAP and IoU, but a minor loss for F1. However, if we choose γ such that F1 score is prioritized, we achieve an F1 win while retaining a win in MAP and IoU. These results show that optimizing MDA for insertion and deletion extends to high performance in other metrics, making MDA more valuable.

Positive and Negative Perturbation Evaluation: The positive and negative perturbation tests functional similarly to the optimized metrics, but measure prediction accuracy instead of softmax score. They measure an attribution's ability to explain the model with a two-stage process. First, for a given set of validation images, an attribution is made for all images. Following this, the pixels of the images are gradually masked, and the resulting accuracy is recorded at each step, forming a curve. In positive perturbation, pixels are masked in largest attribution order, and in negative perturbation they are masked in lowest attribution order, thus lower and higher AUC scores are better, respectively. We perform these tests with the default code available in the original author's repository (Chefer et al., 2021b) using the ImageNet validation set (Russakovsky et al., 2015) on the ViT-base 32 × 32 model and we show the results in Table 3. In this table we clearly see MDA outperforms the SOTA methods in both tests. These indicate not only that MDA best orders the most important features but it also best orders the least important features. Since this test measures via model accuracy and not softmax score, we can confirm MDA was not optimized for this test, yet still outperforms all methods.

Table 3: Positive and negative perturbation tests for the ViT-base 32 × 32 model.

Metric	Positive (↓)	Negative (↑)
GC (Selvaraju et al., 2017)	0.199	0.637
IG (Sundararajan et al., 2017)	0.128	0.698
ViT-CX (Xie et al., 2023)	0.143	0.708
T-Attn (Yuan et al., 2021)	0.140	0.714
T-Attr (Chefer et al., 2021a)	0.140	0.708
Bi-Attn (Chen et al., 2023)	0.124	0.733
TIS (Englebert et al., 2023)	0.126	0.741
MDA (ours)	**0.122**	**0.789**

Runtime Evaluation: We measure the mean runtime of MDA to be 13.34s, 3.64s, and 1.13s, for the ViT-base 16 × 16, ViT-tiny 16 × 16, and ViT-base 32 × 32 models, respectively. We provide a comparison of these times with SOTA attributions in the Appendix and provide further discussion about the value of online and offline attribution methods.

4.2 QUALITATIVE EVALUATION

In this section we perform qualitative MDA γ selection and compare MDA against the SOTA methods. In the Appendix, we extended these evaluations and explore seed attribution selection.

Evaluation of γ: We compare MDA with $\gamma = 0$ and $\gamma = 1$ on the ViT-base 32 × 32 model against Bi-Attn in Figure 5 and we make three distinctions. First, existing SOTA methods suffer from attribution on unimportant features, hurting quantitative and qualitative value. Second, at a low γ, MDA provides minimal information attributions, indicating only the most important features. Third,

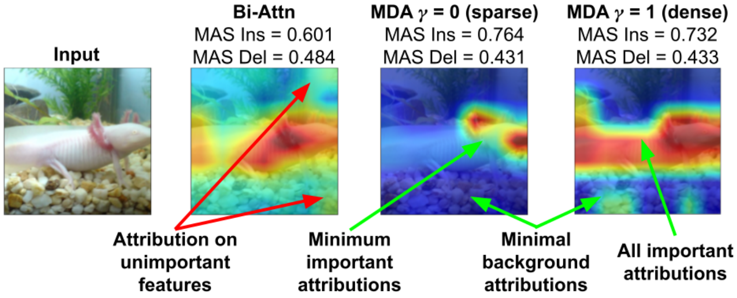


Figure 5: MDA has major visual improvements over the state-of-the-art methods. Unlike Bi-Attn (Chen et al., 2023), MDA does not provide attributions for any features unrelated to the class subject. In addition, it can provide a sparse or dense attribution while scoring more favorably.

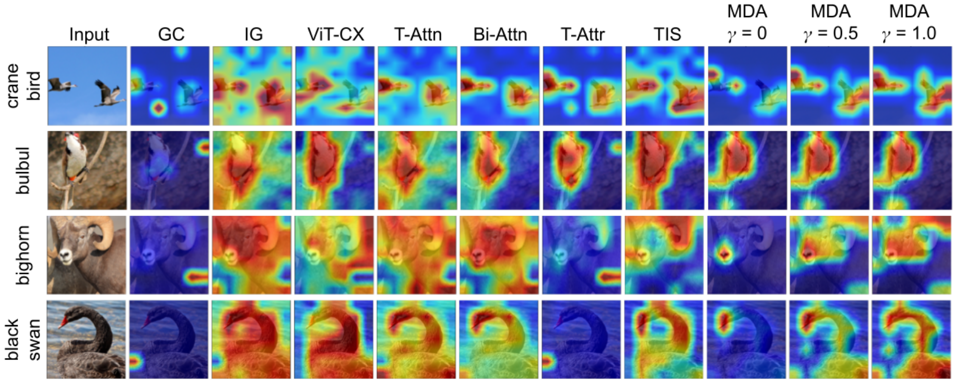


Figure 6: In the selected ImageNet examples we compare MDA with $\gamma = 0$, $\gamma = 0.5$, and $\gamma = 1$ against GC (Selvaraju et al., 2017), IG (Sundararajan et al., 2017), T-Attn (Yuan et al., 2021), T-Attr (Chefer et al., 2021a), ViT-CX (Xie et al., 2023), TIS (Englebert et al., 2023), and Bi-Attn (Chen et al., 2023). We see in all examples, MDA $\gamma = 0$ provides the best sparse attribution with no background attributions. As γ increases, MDA provides dense attributions which lack unimportant attributions.

MDA at a high γ provides dense attributions, highlighting all important features, while avoiding unimportant features. We see this behavior continues in additional examples in the Appendix.

SOTA Comparison: We present four qualitative comparison examples from the ImageNet dataset in Figure 6 for the ViT-base 32×32 model. We show MDA with $\gamma = 0$, $\gamma = 0.5$, and $\gamma = 1$ across a mix of images with a large and small subject. For all examples, $\gamma = 0$ provides a sparse, minimal information attribution which highlights the most important features. As γ moves to 1, attributions cover more of the image subject and stay retained to the image subject. In all examples, MDA provides reduced attributions on unimportant features when compared to the existing state-of-the-art ViT attribution methods. We show 255 more examples in Appendix A.10 equally divided among the three models. We see the behavior shown in this figure continues across the images presented, indicating that MDA provides more valuable attributions than state-of-the-art methods.

5 CONCLUSION

We propose the first-of-its-kind Metric-Driven Attribution for the vision transformer. Through two core optimization objectives, MDA generates attributions which perform highly in the insertion, deletion, and MAS attribution quality metrics. The first objective finds an ideal patch order which jointly maximizes insertion and deletion scores. The second objective finds the ideal magnitude of each patch to jointly maximize the MAS insertion and deletion scores. The result is a high-scoring, sparse explanation for only the most important features in an image. Through a modification of the objective, MDA can additionally create dense attributions, allowing a user to choose the sparse-dense balance. Across three ViT models, MDA shows significant performance gains over SOTA methods, further motivating optimization for attribution metrics. Due to its design, MDA can be applied to any input perturbation metric. The problem of defining the ideal metric is still open, and MDA can evolve with new metrics to create stronger attributions.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL `https://aclanthology.org/2020.acl-main.385`.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *NeurIPS*, 31, 2018.

Angelika Ando, Spyros Gidaris, Andrei Bursuc, Gilles Puy, Alexandre Boulch, and Renaud Marlet. Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5240–5250, 2023.

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the 2020 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2020.

Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*, pp. 63–71. Springer, 2016.

Ali Borji, Dicky N. Sihite, and Laurent Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1): 55–69, 2013. doi: 10.1109/TIP.2012.2210727.

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*, pp. 782–791, 2021a.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer attribution code repository, 2021b. Available at `https://github.com/hila-chefer/Transformer-Explainability`.

Jiamin Chen, Xuhong Li, Lei Yu, Dejing Dou, and Haoyi Xiong. Beyond intuition: Rethinking token attributions inside transformers. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=rm0zIzlhcX`.

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 Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL `https://aclanthology.org/2020.acl-main.408`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alexandre Englebert, Sédrick Stassin, Géraldin Nanfack, Sidi Ahmed Mahmoudi, Xavier Siebert, Olivier Cornu, and Christophe De Vleeschouwer. Explaining through transformer input sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 806–815, October 2023.

Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):5, 2021.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110:328–348, 2014.

George To Sum Ho, Yung Po Tsang, Chun Ho Wu, Wai Hung Wong, and King Lun Choy. A computer vision-based roadside occupation surveillance system for intelligent transport in smart cities. *Sensors*, 19(8):1796, 2019.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viegas, and Michael Terry. Xrai: Better attributions through regions. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4947–4956, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. doi: 10.1109/ICCV.2019.00505. URL `https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00505`.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2019. doi: 10.48550/arXiv.1912.01703. URL `https://doi.org/10.48550/arXiv.1912.01703`.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

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 *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3319–3328. JMLR.org, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chase Walker, Dominic Simon, Kenny Chen, and Rickard Ewetz. Attribution quality metrics with magnitude alignment. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 530–538. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/59. URL `https://doi.org/10.24963/ijcai.2024/59`. Main Track.

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

Weiyan Xie, Xiao-Hui Li, Caleb Chen Cao, and Nevin L. Zhang. Vit-cx: Causal explanation of vision transformers. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 1569–1577. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/174. URL `https://doi.org/10.24963/ijcai.2023/174`. Main Track.

Tingyi Yuan, Xuhong Li, Haoyi Xiong, Hui Cao, and Dejing Dou. Explaining information flow inside vision transformers using markov chain. In *eXplainable AI approaches for debugging and diagnosis.*, 2021. URL `https://openreview.net/forum?id=TT-cf6QSDaQ`.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

A APPENDIX

In this section, we present extended quantitative and qualitative results. First, we present the additional tables referenced in the experimental evaluation to present results across all three models. Then, we present detailed and extensive qualitative results across all three models.

A.1 EXTENDED OPTIMIZATION METRIC EVALUATION

We include Tables 4 and 5 which provide the results of the main experimental comparison for the ViT-base 32 × 32 and ViT-tiny 16 × 16 models. The results seen in these tables are consistent with the Table 1. The only difference is in Table 5 where the winner of deletion is ViT-CX. However, referencing Figure 7, we can still conclude this is due to our joint optimization of insertion and deletion, and MDA would clearly win over ViT-CX in deletion if optimized solely for deletion. In Table 4, we note a maximum improvement of 73% in MAS insertion - deletion.

Table 4: Evaluation of MDA on the optimized metrics for the ViT-base 32 × 32 model.

	Metric From Petsiuk et al. (2018)			Metric From Walker et al. (2024)		
Test Type	**Ins** (↑)	**Del** (↓)	**Ins - Del** (↑)	**Ins** (↑)	**Del** (↓)	**Ins - Del** (↑)
GC (Selvaraju et al., 2017)	0.783	0.262	0.520	0.695	0.339	0.356
IG (Sundararajan et al., 2017)	0.805	0.206	0.598	0.575	0.425	0.151
ViT-CX (Xie et al., 2023)	0.813	0.208	0.605	0.615	0.389	0.226
T-Attn (Yuan et al., 2021)	0.797	0.222	0.576	0.578	0.427	0.151
T-Attr (Chefer et al., 2021a)	0.798	0.221	0.576	0.650	0.358	0.292
Bi-Attn (Chen et al., 2023)	0.826	0.205	0.621	0.613	0.396	0.216
TIS (Englebert et al., 2023)	0.847	**0.176**	0.671	0.643	0.370	0.273
MDA (ours)	**0.908**	0.200	**0.708**	**0.851**	**0.235**	**0.616**

Table 5: Evaluation of MDA on the optimized metrics for the ViT-tiny 16 × 16 model.

	Metric From Petsiuk et al. (2018)			Metric From Walker et al. (2024)		
Test Type	**Ins** (↑)	**Del** (↓)	**Ins - Del** (↑)	**Ins** (↑)	**Del** (↓)	**Ins - Del** (↑)
GC (Selvaraju et al., 2017)	0.729	0.252	0.477	0.623	0.323	0.299
IG (Sundararajan et al., 2017)	0.745	0.233	0.512	0.627	0.355	0.272
ViT-CX (Xie et al., 2023)	0.719	**0.155**	0.563	0.548	0.387	0.161
T-Attn (Yuan et al., 2021)	0.752	0.239	0.513	0.635	0.354	0.281
T-Attr (Chefer et al., 2021a)	0.745	0.243	0.502	0.643	0.339	0.304
Bi-Attn (Chen et al., 2023)	0.764	0.228	0.536	0.654	0.328	0.326
TIS (Englebert et al., 2023)	0.764	0.207	0.557	0.617	0.389	0.228
MDA (ours)	**0.858**	0.241	**0.617**	**0.777**	**0.289**	**0.488**

A.2 MONOTONICITY EVALUATION

The last quantitative attribution metric we include is the monotonicity test (Arya et al., 2019). This test measures if inserting all features in order of largest attribution value will lead to a monotonically increasing curve model confidence. Starting from a blurred image, features are inserted and the confidence of the image class prediction is measured. Over all features, this creates the confidence vector $\mathbf{c}$. This vector is then compared to a feature order vector $\mathbf{o}$ which denotes the descending token importance. Monotonicity is then measured as $MONO = p(\mathbf{o}, \mathbf{c})$, where p(.) is the Spearman correlation of the vectors, and a higher score means $\mathbf{c}$ is more monotonic. We report the results in Table 6 for the ViT-base 32 × 32 model using 100 ImageNet validation dataset images. We see MDA significantly improves over the SOTA methods. This further proves that *MDA meets objectives outside of its optimization targets*, showing its value as an effective attribution for ViT models.

Table 6: Monotonicity comparison on the ViT-base 32 × 32 model.

Metric	**Monotonicity (↑)**
GC (Selvaraju et al., 2017)	0.798
IG (Sundararajan et al., 2017)	0.791
ViT-CX (Xie et al., 2023)	0.775
T-Attn (Yuan et al., 2021)	0.782
T-Attr (Chefer et al., 2021a)	0.788
Bi-Attn (Chen et al., 2023)	0.783
TIS (Englebert et al., 2023)	0.770
MDA (ours)	**0.820**

A.3 QUANTITATIVE τ EVALUATION

We now explain the deletion loss in Tables 1, 4, and 5, with an ablation study of τ in Figure 7. In graph (a), the dots represent the insertion and deletion scores of the attributions and in (b), the dots are the MAS scores. Better overall scores are closer to the origin. The 11 dots are generated by varying τ from 1 to 0 (decreasing by 0.1 left to right). As τ decreases, the deletion score improves and insertion suffers. When $\tau = 1$, MDA is optimized solely for insertion and when $\tau = 0$, it is optimized for deletion only. We circle $\tau = 0.9$ in red and draw a boundary across the MDA dots. We observe no SOTA methods pass the boundary, signifying *there exists a configuration of MDA which can win both tests*. In (c), we show selecting $\tau = 0.9$, on average, provides a peak in the joint scores, indicating it is the correct choice. This analysis was performed on 100 ImageNet images for the ViT-base 32 × 32 model.

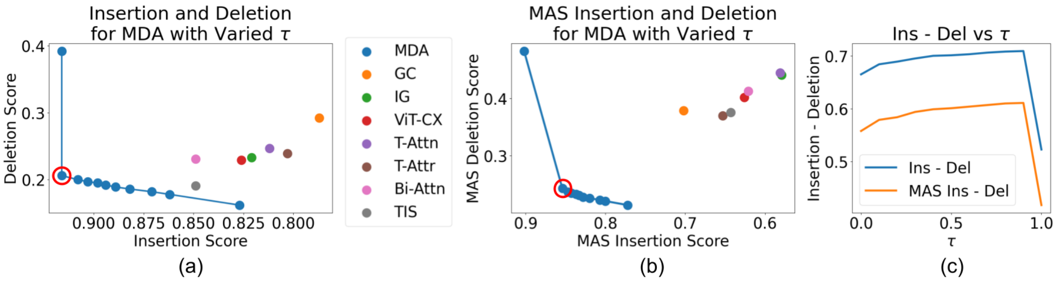


Figure 7: We illustrate the impact of τ for the insertion, deletion (a) MAS tests (b), and insertion - deletion (c). In (a) and (b), a score near the origin is better and we circle $\tau = 0.9$ in red. We draw a boundary across all MDA points and compare the SOTA attributions. We see no method breaks this boundary, indicating there is always a configuration of MDA which will perform best. We show $\tau = 0.9$ provides the peak score (c), verifying it is the best choice for joint optimization.

We present in Tables 7, 8, and 9, tabular data of this τ experiment for the ViT-base 16 × 16, ViT-tiny 16 × 16, and and ViT-base 32 × 32 models when only MDA is considered. As expected, we see for all models that MDA optimized for insertion, deletion, or with joint optimization will perform the best on the insertion, deletion, or insertion - deletion tests, respectively.

Table 7: Evaluation of MDA optimization of insertion or deletion on the ViT-base 16 × 16 model.

	Metric From Petsiuk et al. (2018)			Metric From Walker et al. (2024)		
Test Type	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**
MDA Insertion	**0.888**	0.501	0.387	**0.875**	0.649	0.226
MDA Deletion	0.710	**0.151**	0.559	0.581	**0.216**	0.364
MDA (ours)	0.864	0.259	**0.605**	0.781	0.309	**0.472**

Table 8: Evaluation of MDA optimization of insertion or deletion on the ViT-tiny 16×16 model.

	Metric From Petsiuk et al. (2018)			Metric From Walker et al. (2024)		
Test Type	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**
MDA Insertion	**0.896**	0.236	0.660	**0.880**	0.313	0.567
MDA Deletion	0.706	**0.062**	0.644	0.640	**0.143**	0.496
MDA (ours)	0.882	0.140	**0.742**	0.829	0.163	**0.666**

Table 9: Evaluation of MDA optimization of insertion or deletion on the ViT-base 32×32 model.

	Metric From Petsiuk et al. (2018)			Metric From Walker et al. (2024)		
Test Type	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**
MDA Insertion	**0.916**	0.392	0.523	**0.902**	0.483	0.418
MDA Deletion	0.827	**0.162**	0.665	0.772	**0.214**	0.558
MDA (ours)	0.916	0.206	**0.709**	0.854	0.243	**0.611**

Lastly, we provide a visual example to further explain why we must use $\tau = 0.9$ to perform a joint insertion and deletion process. To do this, we illustrate that the resulting patch orders of the insertion and deletion optimizations do not agree. We show the effects of optimizing solely for insertion or deletion, or performing a joint optimization in Figure 8. In this image, the darkest red represents the most important patch and darkest blue represents the least important. We can see insertion-only optimization ($\tau = 1$) considered the patches on the animal's head to be most important while deletion-only optimization ($\tau = 0$) considers those on the tail most important. However, when we perform joint optimization there is an agreement that the whole body is important, but the head is the most important feature. From this study we confirm that both *a joint optimization and a proper selection of* τ *is necessary* for the best patch ordering to be obtained.

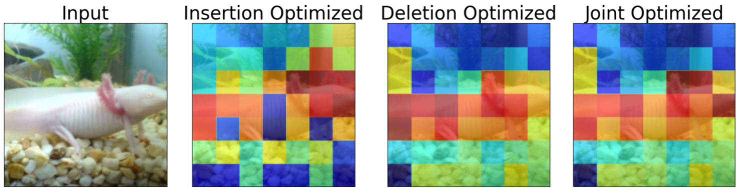


Figure 8: A visual comparison of the varied patch orders resulting from optimization of only insertion, only deletion, or both insertion and deletion. The darkest red represents the first patch in the order, and the darkest blue represents the last.

A.4 RUNTIME EVALUATION

We perform a runtime evaluation of all attribution methods presented in the main paper. These results are gathered over 100 ImageNet (Russakovsky et al., 2015) images for the ViT-base 16×16, ViT-tiny 16×16 and 32×32 models. We report the mean runtime and standard deviation across all images for each attribution method in seconds. These results are shown in Table 10. We see in the results that MDA is clearly slower, but its runtime in the range of 1.1 to 13.3 seconds is still *very reasonable for real-world use*. The reviewers discuss that attribution methods should be real-time. However, we argue that the use of attributions should not be limited to real-time scenarios only. Specifically in the case of MDA, we believe its slower runtime is a fair trade off not only for its performance, but also its flexibility which benefits offline utilization, and its lack of need for internal model information. MDA not only achieves improvements of up to 73% over SOTA methods, but it is the sole attribution method which is flexible to user preferences in providing sparse or dense attributions. This second attribute is valuable for offline evaluation of a neural model, especially as γ is applied in post without model calls such that only one MDA generation must occur to have access to a varied set of attributions from $\gamma \in [0, 1]$. The flexibility could assist in debugging in ways that current attribution methods cannot.

Table 10: Runtime comparison on the ViT-b 16 × 16, ViT-t 16 × 16 , and ViT-b 32 × 32 models.

	Mean Runtime per Image (s)		
Model	**ViT-b 16×16**	**ViT-t 16×16**	**ViT-b 32×32**
GC (Selvaraju et al., 2017)	$\mathbf{0.014} \pm 0.006$	$\mathbf{0.008} \pm 0.005$	$\mathbf{0.010} \pm 0.006$
IG (Sundararajan et al., 2017)	0.226 ± 0.007	0.116 ± 0.012	0.144 ± 0.007
ViT-CX (Xie et al., 2023)	0.519 ± 0.078	0.189 ± 0.020	0.609 ± 0.058
T-Attn (Yuan et al., 2021)	0.225 ± 0.007	0.122 ± 0.012	0.143 ± 0.008
T-Attr (Chefer et al., 2021a)	0.049 ± 0.006	0.048 ± 0.008	0.049 ± 0.008
Bi-Attn (Chen et al., 2023)	0.226 ± 0.007	0.118 ± 0.011	0.141 ± 0.008
TIS (Englebert et al., 2023)	1.104 ± 0.015	0.171 ± 0.011	0.359 ± 0.015
MDA (ours)	13.344 ± 0.308	3.639 ± 0.649	1.130 ± 0.249

A.5 QUANTITATIVE γ EVALUATION

We compare MDA with $\gamma = 0.0$, $\gamma = 0.5$, and $\gamma = 1.0$ on 100 ImageNet images. Table 11 shows the ViT-base 16 × 16 model results. As expected, MDA $\gamma = 0$ achieves the best scores in all tests. However, the score penalty of increasing γ is not extreme, indicating dense attributions are still of value. We provide results for the ViT-tiny 16 × 16 and ViT-base 32 × 32 models in Tables 12 and 13 below. We find the same results to hold true. We later provide more detailed qualitative examples to compare MDA against SOTA methods and further analyze the effect of γ across the models.

Table 11: Evaluation of MDA with varying levels of γ on the ViT-base 16 × 16 model.

	Metric From Petsiuk et al. (2018)			Metric From Walker et al. (2024)		
Test Type	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**
MDA (dense) $\gamma = 1$	0.812	0.350	0.462	0.701	0.466	0.235
MDA $\gamma = 0.5$	0.849	0.262	0.587	0.781	0.312	0.472
MDA (sparse) $\gamma = 0$	**0.864**	**0.259**	**0.605**	**0.781**	**0.309**	**0.472**

Table 12: Evaluation of MDA with varying levels of γ on the ViT-tiny 16 × 16 model.

	Metric From Petsiuk et al. (2018)			Metric From Walker et al. (2024)		
Test Type	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**
MDA $\gamma = 1$	0.861	0.159	0.703	0.792	0.194	0.599
MDA $\gamma = 0.5$	0.882	0.140	0.742	0.827	0.163	0.665
MDA $\gamma = 0$ (ours)	**0.882**	**0.140**	**0.742**	**0.829**	**0.163**	**0.666**

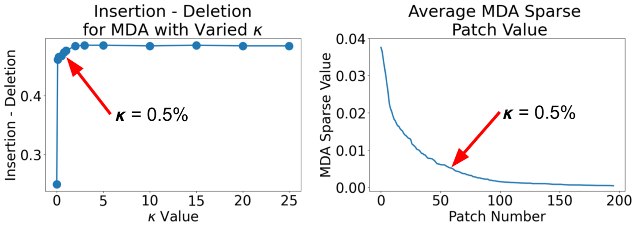
Table 13: Evaluation of MDA with varying levels of γ on the ViT-base 32 × 32 model.

	Metric From Petsiuk et al. (2018)			Metric From Walker et al. (2024)		
Test Type	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**	**Ins (↑)**	**Del (↓)**	**Ins - Del (↑)**
MDA $\gamma = 1$	0.914	0.214	0.699	0.834	0.255	0.578
MDA $\gamma = 0.5$	0.916	0.206	0.709	0.841	0.243	0.599
MDA $\gamma = 0$ (ours)	**0.916**	**0.206**	**0.709**	**0.854**	**0.243**	**0.611**

A.6 QUANTITATIVE κ EVALUATION

In this section we perform an ablation study on κ to motivate our choice of $\kappa = 0.005 = 0.5\%$. In this test we set $\gamma = 1$ to activate the application of Eq (11), and we vary κ. From Eq (11), κ determines the threshold of model importance that a patch must have for it to be assigned a significant attribution value, thus by increasing κ the density of the attribution will decrease. In the paper, we choose the very small value of 0.5% importance as we intend for all patches with any model importance to be attributed. When $\kappa = 0\%$ every patch receives a value relative to its order. When κ is greater than or equal to the importance of the most important patch, $A^{dense} == A^{sparse}$. In this test, we only report MAS scores (Walker et al., 2024), as κ does not change the order of patches, thus all insertion and deletion scores (Petsiuk et al., 2018) will be the same regardless of κ selection.

We use $\kappa = [0, 0.1, 0.25, 0.5, 0.75, 1, 2, 3, 5, 10, 15, 20, 25]$ where each value is a percentage and we evaluate over 100 ImageNet images. In Figure 9(a) we plot MAS insertion - deletion scores against the choice of κ. We see when $\kappa = 5\%$ the attribution has approached A^{sparse} as the score stagnates, and when it is 0% the attribution scores poorly. We choose κ such that the attribution scores well and attributes more patches than A^{sparse}. In (b), we plot the model importance of the highest to lowest value patches across all 100 images. This provides insight into what portion of patches are being selected with different values of κ. Since the largest value here is 0.04 or 4%, we confirm $\kappa > 5\%$ yields A^{sparse} on average over these images. Furthermore, we see $\kappa = 0.5\%$ results in Eq (11) applying to nearly $60/196$ patches. Overall, we observe $\kappa = 0.5\%$ provides a balance between only attributing patches which are valuable to the model (the score is still high) and attributing a significant number of patches (the attribution is dense). Finally, we can visually confirm this behavior with Figure 10 where we see $\kappa = 0\%$ results in poor attributions with too many background attributions, increasing κ leads to A^{sparse}, and $\kappa = 0.5\%$ provides the best balance for A^{dense}.



(a) (b)

Figure 9: Quantitative comparison of MDA with varied κ. We see $\kappa = 5\%$ provides the best balance of (a) insertion - deletion score and (b) number of patches strongly attributed.

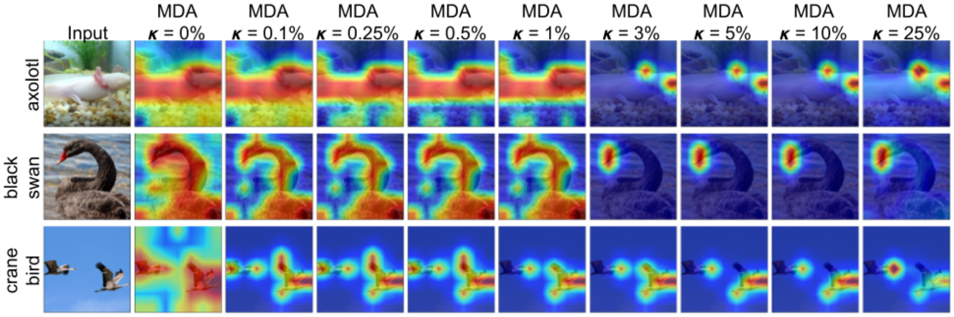
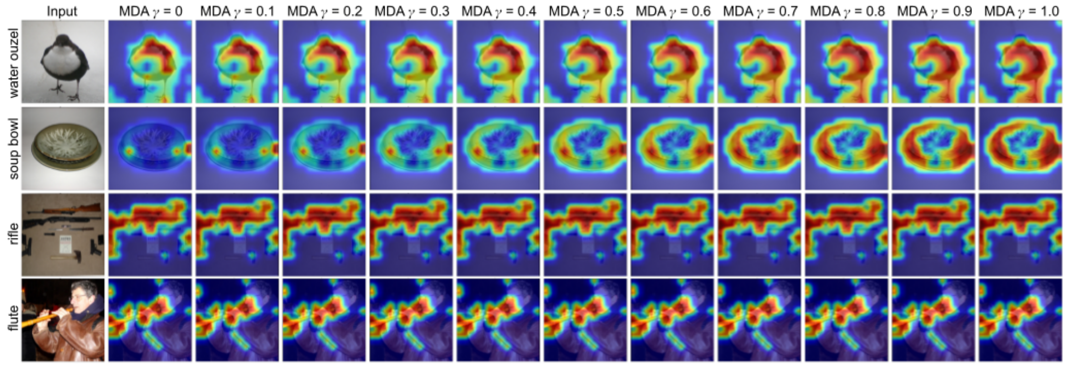


Figure 10: Qualitative comparison of MDA with varied κ. We see $\kappa = 5\%$ provides is the best choice as it consistently results in attributing all important features without attributing the background.

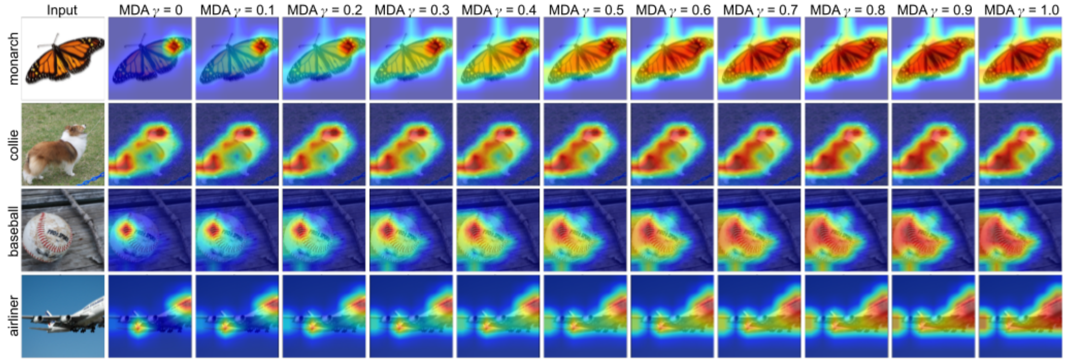
A.7 EXTENDED QUALITATIVE γ EVALUATION

First, we provide detailed qualitative evaluation of the visual effects as MDA attributions move from sparse to dense with varying γ across all three models. We perform this analysis in Figure 11 for the ViT-base 16×16 model in (a), the ViT-base 32×32 model in (b), and the ViT-tiny 16×16 model in (c). We show the results for 4 images per model as γ transitions from 0 to 1 by steps of 0.1. For each model, we present a mixture of images with small and large subjects. In Figure 11 (a), we see the two images with small subjects of class “rifle" and “flute" have consistent attributions for all γ. This behavior indicates, for these images, the model requires all of the subject’s features for classification and no less. For the other three images, we see the attributions across the subjects grow in intensity and density with increasing γ, indicating only a subset of the subject's features provide a decision. In (b) we see identical behavior to (a), however, we see a larger change between low and high γ. This could be explained by the larger patches used by the 32×32 model which contain more information. Lastly, in (c) we observe the change in attribution density with increase in γ is much less pronounced for the ViT-tiny model. This could be attributed to the fewer layers in the model which results in learning less specific features, thus more features of any subject are needed for a decision to be made.

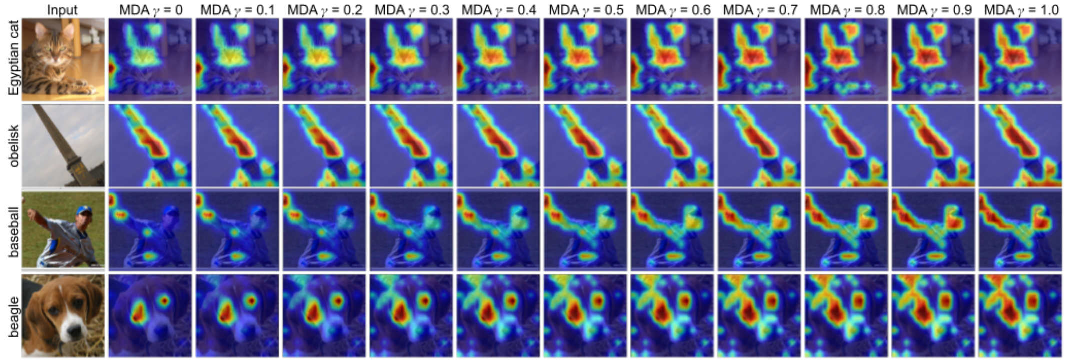
This study shows the effectiveness of MDA for creating both sparse and dense attributions. Additionally, we see studying MDA attributions with varying γ can show behaviors otherwise unseen through analysis of current attributions which lack sparsity control. MDA can provide an understanding of the minimum required information a model needs to make a decision and how this varies with model parameters, which can be valuable for choosing the best model for a given application.



(a) ViT-Base 16 x 16



(b) ViT-Base 32 x 32



(c) ViT-Tiny 16 x 16

Figure 11: We analyze how changes in γ effect the sparsity to density transition of an MDA attribution. We see for the first two ViT-base models (a) and (b) that the sparsity to density transition is smooth and significant, meaning the models find many features important, but only need a few. However, the minimal transition for ViT-tiny (c), indicates the model requires more features for a decision.

A.8 QUALITATIVE EVALUATION OF MDA SEED ATTRIBUTION

MDA as presented and evaluated in this work uses a high-quality input attribution to reduce its search space. We present a small selection of examples to indicate the impact this “seed" attribution has on the final MDA result. In Figure 12, we present three examples which include the following: an input image, GC (Selvaraju et al., 2017), IG (Sundararajan et al., 2017), T-Attn (Yuan et al., 2021), T-Attr (Chefer et al., 2021a), ViT-CX (Xie et al., 2023), TIS (Englebert et al., 2023), and Bi-Attn (Chen et al., 2023) attributions, the output of MDA with each of these attributions as an input, labeled as “MDA seed", and MDA without a seed. All examples are for the ViT-base 32 $\times$ 32 model and the images are from ImageNet. In the first example (a) we see a wide variety in the input attributions. GradCAM provides no attribution on the bird subject, IG provides attributions on a large portion of the image, and the remaining methods provide attributions on the bird with varying degrees of background attribution. However, we see each MDA attribution seeded by these inputs remains both fairly consistent, and loyal to MDA without a seed. This is desirable, as the input attribution is shown to not destroy the performance of MDA . In (b) we see the same behavior as before. The input attributions vary widely in appearance and quality, yet the MDA outputs maintain a level of consistency, although reduced from the consistency of (a). We see this pattern continue in example (c). Overall, MDA with a seed provides the benefit of greatly reduced runtime without a marginal loss in quality. MDA is best used with a high-quality attribution, and future higher-quality attributions could provide a better seed for better results.

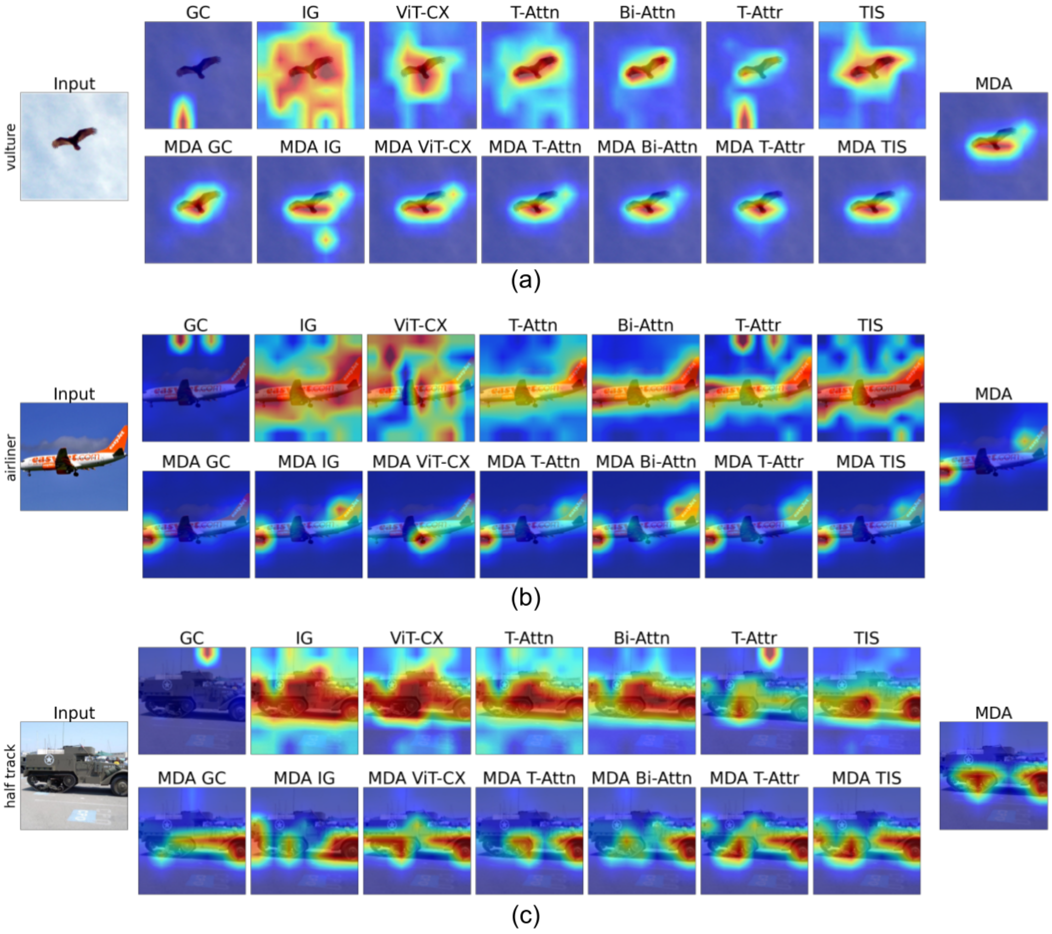


Figure 12: We qualitatively compare the performance of MDA with various “seed" attributions used as input against MDA without a seed. Regardless of seed quality, MDA provides an output consistent with the unseeded version, indicating the input attribution does not have a large impact on final result.

A.9 EVALUATION OF MDA VS SHAP

The Shapely value sampling attribution SHAP (Lundberg & Lee, 2017) is a popular perturbation-based method that can be considered in our patch order optimization problem. To make a comparison of our MDA method to SHAP, we would first define our approach as existing between the occlusion perturbation (Zeiler & Fergus, 2014) and SHAP (Lundberg & Lee, 2017) approaches. The occlusion method generates attributions by measuring the impact of masking only one input patch at a time. However, this results in issues where relationships between input features cannot be determined, but a $\mathcal{O}(M^2)$ runtime is achieved, where M^2 is the total number of patches. On the other hand, SHAP, if discussed in our context, would generate every possible ordering of the patches in the input image *without regard to the metrics* and then an evaluation of each of the resulting M^2 attributions would be made and the best scoring one would be chosen. This leads to SHAP having a runtime of $\mathcal{O}(2^{M^2})$. As a benefit of our problem definition and methodical approach, we are able to greedily select the best patch at every search step, preventing an iteration over all orderings, and yielding a $\mathcal{O}(M^2)$ runtime. *SHAP has a runtime of* $2.4s$ *compared to the* $1.1s$ *of MDA* on the ViT-base 32 × 32 model.

In addition to this, SHAP has significantly worse performance both quantitatively and qualitatively. When performing the segmentation test from Table 2, we see SHAP performs significantly worse than MDA in Table 14 below, losing in every test by a large margin. In addition, in Figure 13 below, we see that MDA consistently provides attributions which are both focused on the interesting features of the image subject and do not focus on the unimportant background features, while SHAP struggles with localization and has significant background attributions.

Table 14: SHAP vs MDA on segmentation test for the ViT-base 32 × 32 model

Metric	**MAP**	**IoU**	**F1**
SHAP (Lundberg & Lee, 2017)	0.648	0.512	0.405
MDA (ours)	**0.796**	**0.702**	**0.487**

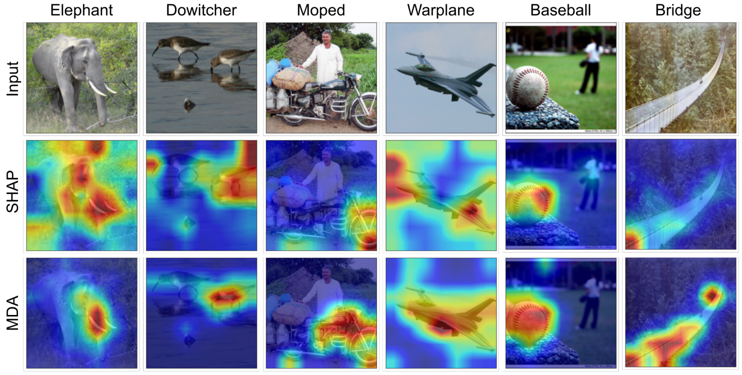
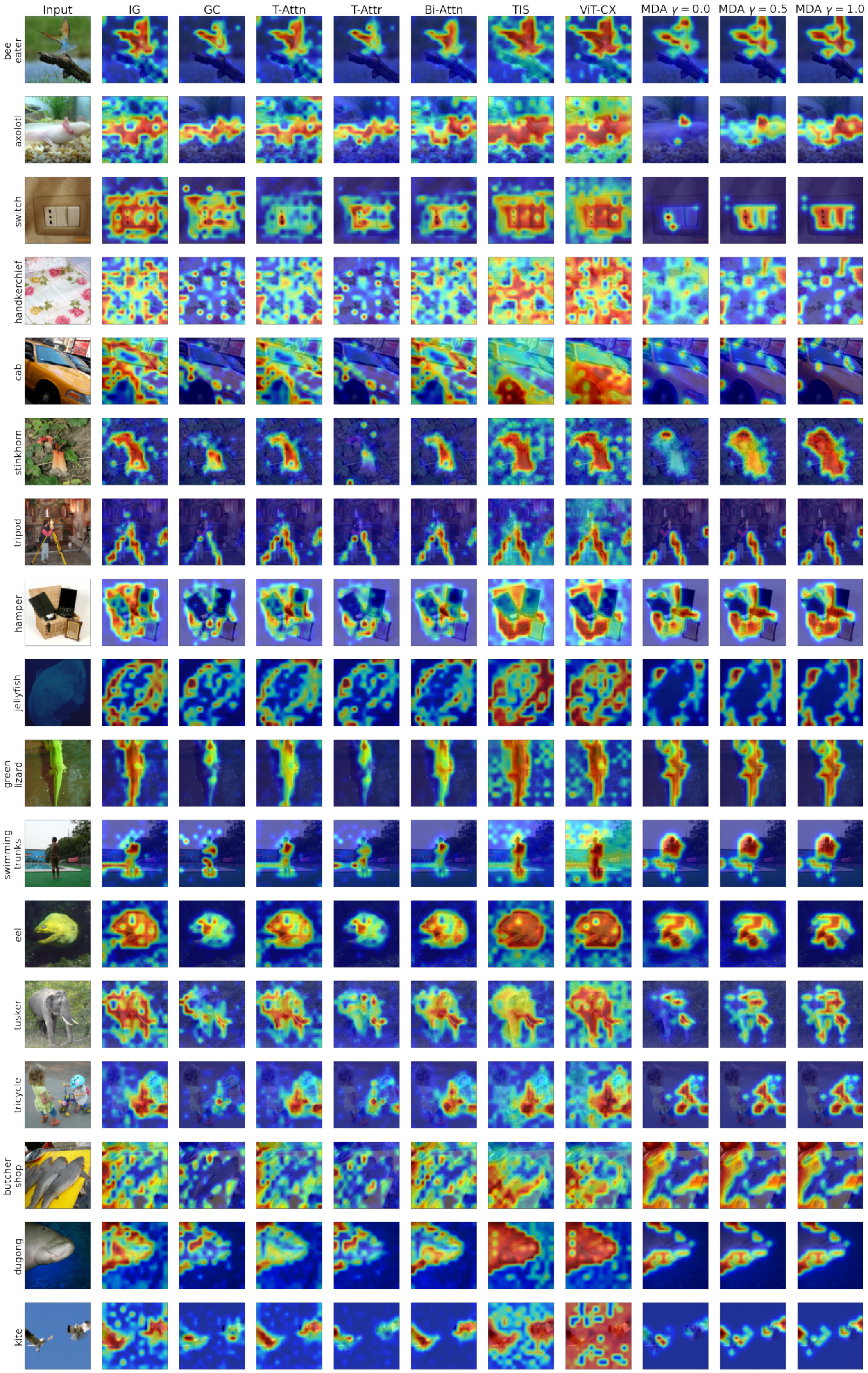
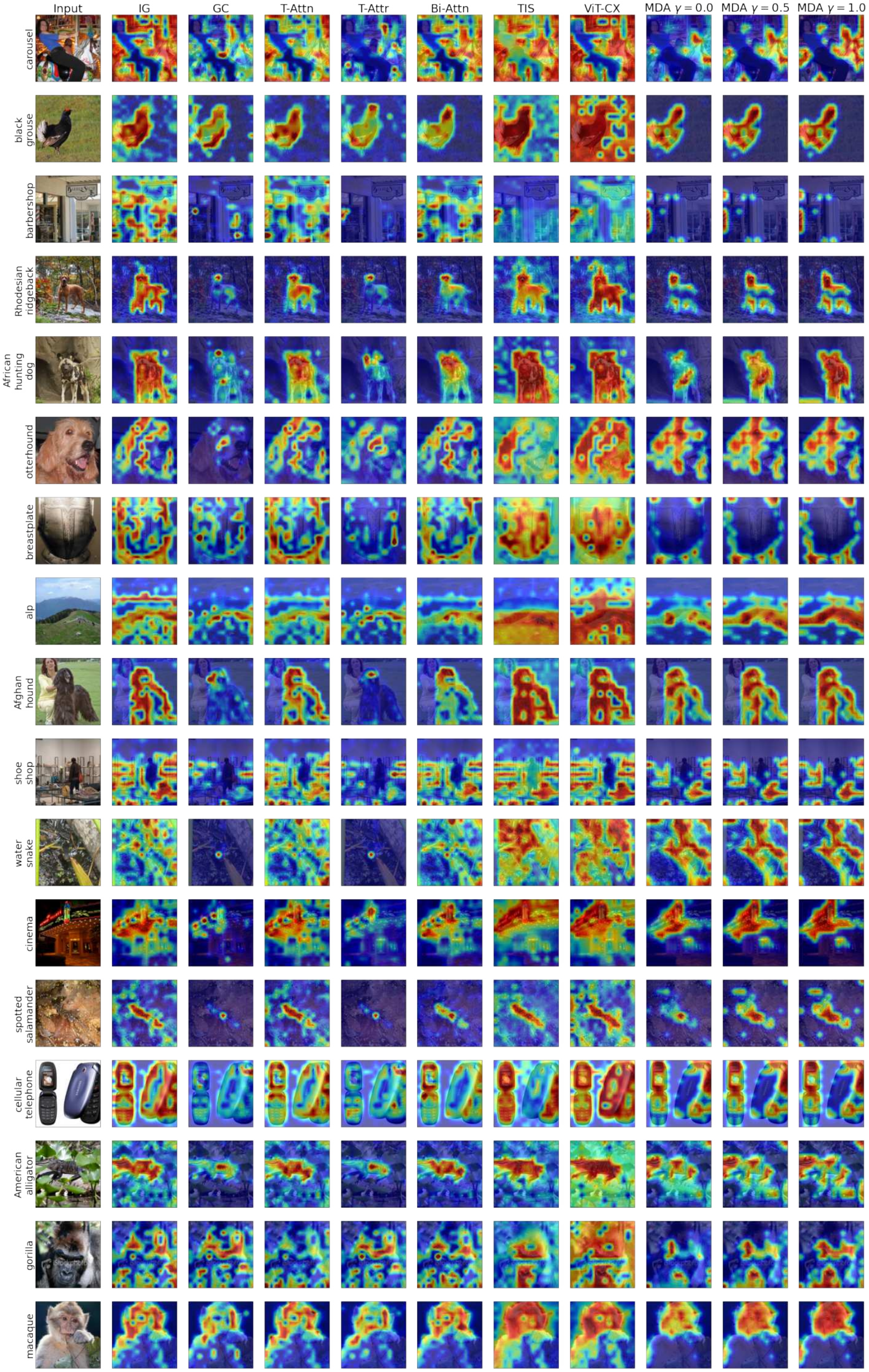


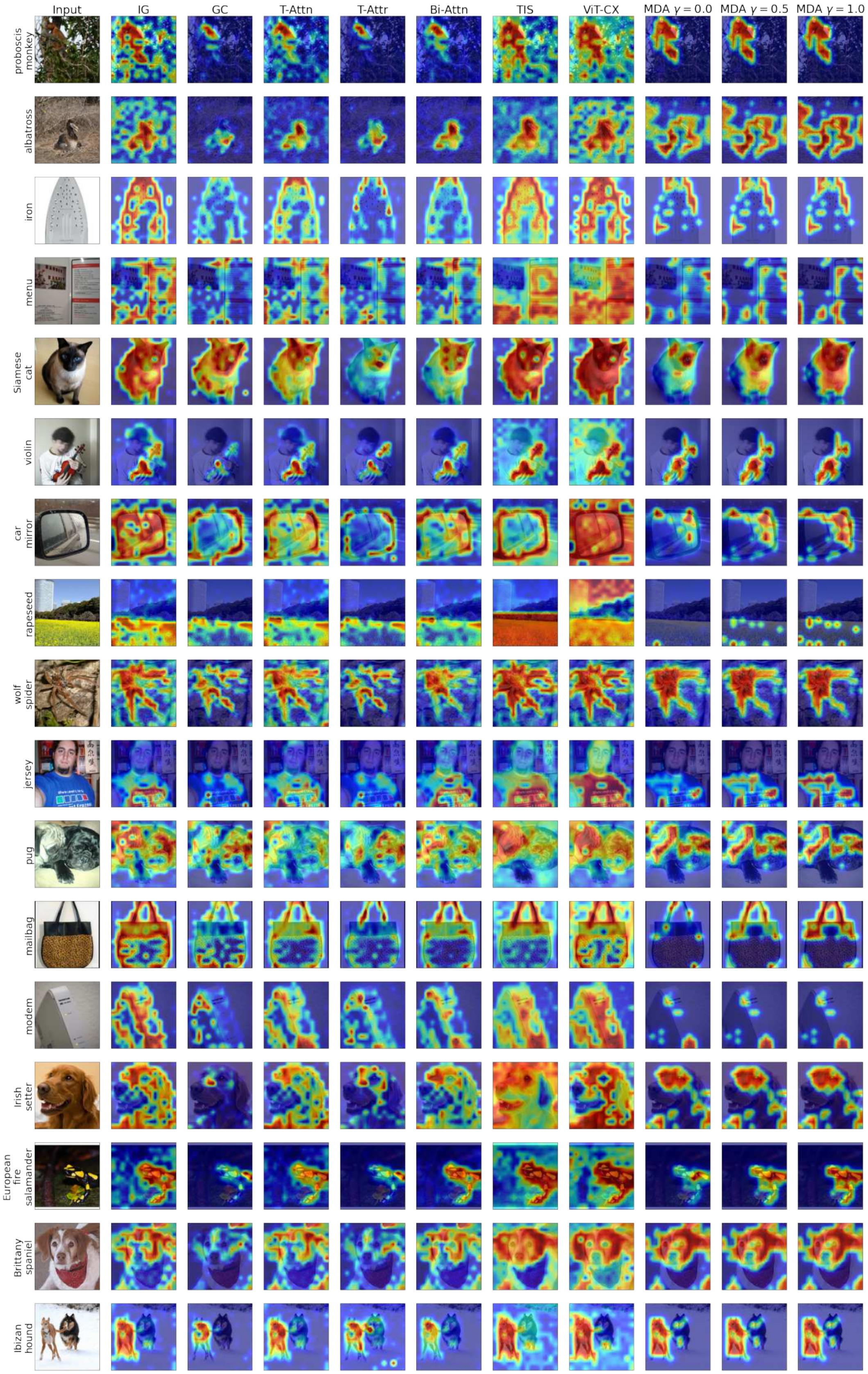
Figure 13: A visual comparison of MDA with SHAP. We see clear improvements by MDA over SHAP in all examples. MDA shows reduced background attributions and better subject localization.

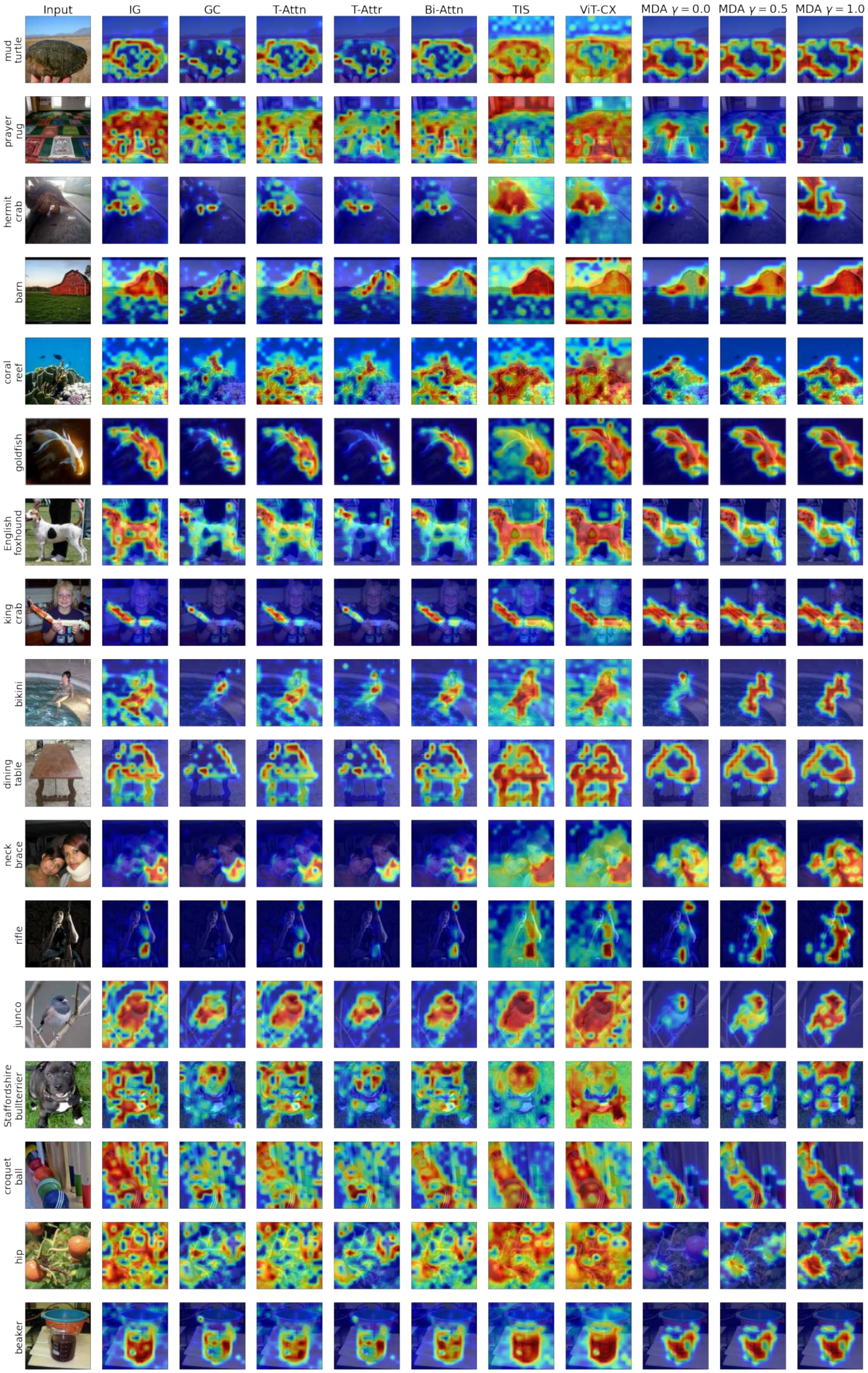
A.10 EXTENDED QUALITATIVE VISUAL PERFORMANCE EVALUATION

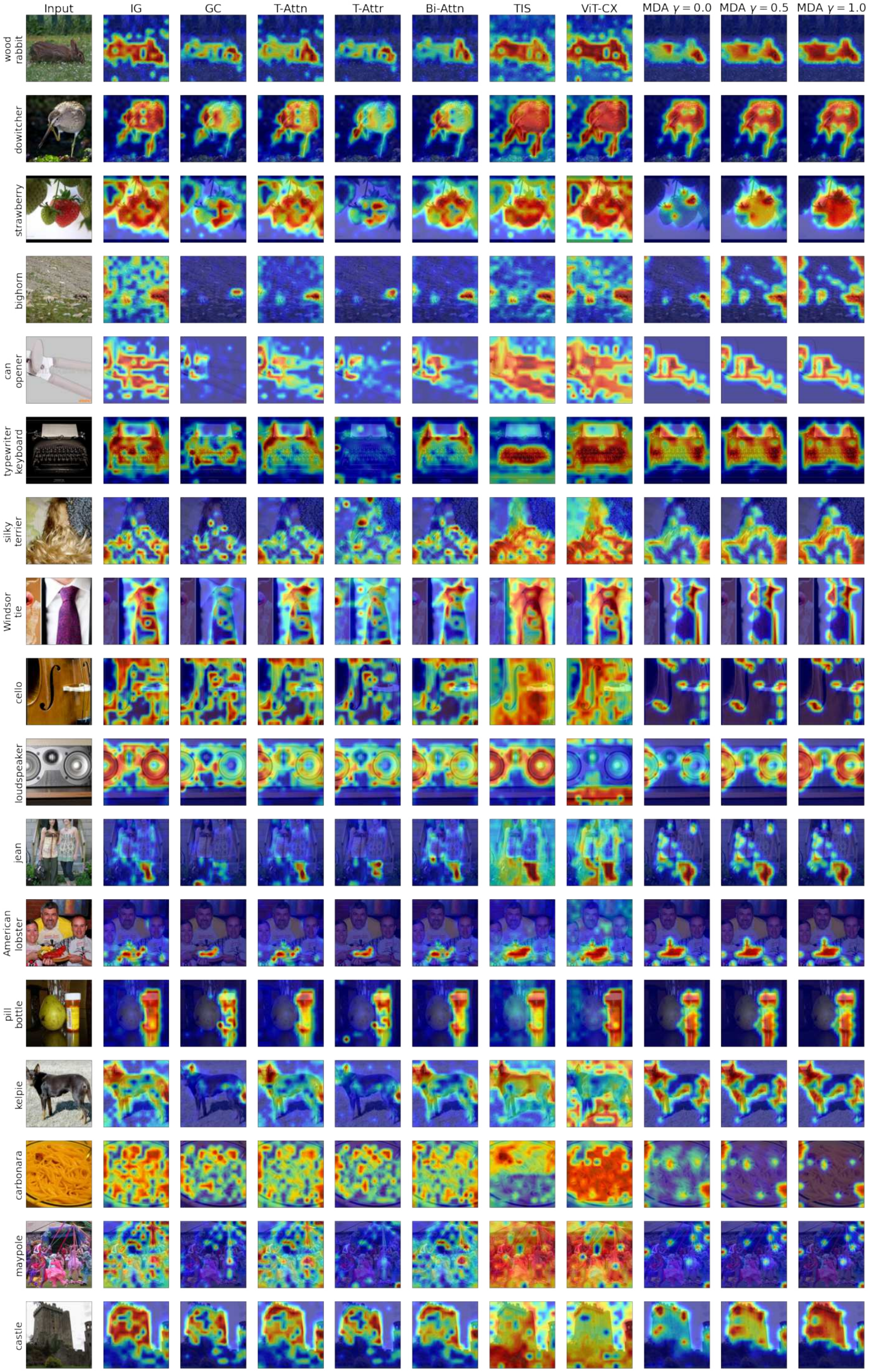
We now present a large selection of attribution comparisons generated from the ImageNet validation dataset which extend the comparisons in Figure 6. We contrast MDA with $\gamma = 0$, $\gamma = 0.5$, and $\gamma = 1$ against IG (Sundararajan et al., 2017), GC (Selvaraju et al., 2017), T-Attn (Yuan et al., 2021), T-Attr (Chefer et al., 2021a), Bi-Attn (Chen et al., 2023), TIS (Englebert et al., 2023), and ViT-CX (Xie et al., 2023). The examples are broken in to three sections, in groups of five pages. From page 22 to 26, we display attributions for the ViT-base 16 × 16 model. Pages 27 - 31 contain attributions for the ViT-base 32 × 32 model. Lastly, pages 32 - 36 contain attributions for the ViT-tiny 16 × 16 model. Across 15 pages we present 255 unique images. MDA presents the most consistent, high-quality attributions across the images and models. The transition from $\gamma = 0$ to $\gamma = 1$ consistently results in a move towards a dense attribution when the model does not require all features for classification.

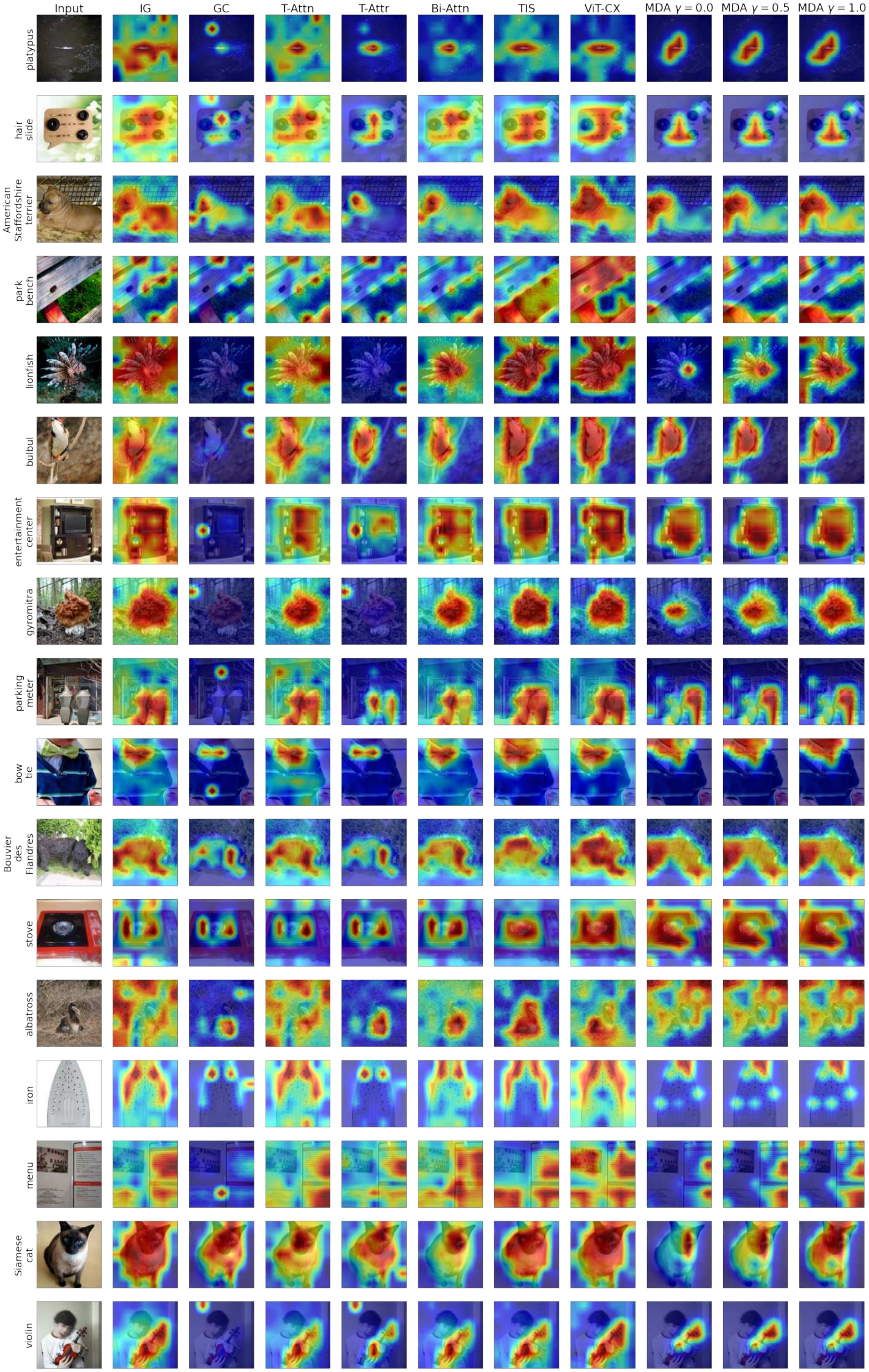

Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
bee eater
axolotl
switch
handkerchief
cab
stinkhorn
tripod
hamper
jellyfish
green lizard
swimming trunks
eel
tusker
tricycle
butcher shop
dugong
kite

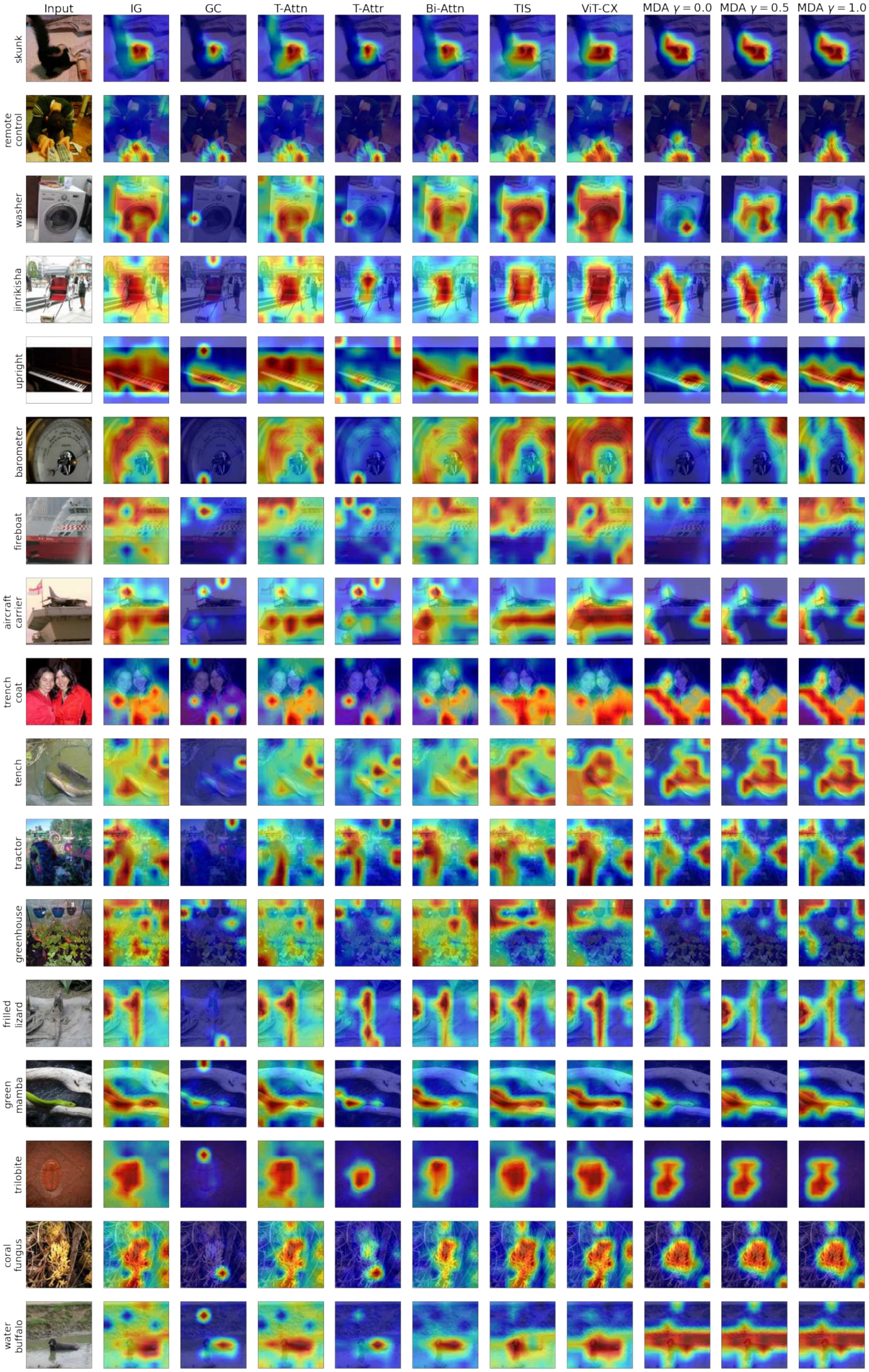

Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
carousel
black grouse
barbershop
Rhodesian ridgeback
African hunting dog
otterhound
breastplate
alp
Afghan hound
shoe shop
water snake
cinema
spotted salamander
cellular telephone
American alligator
gorilla
macaque

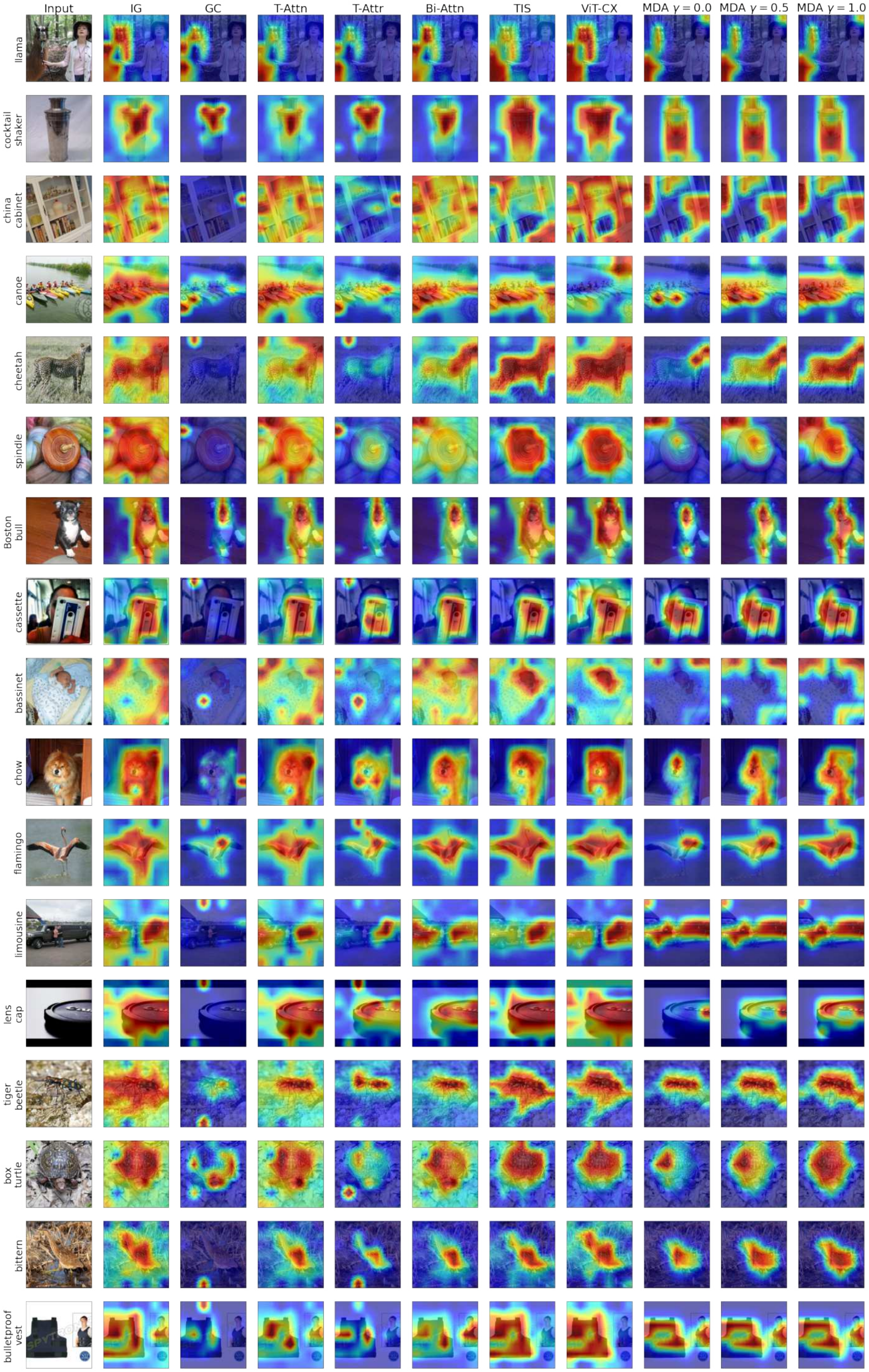

Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
proboscis monkey
albatross
iron
menu
Siamese cat
violin
car mirror
rapeseed
wolf spider
jersey
pug
mailbag
modem
Irish setter
European fire salamander
Brittany spaniel
Ibizan hound

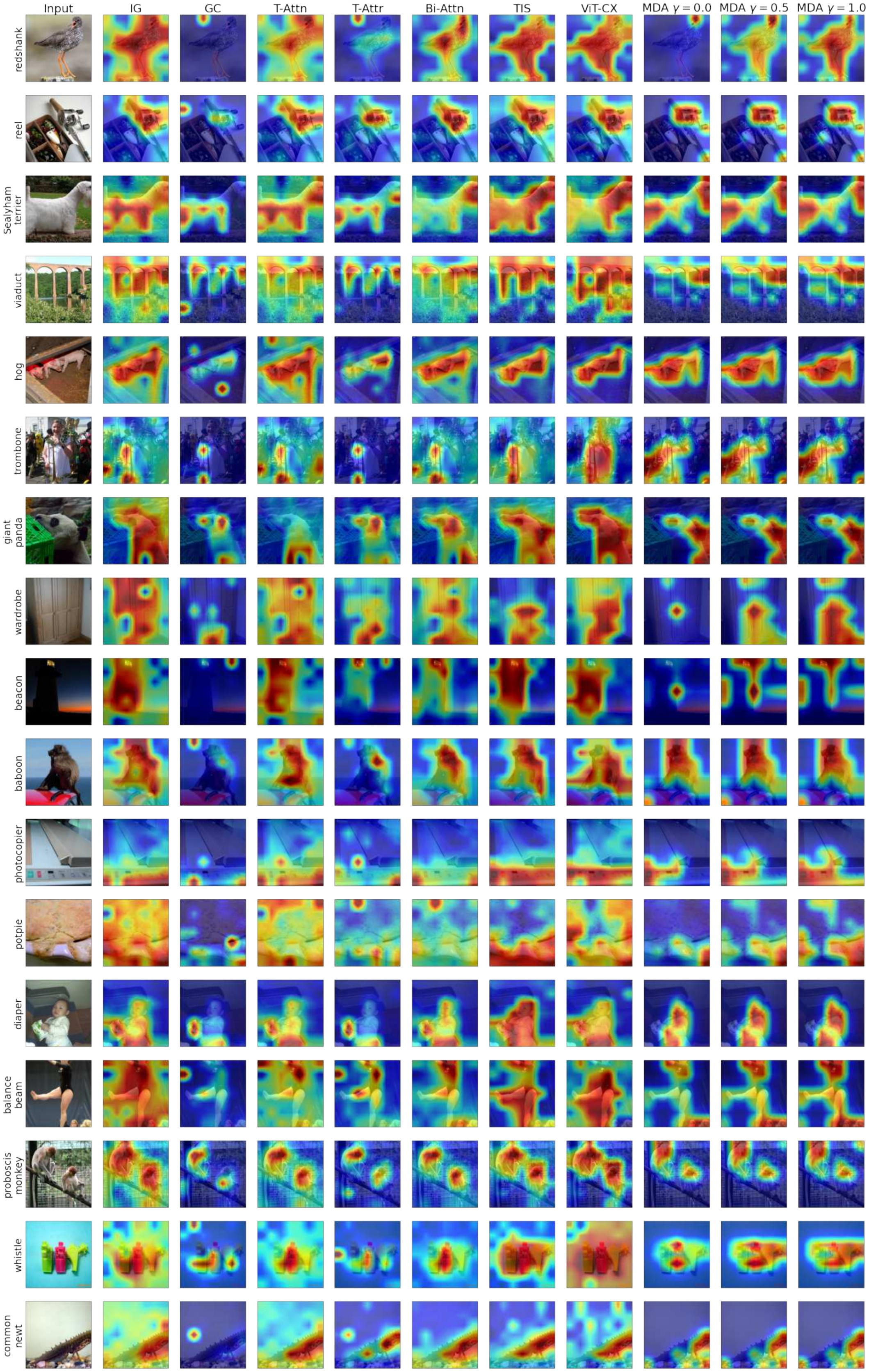

Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
mud turtle
prayer rug
hermit crab
barn
coral reef
goldfish
English foxhound
king crab
bikini
dining table
neck brace
rifle
junco
Staffordshire bullterrier
croquet ball
hip
beaker

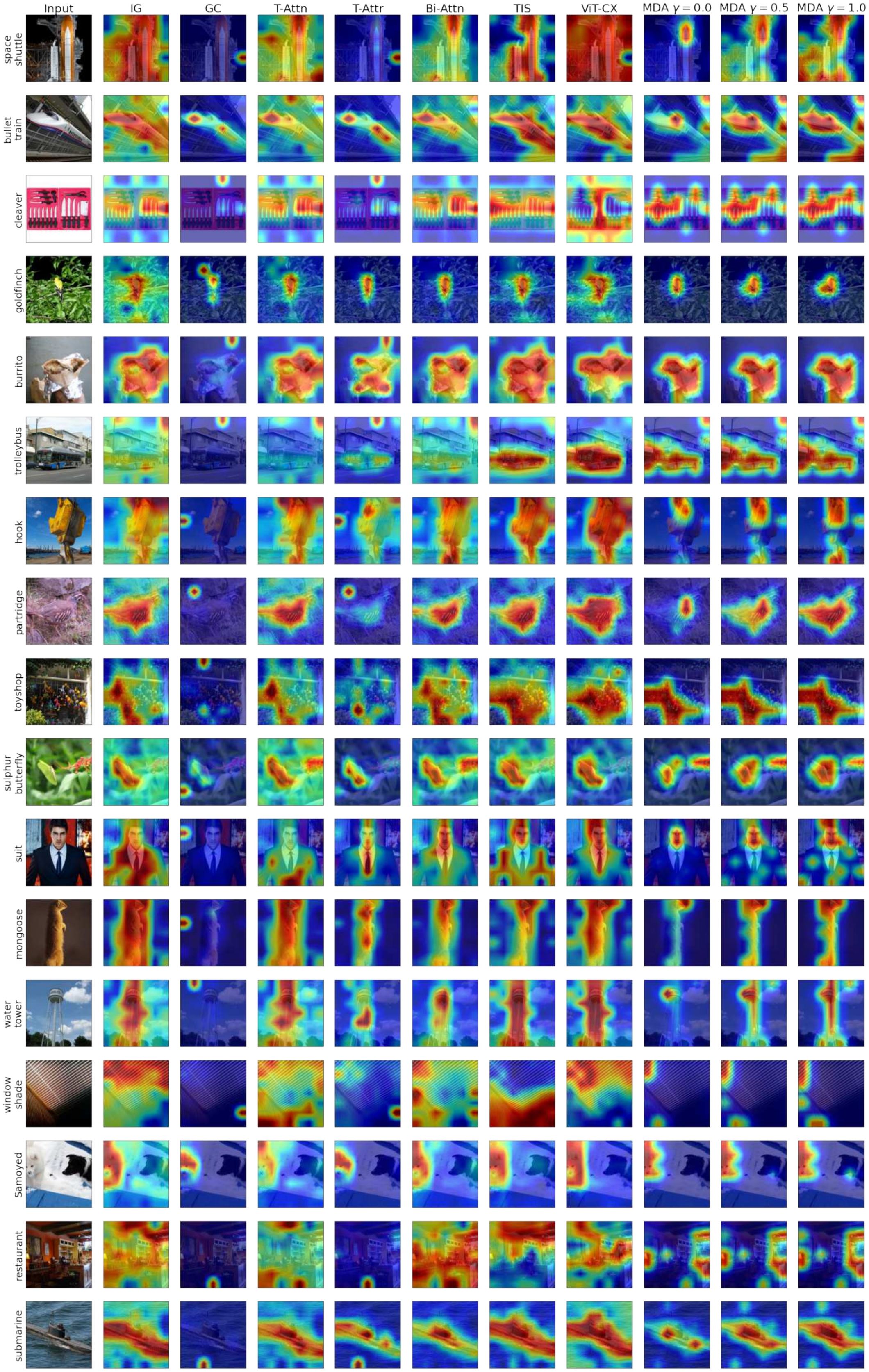

Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
wood rabbit
dowitcher
strawberry
bighorn
can opener
typewriter keyboard
silky terrier
Windsor tie
cello
loudspeaker
jean
American lobster
pill bottle
kelpie
carbonara
maypole
castle

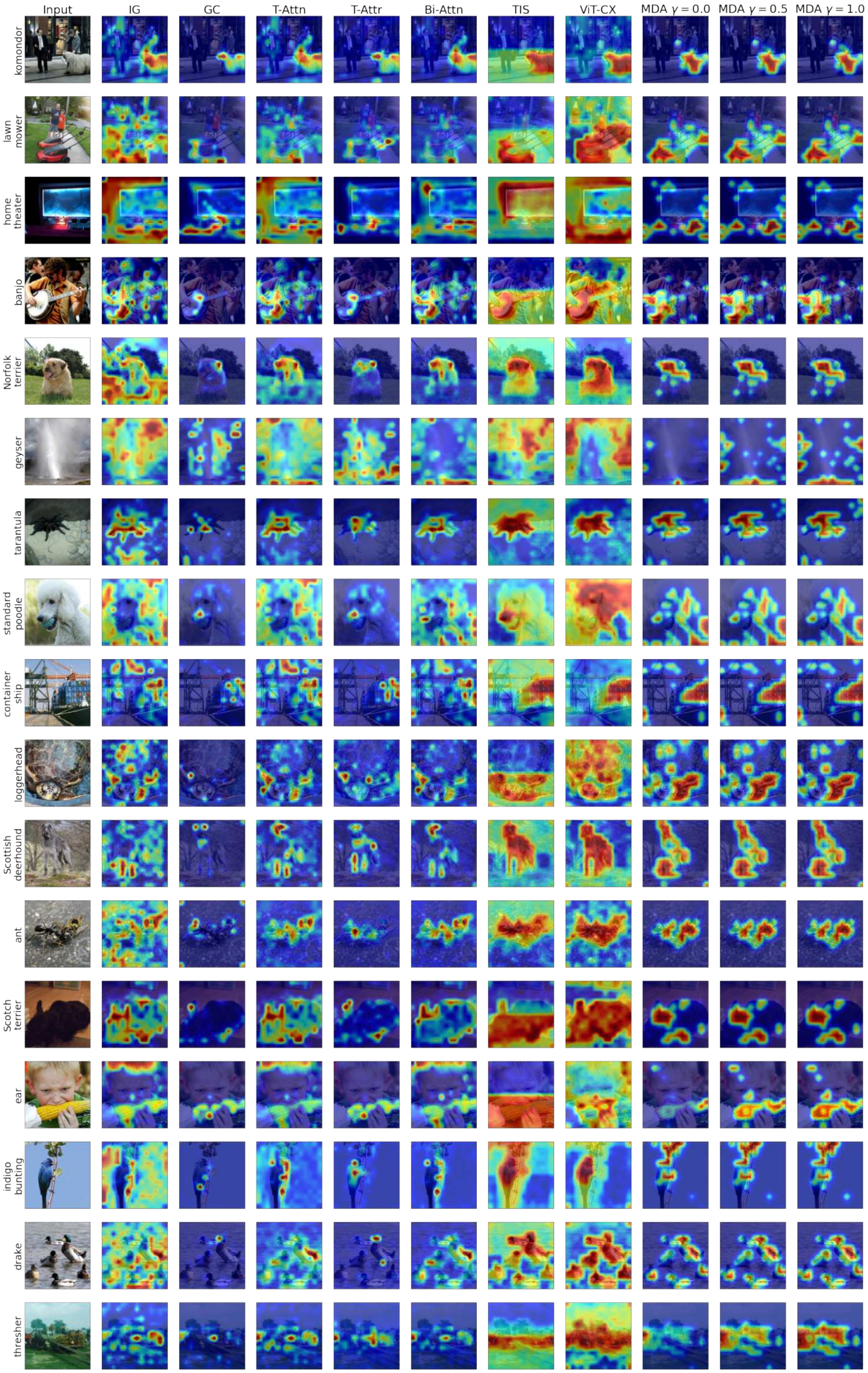

Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
platypus
hair slide
American Staffordshire terrier
park bench
lionfish
bulbul
entertainment center
gyromitra
parking meter
bow tie
Bouvier des Flandres
stove
albatross
iron
menu
Siamese cat
violin

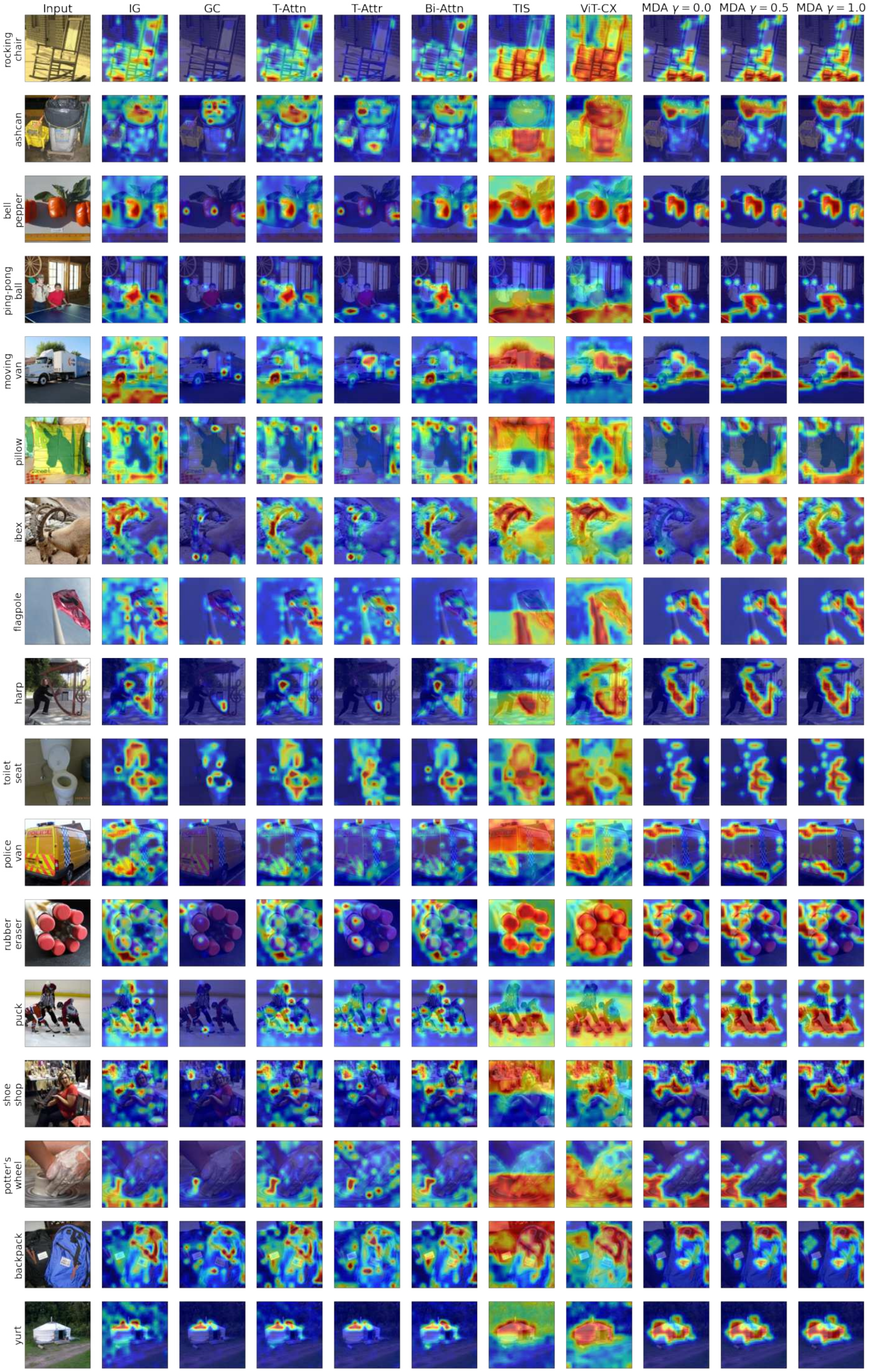

Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA γ = 0.0
MDA γ = 0.5
MDA γ = 1.0
skunk
remote control
washer
jinrikisha
upright
barometer
fireboat
aircraft carrier
trench coat
tench
tractor
greenhouse
frilled lizard
green mamba
trilobite
coral fungus
water buffalo

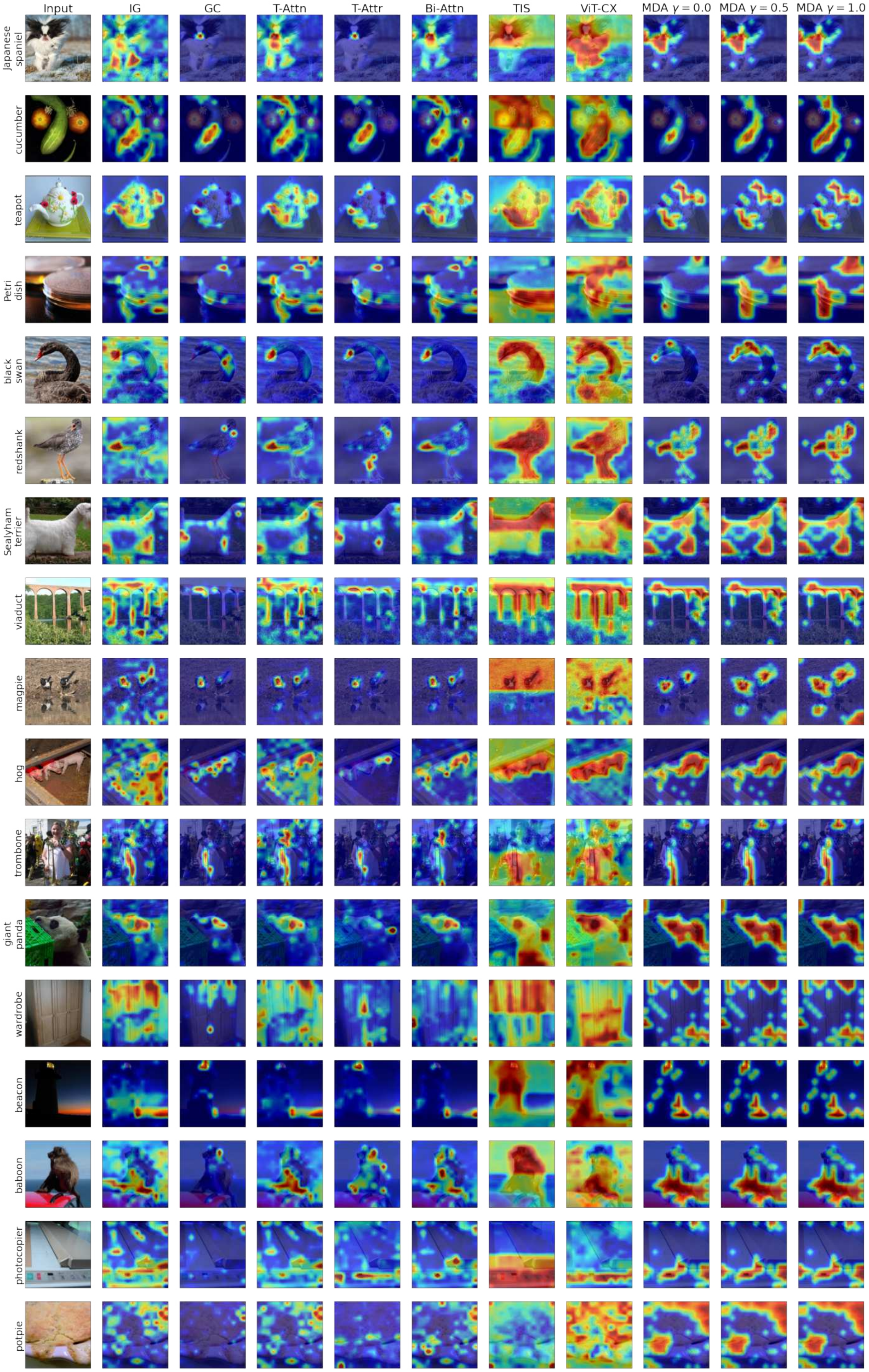

Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
llama
cocktail shaker
china cabinet
canoe
cheetah
spindle
Boston bull
cassette
bassinet
chow
flamingo
limousine
lens cap
tiger beetle
box turtle
bittern
bulletproof vest

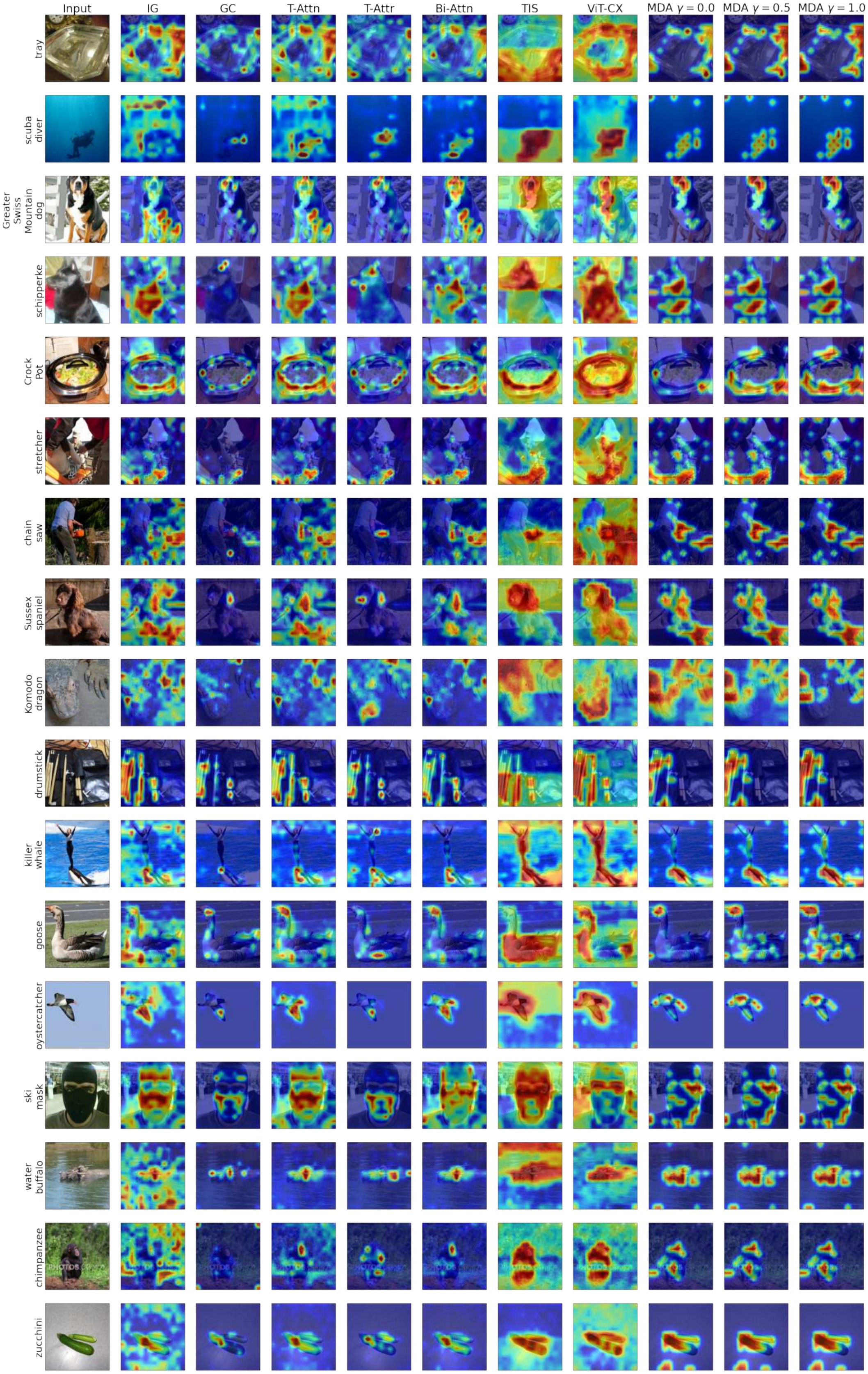

Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
redshank
reel
Sealyham terrier
viaduct
hog
trombone
giant panda
wardrobe
beacon
baboon
photocopier
potpie
diaper
balance beam
proboscis monkey
whistle
common newt

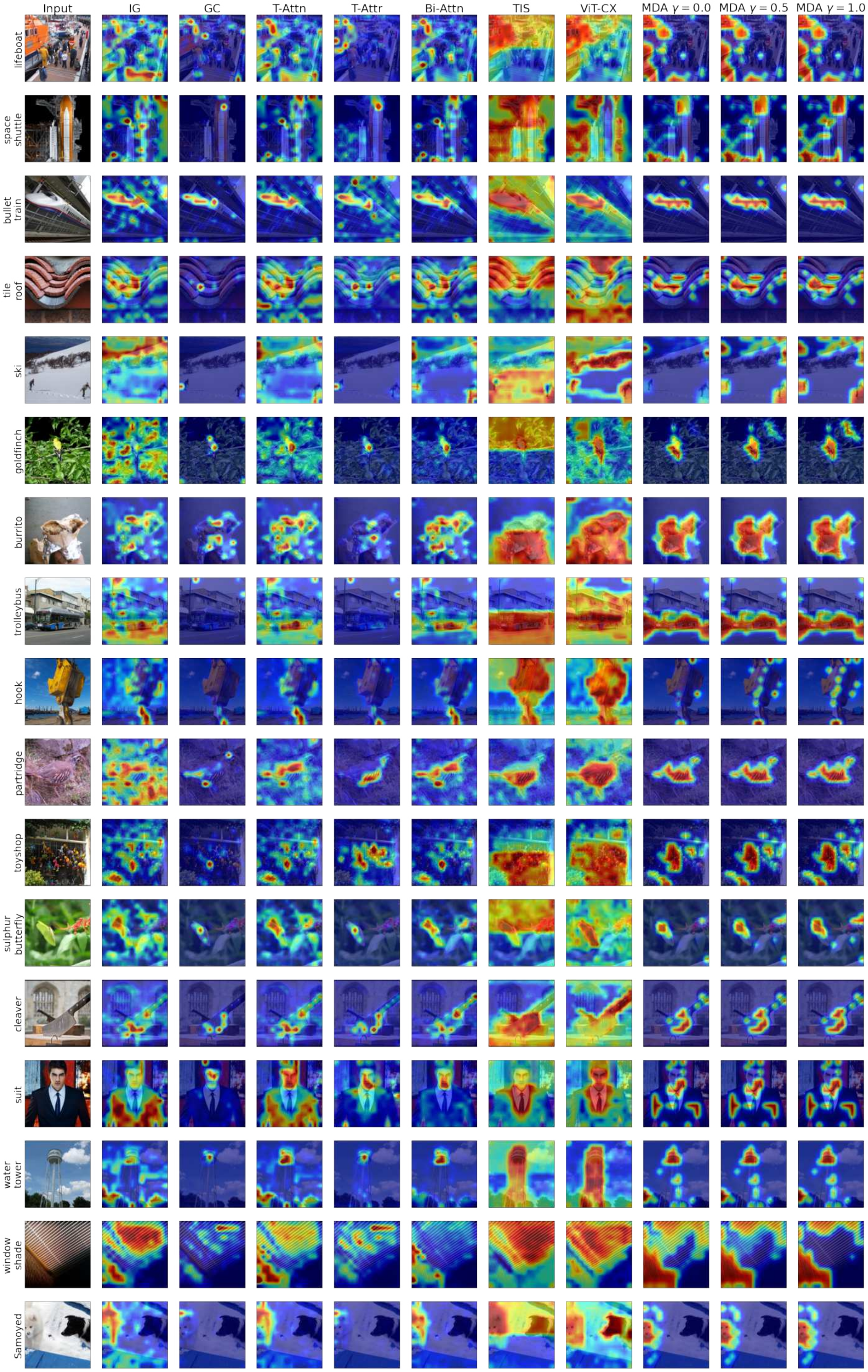

Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
space shuttle
bullet train
cleaver
goldfinch
burrito
trolleybus
hook
partridge
toyshop
sulphur butterfly
suit
mongoose
water tower
window shade
Samoyed
restaurant
submarine


Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
komondor
lawn mower
home theater
banjo
Norfolk terrier
geyser
tarantula
standard poodle
container ship
loggerhead
Scottish deerhound
ant
Scotch terrier
ear
indigo bunting
drake
thresher


Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
rocking chair
ashcan
bell pepper
ping-pong ball
moving van
pillow
ibex
flagpole
harp
toilet seat
police van
rubber eraser
puck
shoe shop
potter's wheel
backpack
yurt


Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
Japanese spaniel
cucumber
teapot
Petri dish
black swan
redshank
Sealyham terrier
viaduct
magpie
hog
trombone
giant panda
wardrobe
beacon
baboon
photocopier
potpie


Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
tray
scuba diver
Greater Swiss Mountain dog
schipperke
Crock Pot
stretcher
chain saw
Sussex spaniel
Komodo dragon
drumstick
killer whale
goose
oystercatcher
ski mask
water buffalo
chimpanzee
zucchini


Input
IG
GC
T-Attn
T-Attr
Bi-Attn
TIS
ViT-CX
MDA $\gamma = 0.0$
MDA $\gamma = 0.5$
MDA $\gamma = 1.0$
lifeboat
space shuttle
bullet train
tile roof
ski
goldfinch
burrito
trolleybus
hook
partridge
toyshop
sulphur butterfly
cleaver
suit
water tower
window shade
Samoyed