In search of a definition of importance: Do attributions capture it?

Antonia Marcu **Jonathon Hare**
Annika Catulli **Damian Smith** **Srinandan Dasmahapatra**
Vision, Learning, and Control Research Group
The University of Southampton
{a.marcu, j.s.hare, abc1g22, ds5n23, srinanda}@soton.ac.uk

Abstract

Attribution methods are extensively used to identify which parts of the input are important for a model's decision. However, despite recent efforts there is still no consensus on what exactly they capture and how to evaluate them. We start addressing this problem by explicitly modelling the implicit assumptions that sit at the foundation of attribution methods. We define two types of importance contributions an input region can have. We then create a framework that decouples the process of **verifiably** establishing contributions from that of evaluating attributions. This shows that current attribution methods struggle to reliably capture input contributions. Further investigation shows a stark difference in attribution performance between model behaviours, exposing a previously overlooked aspect in the attribution literature. Our work calls for clear, quantifiable statements about what attribution methods aim to capture, along with rigorous evaluation frameworks.

1 Motivation

Attribution methods seek to uncover what parts of an input are the most "important" for a model's decision. In this paper we focus on image classification tasks. Despite early attribution methods for vision problems having been introduced over 10 years ago [e.g., 38], the community is still far from having faithful attributions [e.g., 48, 30, 44, 29, 1, 7]. We argue that what is needed for a long-awaited breakthrough are clear definitions for importance, coupled with controlled benchmarks. Whilst steps have been taken in this direction [29, 21, 4, 6], we argue that key ingredients are still missing. In this paper, we search for a definition of input pattern importance that is decoupled from attribution evaluation. To do this, we identify some of the assumptions that are at the foundation of attribution, challenging some of them and explicitly modelling others. Then, we propose two definitions of importance. To quantify these, we propose a carefully designed framework. We then illustrate how our framework can be used for attribution evaluation by analysing the extent to which existing attributions capture our notions of input importance in a number of example settings. Our analysis shows that existing attributions struggle to consistently capture our notions of input importance, despite significantly relaxing our evaluation setting in their favour.

Problem: Many different attribution methods have been proposed and systematically reviewed [e.g., 18, 41, 29, 1]. While it is common in the literature for the output of attributions to be interpreted as the "importance" of image regions [e.g., 45, 6, 22], it is unclear what this means, with an overwhelming majority of methods lacking explicit definitions. This makes it difficult to rigorously evaluate and benchmark them. A common approach to evaluating attributions involves transforming the input whose classification we want to explain, and then measuring the change in output. As noted [e.g., 23, 22, 49], these transformations often take the input out of distribution. Moreover, the methodologies typically require design decisions that impact the results of the evaluation, with no consensus on

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

how to choose these. Therefore, **it is difficult to draw reliable and clear conclusions based on typical attribution outputs.** This is both because, in the absence of a definition, their output cannot be clearly interpreted and because it is challenging and unclear how to verify what they do capture.

Solution: We argue that one of the root causes is that implicit definitions of input importance are entangled with attribution evaluation. This leads to a lack of specification. In this paper, **we disentangle input importance from attribution evaluation**. We first focus on quantifying the contribution of an input region for a model's decision. We argue that input parts can contribute to a model's decision in several different ways, and therefore the concept of "importance" should be broken down into a number of measurable quantities. We propose **definitions for two aspects of input importance**. One aims to capture the contribution of an image part when seen in isolation (individual contribution). The other captures the contribution of a part when seen in the context of the entire image (contextual contribution). Computing these, however, requires performing certain transformations to the input data. To perform these transformations in a way that doesn't make unverified assumptions about what the model learns from the data, we need a number of conditions to hold. We refer to these as "validity conditions". Starting from these conditions, we use **controlled settings** to create a framework in which contributions can be reliably established. It is only then that we evaluate how closely attributions capture the defined contribution. We explore the extent to which popular attributions implicitly capture importance in terms of these contributions.

Methodology: When creating controlled settings we aim to model aspects of real datasets, while at the same time leading to as many cases that successfully pass the validity conditions as possible. This is to ensure that we have a sufficient number of instances in our analysis where contributions can be reliably established, such that conclusions can be empirically substantiated. We start by **modelling settings based on underlying assumptions behind current attribution methods**. One such assumption is that an input image can be seen as a composition of subsets of pixel values that carry different levels of information for the model, including subsets that carry negligible information [e.g., 21, 6]. We generate a family of settings which allow us to obtain models with different behaviours. Then, we apply two validity conditions designed to **filter out those cases for which our modelling assumptions do not hold**. In those cases, contributions cannot be readily established, and therefore they are removed from our analysis. Importantly, we link our validity conditions to a number of observations, highlighting limitations of prior attribution evaluation settings. For the cases where our modelling assumptions are verifiably in line with the model's behaviour, we establish input contributions. Then, once these contributions are reliably established, we can use this framework to explore the extent to which attribution methods capture these aspects of importance in some example settings.

Impact: To the best of our knowledge, our framework is the first to define and **verifiably** establish the contribution of an image region. This framework allows us to carry out an initial exploration of popular attributions and characterise the types of rules learned by models, providing a more informed and structured analysis. Our primary objective is not to propose an exhaustive evaluation for attributions. Instead, we aim to illustrate the usage of our framework, which we see as a stepping stone towards systematic evaluation. Our findings indicate a potential need to rethink the sensitivity-centric way in which attributions and their evaluations are approached. While we see our work as broadly complementing existing tests for attribution [e.g., 1, 45], we believe it provides an entire platform for attribution evaluation development. We believe that having attribution methods that are reliable in controlled settings is a necessary step towards more trustworthy attributions. Our framework calls on the community to build more benchmarks for controlled evaluation.

Our contributions are:

- We decouple the process of establishing importance from attribution evaluation and define two notions of importance: the contextual and individual contribution. (Section 2.2)
- We create a framework to reliably compute input contributions in controlled settings (Section 3). We observe that: 1. One image part can be the most important in a contextual sense, while another can be the most important in an individual sense. Definitions of importance are needed when attributions are proposed. 2. Assumptions about what the model has learned based on data set constructions might not hold. Importance needs to be verifiably established.
- We explore the extent to which existing attribution methods capture contextual and individual contributions (Section 4). We find that: 1. Attribution success on these tasks is dependent on

the learning setting (e.g., learned diversity), with poorer raw performance[1] in settings where there isn't a single input pattern that dominates learning. 2. A lot of the apparent success of existing attributions could be due to an overlap between sensitivity and importance caused by the coupling between attribution evaluation and establishing importance.

2 The importance of an image part for a model's decision

While the notion of importance is central to attributions, its definition largely remains unspecified. In this paper, we search for a definition of importance starting from common assumptions and approaches in the field. We propose two complementary quantities that relate to how importance is implicitly defined in the literature and explore the extent to which attributions capture those notions. Our objective is not to propose an exhaustive evaluation of attribution methods. Instead, we analyse the extent to which they align with potential definitions of importance implied in the literature.

We separate the processes of empirically establishing the importance of each image part from evaluating the attribution, treating them as two independent steps. This allows us to focus on how image parts contribute to a model's decision, rather than on desirable qualities of an explanation [e.g., 1, 4, 21, 29]. More importantly, by focusing on quantifying the interaction between image parts and the model's decision, we are able to propose definitions of importance in a way that is agnostic to attribution assignments and their evaluation.

In contrast, the attribution literature often links identifying the importance of input parts with attribution evaluation. For example, most insertion/deletion protocols do not first establish the importance of image parts, and then independently quantify the extent to which attributions capture these [e.g., 30, 24, 12]. Instead, these protocols are purely comparative. As such, they aim to evaluate when an attribution captures a "better" notion of importance, not when importance has been captured. Appendix A provides a broader discussion of the relevant literature. We believe that decoupling the specification of what importance is, from an attribution method that seeks to recover it, is key to advancing the field. To do this, we start by conceptualising images in a way that allows us to decouple the two. We start from ideas and intuitions implicit in the literature.

2.1 Conceptualising images

Despite variations in framing, attributions start from the assumption that not all parts of an input contribute equally to a given model's decision. For example, it is widely considered in the literature that if an image part is important, either deleting it from or adding it to the input should have an effect on the model's output [e.g., 18, 2, 36]. Implicit to this is the assumption that an image can be broken down into parts that contribute to the model's decision to different extents. This contribution is typically thought of as the "importance" of each image part. Based on this assumption, the objective of attributions is to help us understand which parts of the input are more important than others for the model's decision. Moreover, in computer vision, it is commonplace to conceptualise images as compositions of objects. We abstract away from the concept of "objects" and simply consider images as compositions of input patterns, which can be thought of simply as pixel subsets.

Therefore, **we consider an image to be a composition of non-overlapping pixel subsets, and a model's decision to be an operation which aggregates the features extracted from one or more pixel subsets.** Importantly, the input subsets that contribute to a decision could, but **do not have to align with human-interpretable concepts.** We also do not make any assumptions about the general nature of these subsets. We do not require that the subset represents a connected component, and it can be of arbitrary size, including containing a single pixel (effectively leading to a pixel-based view). Starting from this conceptualisation, we next define two aspects of importance.

2.2 Contextual and individual contribution: aspects of decision importance

Common practice in attribution evaluation fundamentally boils down to "adding" and/or "deleting"[2] pixels to attest their impact on a model's decision [e.g., 45, 49, 44, 30, 46, 23]. We argue that

[1] As opposed to performance relative to a *choice* of baseline. Full discussion in Section 4 and Appendix H

[2] In practice, this means replacing the RGB values with another set of RGB values. In our framework, we pay particular attention to how these values are chosen.

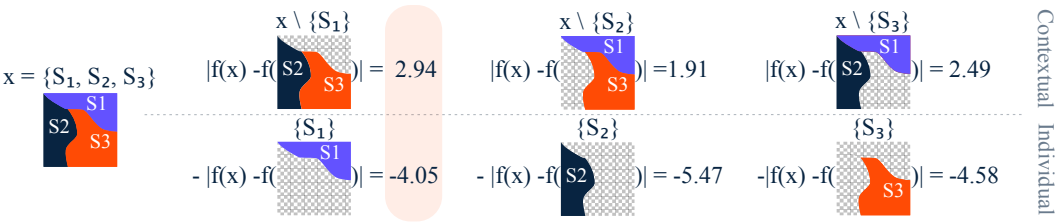


Figure 1: Conceptual illustration of establishing importance. For each of the three subsets that make up the image, we compute contributions. The subset that has the highest contribution in both a contextual and individual sense, S_1, is denoted the highest contributor. Note that in this illustrative example we use a chequered pattern to represent a pattern that contributes negligibly to a given model's decision (i.e. to indicate no visual information) and it is not the pattern we use in our experiments. Our framework (Section 3) pays particular attention to how these transformations are carried out and Figure B.2 provides a concrete example.

this is rooted in the intuition that at least two scenarios need to be accounted for: 1. How the subset contributes to the decision when seen next to all other subsets present in the input (contextual contribution) and 2. How the subset contributes to the decision on its own (individual contribution).

Given a model f and a sample x, we write the model output for class c as $f_c(x)$. Let the image x be composed of k disjoint subsets of pixels with the i-th subset denoted S_i. We define the contextual and individual importance of subset S_i for the original classification decision $\hat{y} = \arg\max_c f_c(x)$ as:

$$I_{ctxt}(S_i) = |f_{\hat{y}}(x) - f_{\hat{y}}(x \setminus S_i)| \qquad I_{indv}(S_i) = -|f_{\hat{y}}(x) - f_{\hat{y}}(S_i)|\,.$$

Intuitively, if removing subset S_i leads to a large change in the model's prediction, then the model must have relied on the presence of that subset for its prediction. Note that we take the absolute value of the difference, considering all possible contributions to a decision, both positive and negative. Conversely, if only having subset S_i present in the image gives a prediction that is close to the model's prediction on the original sample (small output difference), then we consider subset S_i to be a strong individual contributor to the decision. We multiply the absolute difference by -1 such that a **higher** contribution (smaller gap) corresponds to a **higher** $I_{indv}(S_i)$. Note that in our analysis we use the relative contributions of the subsets rather than the raw contribution values (see output considerations below). We provide a visual illustration of this process in Figure 1.

Input considerations: Note that these definitions require the ability to remove one or more pixel subsets from an image without introducing new information, or the model treating the transformed sample as out of distribution. These are extensively discussed challenges faced when aiming to establish the importance of a subset [e.g., 49, 23, 22] and will directly inform the design of the framework we create in Section 3. We also design a series of validity checks to filter out those cases where contributions cannot be reliably established due to transformation artifacts.

Output considerations: From the output perspective, there are a number of factors to consider, such as the model output type (e.g., logits or softmax probabilities) and selection (e.g., predicted class, true label, or aggregate across classes). Just as with contextual and individual contributions, we argue these provide alternative insights into models, with no universally "correct" quantity. We believe new attribution methods, however, need to be explicit about the type of contribution they aim to capture. In line with most attribution analyses, for the selection of output, we focus on the predicted class. For the output type, we consider both the logits and the softmax probabilities in the main paper. To allow for this flexibility, we use relative contributions rather than raw values.

Existing attribution methods do not specify which type of contribution they target. To not disadvantage attributions that capture one type of contribution and not the other, in the main paper we carry out a simplified, more inclusive analysis. For this, we only consider those samples for which there is a consensus on subset contribution according to both the individual and contextual contribution, for softmax and logits. Therefore, irrespective of whether an attribution captures one type or the other, it will be considered correct. We do a more in-depth analysis where we look at the cases when they conflict, to attest whether an attribution has a tendency to capture one type of contribution over the others (Figure F.1) but conclude that this is not the case. For a full discussion, see Appendix E.1.

3 Framework

Attesting the contextual and individual contributions of pixel subsets requires a careful consideration of the transformations that need to be performed. We create a series of controlled data sets designed to encourage models to learn rules that fulfil our requirements. We then design a series of "validity checks" to verify that the requirements hold. We therefore train a series of models on our data sets and then filter out those cases that do not pass our checks. These are the cases where we cannot reliably establish input contributions, and therefore are not suitable for an analysis.

Since we are looking at the decision of a model on a given image, our definitions of contribution are dependent on properties relating to the **sample-model** pair. For example, we require the model to be invariant to certain transformations of the given input image. If the model is not invariant to the transformations needed to establish the contribution of that sample's constituent subsets, then we cannot reliably isolate subset contribution on that sample, for the given model. Therefore, given that our analysis is concerned with the relationship between the sample and the model, we will start with assumptions about how the model acts on the sample from a data construction perspective and use the validity checks to establish if the assumptions hold from a model output perspective.

In distribution from a sample-model perspective. The biggest challenge in establishing subset contribution is ensuring that the transformations preserve the distribution. It is well-known that addition/deletion transformations can become out of distribution [e.g., 49, 23, 22]. As a result, the change in logits cannot be solely attributed to the added/removed information, as it is expected that the distribution shift also contributes to a change in logits. However, there is no clear protocol for establishing when addition/deletion transformations are out of distribution. In fact, identifying when a sample is out of distribution is still an open research question and depends on multiple factors such as the model's training setup and calibration, the nature and magnitude of the shift, etc. Since we focus on the sample-model relationship, in our framework we start by considering what it means to be "in distribution" from a sample perspective and aim to validate it from a model's perspective. In Section 3.2 we will make a number of design decisions to simplify and concretise this validation.

Transformations only add negligible information. One way of removing a subset of pixels without introducing additional information is to have an input pattern that the model is effectively invariant to (i.e. contributes negligibly to a model's decisions), and replace the targeted subset with that pattern. As a starting point, to be able to apply the transformations needed for establishing subset contributions, we encourage the model to be effectively invariant to input subsets depicting random noise. Those sample-model pairs for which we find the random noise contributes to the decision are excluded from our analysis, as they do not allow a clean isolation of the effect of removing a pixel pattern.

3.1 Design

We aim to create settings that model aspects of real-world datasets, while having as large a number of sample-model pairs that fulfil our requirements, such that our analyses can be conclusive. We start with a simple data set, The Squares Problem, which gives the blueprint for our framework. To make the input partition as straightforward as possible from a data perspective, in The Squares Problem each image consists of up to three coloured squares on a black background, with the colour of the squares correlating to the label to different extents. While this is a very simple data set, we argue that it captures more general sample-model interactions. For the purpose of our analysis, each square simply represents a variable. It could, in principle, stand in for an arbitrary input pattern: an object, part of an object, a mixture of objects, or objects and backgrounds. We start out with coloured squares simply because they allow us to easily perform our validity checks, while abstracting away from many problem specifics and therefore being highly generic. We consider more complex patterns in Appendix G.5.

We would like to be able to obtain families of models with different characteristics, allowing us to study different behaviours that are likely to occur in practice. We do this by creating probabilistic rules for generating data samples. Importantly, training models on such data sets is not sufficient to make any inferences about what models have learned. Instead, we will carry out controlled interventions to attest whether validity requirements are respected such that we can reliably establish the contributions of each subset. To model this, we want to create a data set for which given a class c, every sample is composed of multiple pixel subsets that **correlate** with the label to different extents. Therefore

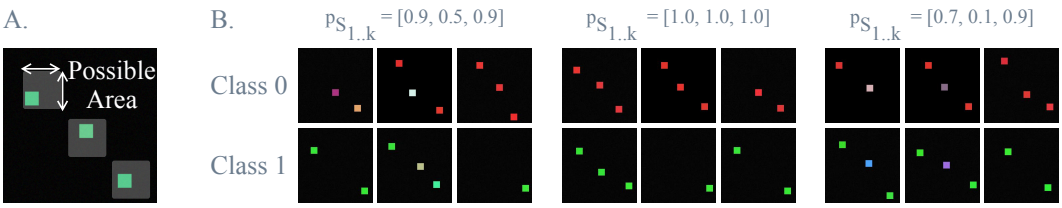


Figure 2: A. Visual illustration of the data set creation. In each of the three possible areas (60×60 px, highlighted in white), we randomly place a square i (20×20 px, shown in green), whose colour correlates with the label with probability p_{S_i}. B. Examples of three randomly-selected samples from training examples belonging to classes 0 and 1 for three different colour correlation settings. In this example, class 0 is associated with the colour red, while class 1 with the colour green. Therefore, deviations from these colours stem from the process of assigning a random colour with probability $1 - p_{S_i}$. In the $[1.0, 1.0, 1.0]$ correlation case, all squares are variants of red and green respectively.

the task is probabilistic. From a classification task perspective, to model realistic settings we allow a small percentage of the data to be mislabelled.

On a black 224×224 canvas we consider three 60×60 possible areas. In these areas, given a class, with a probability p_{S_i}, we will draw a pixel subset S_i that depicts a pattern that correlates with the class, and with probability 1 - p_{S_i} this pattern will be random. We choose these subsets to depict 20×20 squares, but they could be arbitrarily shaped. We also choose the correlation with the class to be given by the colour of the square. Therefore, with probability p_{S_i} the colour of subset S_i will be correlated with the label. We add a small amount of random noise to the entire image. We experiment with data sets that have different correlation probabilities, obtaining models with different behaviours, as described in Section 3.4. We also consider varying the colour variance for each class (results in Appendix G.2). Figure 2 provides visual examples of the images obtained. We train using 60000 samples and our test set has 10000 before the validity checks. For completeness, the test samples have all three squares present before applying the transformations.

Transformations only add negligible information: To encourage sample-model pairs to satisfy this requirement, we create the background pattern and the pattern that replaces a square when it is "removed" to be a black background with a small amount of random noise. As such, both the background and the values that replace the coloured square should add negligible information towards the decision. Sample-model pairs for which this does not hold will be eliminated from the analysis, since in those cases the change in output cannot be entirely attributed to the removal of the square.

Samples with removed squares are in distribution: Encouraging sample-model pairs to satisfy this requirement, in the training data we skip drawing each square with a 10% probability. As such, the training data will have samples with missing squares. We argue that this is reflective of real-world scenarios, where not every image instance will depict all patterns that can correlate with the label.

3.2 Establishing that requirements are satisfied

We next aim to verify that our requirements are satisfied, contributing towards a protocol for establishing subset contributions. We propose a number of "validity checks" which are part of our framework and we apply these to all sample-model pairs. Importantly, our validity checks depend on the data sets we create, while the principle behind them remains. Extending the framework further requires a careful consideration of any additional checks that might be needed to ensure that potential confounding factors have been accounted for. Figure B.1 provides a visual illustration of the checks we apply as part of the framework.

Our first requirement is that the transformations we carry out for establishing the contextual and individual contributions are in distribution from a sample-model perspective. To check this requirement, we first need to divide each image into pixel subsets. While from the data-generating perspective the input can be divided into subsets that correspond to the three squares, we need to verify that this is consistent with patterns the model attends to. That is, we do not want to assume that the model does not attend to areas of the background simply because the data is constructed such that the background is not predictive of the class. We instead want to verify that, for an image instance, the exact value of the background noise does not significantly impact the prediction. If this does not hold, then we are

unable to divide the sample into subsets in a computationally-feasible way and reliably establish their contributions. We therefore exclude those sample-model pairs from our analysis.

This leads us to our first validity check **VC1**: Does varying the noise outside of the three squares (i.e., background noise) lead to a non-negligible change in the model's prediction? We regenerate the background noise 10 times and verify that the predicted class stays the same on the newly obtained versions of the sample (i.e., varying the noise never changes the prediction). We then check that all obtained changes in predicted output are smaller by at least an order of magnitude compared to a baseline. We choose the baseline to be the contribution of a square, as detailed in Appendix B. One limitation of our check is the need for a "negligibility" threshold.

Applying this check, we find that over 75% of the models have samples on which the background contributes non-negligibly. We provide full filtering statistics in Table F.2. Therefore, our first observation is that **O1: despite the background information not being predictive of the label, models can attend to it**. Therefore, checking that patterns expected to be uninformative are indeed not informative for models is a necessary step in the evaluation frameworks that rely on this to hold. Note that we do not claim this is a novel observation. We emphasise it here because it is often unaccounted for in widespread approaches in attribution literature, and because it is central to our approach.

Computing contributions requires removing one or more squares from the image. However, in practice it is not possible to remove a subset of pixels from an image. Instead, the subset is replaced by another subset of pixels. To capture the effect of removing the square, we need to replace it with a pattern to which the mode is effectively invariant. In our framework, the background "pattern" (i.e., black background with random noise) is the same as our "negligible information" pattern. Therefore, we replace the square with the "negligible information" pattern. Importantly, designing the data set such that a pattern corresponding to random noise is not predictive of the label does not exclude the possibility of the model using that pattern in its predictions. This leads to our second validity check **VC2:** For all sample-model pairs we verify that the specific instance of random noise used to replace a square does not significantly impact the output of the model. For this, similarly to VC1, we verify that resampling the random noise 10 times that replaces the square colour when we "delete" the square only changes the model's output negligibly.

3.3 Establishing the highest contributing subset

Having passed our validity checks, the sample-model pairs can be used to establish subset contribution. Given a sample, for each of the three squares we compute the individual and contextual contributions by evaluating the change in the model's output on the different transformations. For an illustrative example, see Figure B.2. Our first observation is that for all models models we obtain, **O2: there exist samples for which one square is the highest contributor in a contextual sense, while another is the highest contributor in an individual sense** (both in the softmax and logits case).

In Section 4 we use our framework to evaluate attributions. However, cases where there is not a single highest contributor according to all definitions are problematic for existing attributions, since, in the absence of a definition of importance, it is unclear which of these they aim to capture. To avoid penalising attributions for not specifying which contribution they capture, we consider a subset to be the "most important" if it has the highest contribution in an individual and contextual sense, according to both logits and softmax probabilities. We refer to this as the highest contributing square. To maintain consistency throughout the paper, we will only present results on those sample-model pairs where a highest contributing square exists. As such, attributions that implicitly capture either of the two definitions would still do well on our analysis. In Appendix F, we perform a separate analysis where we investigate cases where there isn't a single highest contributor according to all definitions.

3.4 Obtained behaviour regimes: diverse versus focal models

We train models across 4 architectures: ResNet-18 [20], ResNet-50 [20], VGG-16 [37] and ViT-B/16 [16]. The main results we present are for 4 different $p_{S_{1..k}}$ values. Unless otherwise stated, we aggregate results over models trained on all these settings. For training details, see Appendix D. For each setting we consider, we train 5 models with different seeds. For completion, we report training and test accuracies (Table F.1) as well as the number of test samples that passed the validity checks (Table F.2). In Appendix F.1.1 we include a brief investigation of the impact of increasing the number

of training samples on the elimination rates for the validity checks. Unless otherwise stated, all our results are computed on the test data. Importantly, note that our filtering process is a post-selection, which, if not treated carefully, can impact our conclusions. However, we confirm that our conclusions hold for models that passed all the validity checks, as much as for models where more than 70% of the samples are eliminated by the checks. In Appendix G.6 we carry out further checks to ensure our conclusions are not a result of the post-selection process.

We hypothesise that some attributions tend to perform better in terms of raw performance on models that predominantly rely on a single square for their predictions (focal) compared to models that learn more diverse classification rules (diverse). To investigate this, we want to identify models for which, across the test set, one of the squares has a consistently higher contribution than the others. Given a model and a data set, we compute the percentage of times each of the three squares is the highest contributor. We plot a histogram of the highest value per setting in the right-hand side of Figure 3. We denote models for which one of the squares was the highest contributor in over 65% of the test samples as focal (orange) and the rest as diverse (blue). Note that this is the only time in our analysis where we characterise a model by its average performance across the entire test set. One limitation is that that we needed to set a threshold to divide the two categories. However, considering the distribution of contributions for the two settings, we obtain an average entropy of $48.14_{\pm 12.90}$ for the focal models, as opposed to $84.32_{\pm 14.67}$ for diverse models, confirming that the two categories capture the intended behaviour. A careful treatment of the exact threshold value is necessary for future investigations where the behaviour description is a crucial element. In this work, we simply aim to show that such classes of behaviour can be obtained and that attributions' performance can depend on the model's behaviour regime.

4 Evaluating attributions

The strength of our framework is that we can define and establish importance in a way that is decoupled from an attribution assignment. So far we focused on defining and establishing aspects of importance. We now focus on attribution evaluation. Our objective is not to highlight the failures of current attribution methods, but to better understand them and their implicit definition of importance. As a result, we make various decisions that favour attributions, in an attempt to create a generous evaluation. We will refer to these decisions as "relaxations" of the setting. Nonetheless, despite relaxing the evaluation framework, we find that existing attributions fall short of consistently capturing one type of contribution, especially in settings where models are not overly-reliant on a single pattern.

We perform a preliminary evaluation of popular attribution methods (e.g., Grad-CAM [35], variants of IntegratedGradients [40]). We focus on heatmap-producing, instance-level attributions for networks with fixed weights. Due to their widespread adoption, we focus on post-hoc methods for explainability of classical "black-box" models, rather than architectures designed to be "inherently explainable" [e.g., 13, 9, 3]. In total, our main results span 80 models (4 architectures, 5 model seeds, 4 data sets) on which we evaluate 15 different attributions, leading to 1200 attribution-model evaluations.

We create a series of simplified tests for attributions. In the first test, rather than quantifying the extent to which attributions correctly rank subsets, we ask how often attributions correctly identify the highest contributing subset. We start by considering an attribution to be correct if its maximally attributed pixel is in the possible area for the square. We observe, however, that in a non-negligible proportion of the instances, the maximally attributed pixel is outside of the three possible areas Figure G.1. **That is, despite the background having a negligible contribution to the model's decision as per our VC1, there are many instances in which attribution maximally assign importance to the background.** We consider these to be attribution failures.

Since the proportion of such failures is significant for some models, we make our evaluation even more accommodating of existing attributions. We set these instances aside, and **report results only for those attribution instances where the maximum attribution is within one of the squares**. We present these results in Figure G.2. Then, relaxing the evaluation even further in favour of existing attribution methods, we propose a second case where we look at the square with the highest average activation across its pixels. Note that both of these cases don't capture failures that result in attributions outside of the square, therefore significantly increasing the probability of an attribution being correct by chance. All main results in the paper correspond to this setting, since we find it to be the most advantageous for existing attributions. Therefore, for the main results of the paper we

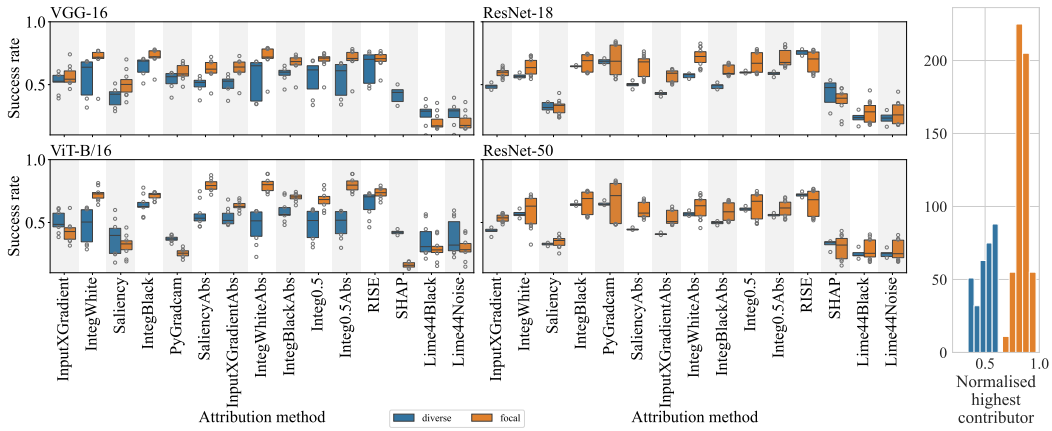


Figure 3: Left: Success rates for different attributions, grouped by behaviour regimes (diverse versus focal models). Shaded bars denote the interquartile range, and the line across is the median. For most attributions, we observe a clear gap in performance between the two regimes, with an accentuated gap visible in less relaxed settings (Figure G.2). Right: Histogram of the average highest-contributor value per model. A large value signals that one of the squares dominates (i.e., the model is focal).

first compute the average attributed value within the square, with a 10px buffer in every direction. If the highest contributing square was assigned the highest average attribution, we consider it an attribution success. Evaluating the attribution on identifying the most important square is a deliberate simplification, and makes the analysis equivalent to The Pointing Game [47]. Appendix E.2 provides sample results for ranking correlation.

We first find that in many cases **O3: there are stark differences in raw attribution performance between models from different behaviour regimes** (Figure 3). Based on raw attribution success, it might be concluded that attributions tend to perform better in focal as opposed to diverse scenarios. Our objective is to provide a **platform** for attribution evaluation rather than benchmarking attribution evaluation as such. For the purpose of benchmarking attribution performance, it is beneficial to define a baseline to compare against. Since in this section we restrict the evaluation to those attributions that fall within one of the three squares, at least two possible baselines can be considered: 1. Uniformly at random picking one of the three squares or 2. Always picking the same square and that square being, by chance, the highest contributor. Relative to baseline 1, attributions typically perform better on focal as opposed to diverse models. Relative to baseline 2, attributions perform better on diverse as opposed to focal models.

Baseline 2 sets a higher standard for focal models, which are, by definition, those models for which one of the squares is the highest contributor in over 65% of the samples. However, it requires either prior knowledge which is not available in real-world scenarios, or making a lucky guess. Irrespective of the choice of baseline, the gap in attribution performance on the two types of models is important, and we discuss its implications in Section 5.

Raw attribution performance highlights an important concern. In real-world scenarios one typically does not know if a model is focal or diverse. As such, one cannot know whether the attribution they are using can be trusted on average 78% of the time or 57% of the time, for example (as is the case of Saliency on ViT-B/16 in our setup). It is therefore important to create attributions that perform consistently across model behaviour.

Our framework makes a series of relaxations in favour of existing attributions (e.g. setting aside the cases where the highest attributions fall outside any of the squares or the regions surrounding them, reporting square rather than pixel-level performance as these results show a better performance of existing attributions, etc.). Nonetheless, we note that **O4: attribution success is highly inconsistent, with no tested method performing consistently well across settings**, with an average raw success rate of $49.48_{\pm 14.41}$ on diverse models and generally high performance variability in many settings.

Are attributions failing because 2 squares are almost equally important? In most cases, no. To answer this question, we carried out a follow-up analysis in Appendix G.3. For incorrect attributions,

we look at the contribution of the incorrectly predicted square relative to the highest contributing square. We conclude that only in a negligible number of cases an attribution failure was caused by the existence of a second square whose contribution is close to that of the highest contributing square.

Do attributions tend to capture one type of contribution? No, when there is a disagreement about the highest contributor, attributions perform equally poorly against both definitions (Figure F.1). Therefore, **O5: attributions do not seem to have an implicit definition of importance**.

5 Discussion

On sensitivity and importance – a call for rethinking attributions and their evaluation: Decoupling the process of establishing input contributions from attribution evaluation allows us to observe a fairly persistent gap in raw attribution performance between model behaviour regimes for many attributions. This highlights an inherent limitation of many perturbation-based attributions and evaluations.

For example, by definition, the gradient captures how a small change in the pixel values impacts the decision. This is a notion of sensitivity, which is different from measuring the contribution of a subset of pixels towards a decision. Note that a subset of pixels that can contribute significantly to a decision without the model being sensitive to small perturbations of the subset. In other words, sensitive subsets are important, but not all important subsets are sensitive.

In the case of focal models, in line with intuitions from the Gradient Starvation [31] and Shortcut Learning [e.g., 14, 17] literature, we might expect more of the important subsets to also be highly sensitive, especially in models that act more linearly. Therefore, we expect at least some of the improvement in raw attribution performance in the case of focal models to be due to a correlation between sensitivity and importance in those cases. We argue that the results of perturbation-based analyses that do not decouple establishing contributions from attribution evaluation can conflate identifying important subsets with identifying sensitive ones. Therefore our paper calls for more comprehensive evaluations that capture the types of behaviours that our example framework allows us to establish.

Limitations: In its current form, our framework is excessively permissive of attributions. This is because our motivation was not to explicitly show the limitations of current attributions. It was instead a search for what their implicit definition of importance might be. As a result, whenever there was a design decision to make, we chose the one that most favoured attributions. More stringent tests can and should be built and our framework offers a platform for doing so.

Another important limitation is that our framework does not provide a methodology for establishing subset importance on real-world data sets. We argue that it should be seen as a minimum viable test: those attributions that do not perform well on our controlled experiments should not be assumed to perform better on real-world data, where contributions cannot be reliably verified.This is especially the case since, as we argue here, it is easy for evaluations to conflate contributions and sensitivity. Therefore, creating more controlled settings where contributions can be established and behaviours can be characterised is an important step towards creating reliable attribution methods. Importantly, to move from a sensitivity-centric perspective, a rethinking of how we approach attributions and their evaluation as a whole might be required.

6 Conclusions

We decouple the process of establishing subset importance from attribution evaluation and propose two definitions of importance. This allows us to avoid making unverified assumptions about what the model learns and preserve the data distribution from a sample-model perspective. This addresses a number of limitations of typical attribution evaluation. More importantly, the decoupling shows a persistent difference in attribution quality between models that are diverse and models that tend to rely on a single input pattern. This prompts us to question the sensitivity-centric approach to attribution. Overall, our paper calls for the development of more controlled settings for developing and evaluating attributions and provides an example framework for doing so.

Acknowledgments and Disclosure of Funding

We thank the VLC reading group for the constructive discussions and support. AM would also like to especially thank Louis Simon for helping prototype very early versions of the Three Squares Problem. The authors acknowledge the use of the IRIDIS X High Performance Computing Facility, and the Southampton-Wolfson AI Research Machine (SWARM) GPU cluster generously funded by the Wolfson Foundation, together with the associated support services at the University of Southampton in the completion of this work. AC is funded by the EPSRC Doctoral Landscape Award (EP/Z534894/1).

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[2] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International conference on learning representations*, 2022.

[3] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.

[4] Alaa Anani, Tobias Lorenz, Mario Fritz, and Bernt Schiele. Pixel-level Certified Explanations via Randomized Smoothing. In *42nd International Conference on Machine Learning*, 2025. URL https://hdl.handle.net/21.11116/0000-0011-66F2-9.

[5] Anna Arias-Duart, Ettore Mariotti, Dario Garcia-Gasulla, and Jose Maria Alonso-Moral. A confusion matrix for evaluating feature attribution methods. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3709–3714, 2023. doi: 10.1109/CVPRW59228.2023.00380.

[6] Leila Arras, Ahmed Osman, and Wojciech Samek. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, May 2022. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.11.008. URL https://www.sciencedirect.com/science/article/pii/S1566253521002335.

[7] Akanksha Atrey, Kaleigh Clary, and David Jensen. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkl3m1BFDB.

[8] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, July 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130140. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140.

[9] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10329–10338, June 2022.

[10] Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkfMWhAqYQ.

[11] Amnon Catav, Boyang Fu, Yazeed Zoabi, Ahuva Libi Weiss Meilik, Noam Shomron, Jason Ernst, Sriram Sankararaman, and Ran Gilad-Bachrach. Marginal contribution feature importance-an axiomatic approach for explaining data. In *International Conference on Machine Learning*, pages 1324–1335. PMLR, 2021.

[12] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[13] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf.

[14] Giacomo De Palma, Bobak Toussi Kiani, and Seth Lloyd. Deep neural networks are biased towards simple functions. *arXiv*, 2018, 2018.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[16] 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.

[17] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[18] Arne Gevaert, Axel-Jan Rousseau, Thijs Becker, Dirk Valkenborg, Tijl De Bie, and Yvan Saeys. Evaluating feature attribution methods in the image domain. *Machine Learning*, 113(9):6019–6064, 2024.

[19] Jacob Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. FunnyBirds: A Synthetic Vision Dataset for a Part-Based Analysis of Explainable AI Methods, August 2023. URL `http://arxiv.org/abs/2308.06248`. arXiv:2308.06248 [cs].

[22] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Benchmarking the Attribution Quality of Vision Models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2024. URL `https://openreview.net/forum?id=XmyxQaTyck&noteId=1FUnA4F1N3`.

[23] 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.

[24] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4948–4957, 2019.

[25] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5050–5058, 2021.

[26] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. URL `http://arxiv.org/abs/2009.07896`. cite arxiv:2009.07896.

[27] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html`.

[28] Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards Unifying Feature Attribution and Counterfactual Explanations: Different Means to the Same End. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 652–663, July 2021. doi: 10.1145/3461702.3462597. URL `http://arxiv.org/abs/2011.04917`. arXiv:2011.04917 [cs.LG].

[29] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, 55(13s):1–42, December 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3583558. URL `http://arxiv.org/abs/2201.08164`. arXiv:2201.08164 [cs].

[30] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[31] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient Starvation: A Learning Proclivity in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 1256–1272. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

[32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938.

[33] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18770–18795. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/rong22a.html.

[34] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1xWh1rYwB.

[35] 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. *International journal of computer vision*, 128:336–359, 2017.

[36] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[37] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, April 2015. URL http://arxiv.org/abs/1409.1556. arXiv:1409.1556 [cs.CV].

[38] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[39] Suraj Srinivas and Francois Fleuret. Rethinking the role of gradient-based attribution methods for model interpretability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=dYeAHXnpWJ4.

[40] 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, pages 3319–3328, Sydney, NSW, Australia, August 2017. JMLR.org.

[41] Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. Gradient based feature attribution in explainable ai: A technical review. *arXiv preprint arXiv:2403.10415*, 2024.

[42] David S Watson, Limor Gultchin, Ankur Taly, and Luciano Floridi. Local explanations via necessity and sufficiency: Unifying theory and practice. In *Uncertainty in Artificial Intelligence*, pages 1382–1392. PMLR, 2021.

[43] Mengjiao Yang and Been Kim. Benchmarking attribution methods with relative feature importance. *arXiv preprint arXiv:1907.09701*, 2019.

[44] Nehal Yasin, Jonathon Hare, and Antonia Marcu. Is saliency really captured by gradient? In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=p20q5sTjEz.

[45] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (In)fidelity and Sensitivity for Explanations, November 2019. URL http://arxiv.org/abs/1901.09392. arXiv:1901.09392 [cs].

[46] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1. doi: 10.1007/978-3-319-10590-1_53.

[47] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down Neural Attention by Excitation Backprop, August 2016. URL http://arxiv.org/abs/1608.00507. arXiv:1608.00507 [cs].

[48] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 9623–9633, 2022.

[49] Zhiyu Zhu, Zhibo Jin, Jiayu Zhang, Bartlomiej Sobieski, Przemyslaw Biecek, Fang Chen, and Jianlong Zhou. Faithfulness under the distribution: A new look at attribution evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=FF14TqjU3e`.

A Broader context

A large number of studies noted the crucial aspect of creating controlled settings for the purpose of evaluating attributions. However, these often do not account for confounding factors introduced by transformations [e.g., 43, 30, 4, 8]. Others often make data-informed assumptions about what the model has learned without verifying that the assumptions indeed hold [e.g., 48, 6, 46, 47]. This is precisely the reason why we emphasises our observation **O1: despite the background information not being predictive of the label, models can attend to it**. Despite this being a known observation, it is often disregarded in the attribution evaluation literature.

Several attribution evaluations aim to reduce the effect of transformations being potentially out-of-distribution by focusing on removing the most weakly attributed pixels [e.g., 30, 49]. We believe that this approach attempts to measure the accuracy with which **"unimportance"** is indicated by the attribution method: This type of attribution definition is also useful and should be clearly defined and evaluated. Our framework can be extended for this purpose.

Some existing techniques aim to avoid the evaluation transformations being out-of-distribution by optimising (according to a variety of criteria) for a particular pattern to occlude information with [e.g., 49, 33], or by training the networks with data that include the perturbations [e.g., 22, 23, 21]. These examples include studies that use their data-generating process to directly create the transformed versions of the input.

Many studies point out the lack of ground truth for attribution evaluation [e.g., 34, 5, 25]. We avoid this terminology since it is both used in studies that aim to explain the model's decision as well as those that aim to explain the domain [e.g., 11, 39]. We address this by decoupling attribution evaluation from the process of establishing importance.

A.1 Other related work

Necessity and Sufficiency: The principles of necessity and sufficiency underlie much of logical reasoning. The novelty of our work is not in proposing these principles, but in decoupling contributions from evaluating explanations and in proposing an example methodological framework for establishing the types of contributions we define.

In the context of XAI, necessity and sufficiency are typically based on variables with known semantics, whereas contextual and individual importance only register numerical changes without attributing any meaning. Necessity and sufficiency are qualities of an explanation, while importances are about a model's decisions **independently of an explanation**. From [28], "Sufficiency means that the feature subset will always lead to the given model output, irrespective of the values of other features". This is very different from an individual contribution, which is much more fine-grained. Most importantly, necessity and sufficiency typically do not account for O1, which is a well-known observation, yet unaccounted for in this setting.

Consider a case where none of the squares is sufficient to explain the output, nonetheless each of them still has a computable individual contribution for that particular sample in our framework. For example, in our experiments there are many cases in which each of the three squares independently leads to the same classification prediction as seeing all three together. That is, if we define $g_c(x) = \arg\max_c f_c(x)$, we find $g(S_1) = g(S_2) = g(S_3) = g(S_1 \cup S_2 \cup S_3)$. This means that each of the S_i squares is sufficient, but none of them is necessary. Despite each of them being sufficient to obtain the original classification, our framework allows us to calculate the contribution that each makes. Equally, despite none of them being necessary for obtaining the original classification, we can calculate the contribution each makes in a contextual sense.

Therefore, while the principles of necessity and sufficiency have been applied in the context of evaluating the quality of an explanation [e.g. 42], our framework's primary aim is to establish the contributions of input-level patterns to a model's decision. An example usage of our framework is evaluating the quality of an explanation.

Partial Information Decomposition:

Our framework seeks to explain the decision on **each individual sample (not variable-level)**, which is finer-grained than partial inforamtion decomposition. The two objectives, the formulation and entire setup are vastly different.

B Validity checks

Figure B.1 provides a visual flow of our validity checks and their objectives. Our validity checks require a baseline against which "negligible" information can be thresholded. One of our main objectives is to not make unverified assumptions about what the model has learned. We observe, for example, that in the $p_{S_{1..k}} = [0.9, 0.5, 0.9]$ case, all models overwhelmingly rely on a single square, despite both squares being predictive. As such, since in our framework we establish the highest contributing square irrespective of the setting, we choose the contextual contribution of the highest contributing square as a baseline. For VC1, we then verify that every background variation leads to a change smaller than an order of magnitude compared to the baseline.

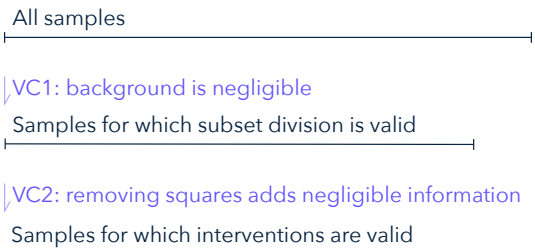


Figure B.1: Flow diagram of our process for selecting those sample-model pairs for which contribution can be reliably established. Given a model we start by considering all test samples. We then apply our validity checks to retain a subset for which our subset division and interventions are valid.

For VC2, we follow the same fundamental rationale. We need to verify that when replacing a given square with a random noise pattern, the random noise itself carries negligible information. That is, if we vary the random noise used to replace the square, the output changes insignificantly with respect to the baseline.

In our paper, we noted a large number of samples that did not pass our validity checks in certain settings, highlighting the importance of these checks Table F.2. For example, our checks eliminate over 70% of the sample-model pairs for the uniform setting ($p_{S_{1..k}} = [1.0, 1.0, 1.0]$) for ViT-B/16. In contrast, in the same setting, no sample-model pairs are eliminated for ResNet-18. Obtaining such different model behaviour is important in ensuring that the selection process does not impact our results, since the conclusions we draw hold for both types of models.

We refer to the cases where there was no agreement on the highest contributor as lacking consensus. We show the number of samples eliminated from our main analysis by a lack of consensus in Table F.2. Note that we investigate these eliminated cases separately (Figure F.1), to see if there is evidence that attributions tend to capture one type of contribution over the other.

B.1 Establishing the most important square (highest contributor)

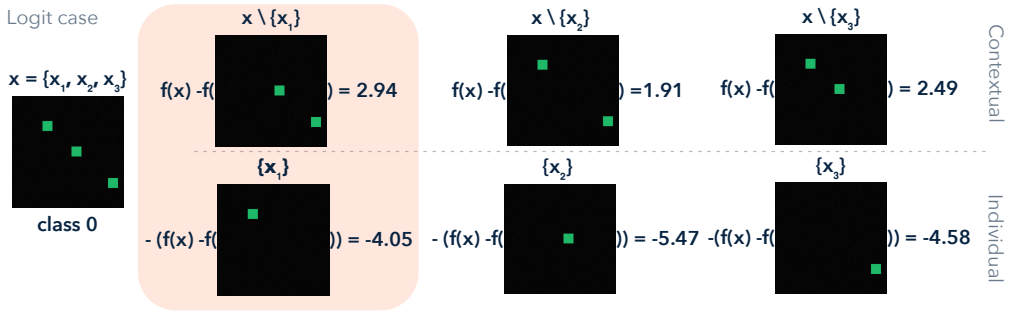


Figure B.2: Illustrative example of our **controlled experiments for determining subset importance** for a model instance f on an input sample x. Removing subset x_1 led to the biggest logit difference (2.94), making it the most important contextual component. Keeping only x_1 led to the smallest logit difference (-4.05), making it the most important individual subset. If x_1 is the most important from a contextual and individual perspective in the softmax case as well, it then becomes the "correct" square to attribute to on sample x. Methods that do not maximally attribute x_1 are considered to have failed on this input instance.

B.2 On subpartitions

Note that the model might still learn to take into account a subset of the square, even though partial squares might be perceived as out of distribution. For the purpose of establishing square importance, this does not impact our analysis. For the purpose of attribution evaluation, we account for this by considering a few different metrics for attribution correctness (the pixel and the square cases, around and within the square).

Note that our partition into subsets is only needed for carrying out the required interventions. In principle, two subsets can have identical contributions, effectively acting as a single subset. Therefore the fact that model features can span multiple squares is already accounted for in our framework. Precisely, a model can have features that treat squares individually, contextually, or a combination thereof. Since we are interested in the overall contribution for the model's prediction rather than a feature-based contribution, our framework accounts for these cases by design.

C Additional framework design justification

To obtain models with different behaviours and learned rules, we created three different possible areas for squares to be positioned in. This adds variability to the training samples but it also mimics a more realistic data set where input patterns are not typically expected to occur at the exact same locations for all samples. To avoid models memorising and relying on exact pixel values, we added a small amount of noise to our samples. This also makes model-sample pairs more likely to pass the VC tests.

C.1 Extension to other tasks

We believe it is possible to extend the existing framework to other types of tasks, such as regression problems. A careful consideration of task-specific requirements is necessary, paying particular attention to creating an equivalent of a "negligible information" pattern that can be resampled for the VC checks.

D Other experimental details

All models are trained using SGD with a momentum of 0.9, weight decay of 10^{-4} and learning rate of 0.001. We train the Coloured Squares models for 10 epochs and the ResNet-18 and ResNet-50 ImageNet models for 30 and 50 epochs respectively.

D.1 Attribution computation

For computing attributions we use the Captum [26] and PyGradcam [19] libraries. A full list of attributions considered (mapping to the names in the figures throughout the paper) can be found in Table D.1.

Table D.1: Attribution methods considered in the paper.

Attribution Method	Description
InputXGradient [36]	The product between the input and the gradient
InputXGradientAbs	Absolute value of the product between the input and the gradient
Integ0.5 [40]	Integrated Gradients with a 0.5 baseline
Integ0.5Abs	Absolute value of Integrated Gradients with a 0.5 baseline
IntegBlack	Integrated Gradients with a black baseline
IntegBlackAbs	Absolute value of Integrated Gradients with a black baseline
IntegWhite	Integrated Gradients with a white baseline
IntegWhiteAbs	Absolute value of Integrated Gradients with a white baseline
Lime44Black [32]	Surrogate method that fits a linear model to predict the model's output from perturbed regions of the input, with a black baseline and a 4x4 pixel grid
Lime44Noise	Surrogate method that fits a linear model to predict the model's output from perturbed regions of the input, with a 0.05 noise baseline and a 4x4 pixel grid
PyGradcam	Grad-CAM
RISE [30]	A perturbation based approach to compute attribution, involving a monte-carlo approach to detecting the sensitivity of the output with respect to features (1000 masks of 4x4)
Saliency [38]	Raw gradient values
SaliencyAbs	Absolute gradients
SHAP [27]	Uses Shapley values to explain machine learning models

D.2 Attribution evaluation

- the **pixel** case: If the maximally attributed pixel is within the area of the square, we consider it an attribution success. We compute the fraction of times an attribution is successful against the total number of times the maximum attribution falls within any of the three possible squares (rather than all valid sample-model pairs).
- the **square** case: For each of the three squares, we compute the average attributed value within the square. If the correct square was assigned the highest average attribution, we consider it an attribution success.

For both the pixel and the square case, we then considered two versions: a "within" version which comprises only the pixels that make up the square and an "around" version, which considers a 10-pixel buffer around the square. **Since the latter consistently provides better results for attributions, we set this version as our default.** Figure D.1 provides an illustrative example of attribution evaluation.

D.3 Computational Resources

Note that due to computational considerations, some of the figures in the appendices contain results only for the attribution methods that we evaluated before rebuttal. Some methods, such as SHAP are prohibitively expensive considering the large number of models we evaluate.

Models were trained, evaluated and attributions computed using a range of Nvidia GPU accelerators including L4s, L40s, A100s, H100s and H200s. For the main experiments on The Squares Problem, we trained each of the 4 model architectures for 10 different seeds (which were subsequently filtered to 5 for the analysis). Training individual models is quick; typically taking of the order of minutes. The most computationally expensive part of our experiments was computing the attributions — particularly the integrated gradient methods which require multiple passes, and this took up to several hours depending on the GPU used.

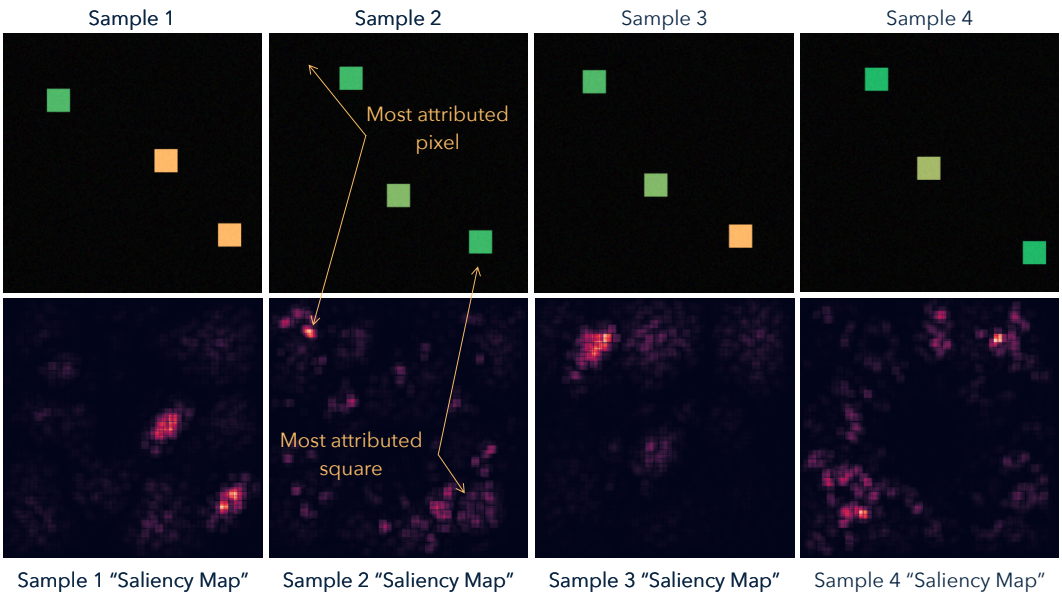


Figure D.1: Data samples and associated attribution map computed using Captum's "Saliency" method (raw gradients). Sample 2 illustrates an example where the **pixel**-based evaluation (highest attributed pixel) is outside the possible area for any of the squares. The square-based evaluation selects the square assigned the highest average attribution. Once an attribution assignment is established, we compare this against the highest contributing square according to our framework.

E Additional Results: Squares Setting

E.1 Logits versus Softmax

Since we focus on classification tasks, another factor to consider when defining importance is the "type" of output with respect to which we define the contribution. In our formulation, the model f can be considered to output either softmax probabilities or logits. All the models used in this paper are implemented to output logits and a decision is simply taken to be the class with the maximum logit. However, during training, the use of softmax is implied through the cross-entropy loss. The prediction remains the same whether using pure logits or softmax probabilities. But importantly, when carrying out interventions, the subsets' contextual and individual contributions could change, as we observe in our experiments.

Since the original task is discriminative, the logit of a given class will naturally capture aspects of what differentiates that given class from the rest. A softmax perspective, on the other hand, will further highlight components that relate to other classes. For example, when removing a component, the logit of the predicted class might change very little, while those of other classes change significantly. One potential case in which this could happen is when the absence of a feature that is associated with other classes is seen as evidence for the class in question.

E.2 Ranking correlation results

While alternative measures of attribution success such as ranking correlation can be considered, the exact choice of success measure is not our primary concern since the purpose of this paper is not to provide an extensive evaluation of attributions, but to illustrate how our framework can be used. In particlar, we demonstrate one of the possible uses, which is to provide a testbed for attribution evaluation.

As an alternative to our highest-contributor notions of attribution success, we consider ranking correlation. For computational reasons, the results in this section are for the main configurations in the paper (4 different $p_{S_{1..k}}$ values: [0.9, 0.5, 0.9], [1.0, 1.0, 1.0], [0.7, 0.5, 0.9] for [5.0, 5.0, 5.0] colour standard deviations).

In the context of rebuttal discussions around ensuring that the VC filters do not introduce a selection bias, we reported the rank correlation when no filtering was applied. We subsequently computed the same quantity when applying the VC checks and we report both of these in Table E.1. For a visual depiction of the latter, see Figure E.1.

When considering rank correlation, we once again note some differences between attribution performance on focal versus diverse models, with more consistent results (lower standard deviation) on focal models.

Table E.1: Rank correlation results for ViT-B/16 on the main data settings.

	Unfiltered		Filtered with VCs	
Attribution	focal	diverse	focal	diverse
InputXGradient	$0.04_{\pm 0.16}$	$0.13_{\pm 0.13}$	$0.07_{\pm 0.18}$	$0.22_{\pm 0.14}$
IntegWhite	$0.53_{\pm 0.06}$	$0.27_{\pm 0.25}$	$0.58_{\pm 0.06}$	$0.26_{\pm 0.28}$
Saliency	$-0.20_{\pm 0.11}$	$-0.03_{\pm 0.17}$	$-0.22_{\pm 0.11}$	$0.02_{\pm 0.27}$
IntegBlack	$0.53_{\pm 0.03}$	$0.37_{\pm 0.15}$	$0.57_{\pm 0.03}$	$0.45_{\pm 0.13}$
SaliencyAbs	$0.50_{\pm 0.04}$	$0.28_{\pm 0.13}$	$0.50_{\pm 0.04}$	$0.37_{\pm 0.16}$
InputXGradientAbs	$0.30_{\pm 0.01}$	$0.20_{\pm 0.08}$	$0.31_{\pm 0.02}$	$0.26_{\pm 0.09}$
IntegWhiteAbs	$0.55_{\pm 0.05}$	$0.24_{\pm 0.26}$	$0.56_{\pm 0.06}$	$0.25_{\pm 0.27}$
IntegBlackAbs	$0.39_{\pm 0.01}$	$0.27_{\pm 0.10}$	$0.41_{\pm 0.01}$	$0.34_{\pm 0.11}$
Integ0.5	$0.44_{\pm 0.11}$	$0.23_{\pm 0.22}$	$0.48_{\pm 0.11}$	$0.22_{\pm 0.26}$
Integ0.5Abs	$0.54_{\pm 0.05}$	$0.25_{\pm 0.25}$	$0.56_{\pm 0.05}$	$0.27_{\pm 0.24}$

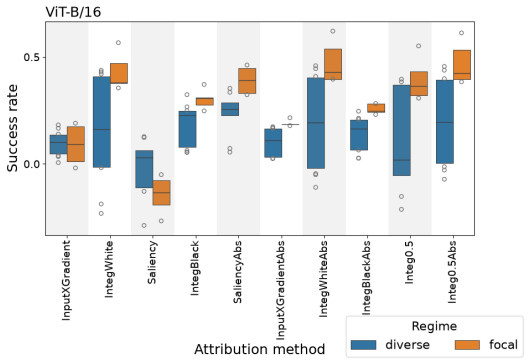


Figure E.1: Rank correlation results for ViT-B/16 on the main data settings. Results after VC checks are applied.

While more extensive results can be collected and analysed, we leave this investigation as future work since it is not the primary concern of this paper.

E.3 Overall change in prediction

As an alternative to our definition of importance based on the change in logit of the predicted class, we considered an alternative based on the overall logit change (i.e. across all classes).

We included a variant where we compute the success rate based on the overall change in prediction across classes, rather than the change in the predicted class only. We computed this for ViT-16, ResNet-18 and VGG-16. On each architecture, we noticed a drop in the average attribution performance of at least 20pp.

E.4 Inverted background

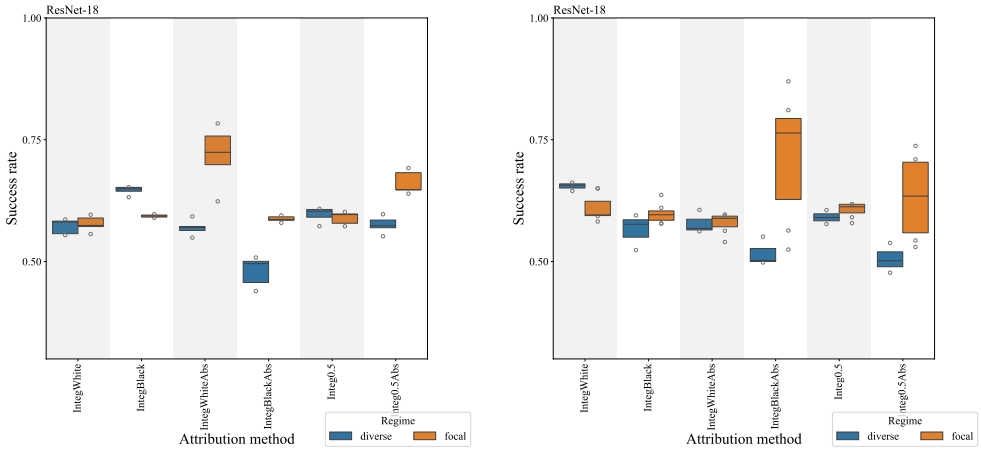
During the rebuttal, the question of whether the choice of black background favours IntegratedBlack over IntegratedWhite variants occurred.

It is first worth noting the existing results (black background). Specifically, note that in Figure 3, the other Integ variants (either white or gray) sometimes already perform on par with IntegBlack,

especially for focal models (see IntegWhite for VGG, ResNet-18 for example). In the black background setting, among the Integrated Gradients variants, IntegWhite has a slightly higher mean for focal models.That is because the difference between the default background and a white baseline is the greatest (leading to higher gradients). Therefore it is IntegWhite that has a slight advantage, but this is minimal. When we ran the white-background variant, we observed the same trend, where IntegBlack was performing slightly better in terms of the mean performance (since now the black baseline had the highest difference). Changing the background to white just leads to an effective swap, with IntegBlack now having a higher average performance on focal models. This swap has no impact on our conclusions.

Note that for simplicity, we implemented a swap as $1-$ the previously generated images. In effect, that means that the colours corresponding to each class are going to be different in this setting. This simplification however suffices for the purpose of investigating the choice of the background in the ranking of Integ variants.

The dependency on a choice of baseline is a known issue with IntegratedGradients and usually requires considering a few different baseline variants (which we have). In a real-world setting, for example ImageNet, the results could be different based on the choice of baseline with no universally acknowledged procedure for choosing a baseline. Therefore, no matter how we would set up the data set, the question of a choice of baseline always remains since, this is not a characteristic of our framework but of Integrated Gradients depending on a baseline.



(a) Black background (b) White background

Figure E.2: Success rate for a subset of ResNet-18 models on images with black background (a) versus images with a white background (b). Note that the white background was simply implemented as $1-$ the original, black-background, images. Therefore, the correspondence between colours and classes is not exact. Nonetheless, these partial results on an approximate white-background version suffice to indicate that the choice of background colour does not impact our broader observations.

F Conflicting contribution

In this section we analyse the cases where there is no consensus between the importance definitions as to which square is the highest contributor. We therefore evaluate attributions separately against each type of contribution. If one attribution implicitly captures one type of contribution, we should see that the performance measured against that type of contribution is consistently better. This would be visible as a gap in the average performance between the two types of contribution. However, we see in Figure F.1 that the means and Interquartile Range overlap; we performed a t-test (null hypothesis: equal population means) for each architecture, and each attribution method. The lowest p-value was 0.13 (and the highest 0.98), with an average p-value of 0.55. Therefore, we find no statistically significant evidence that attributions consistently capture either type of contribution. As such, it is difficult to interpret the results of attribution methods since it is unclear what they intend to capture.

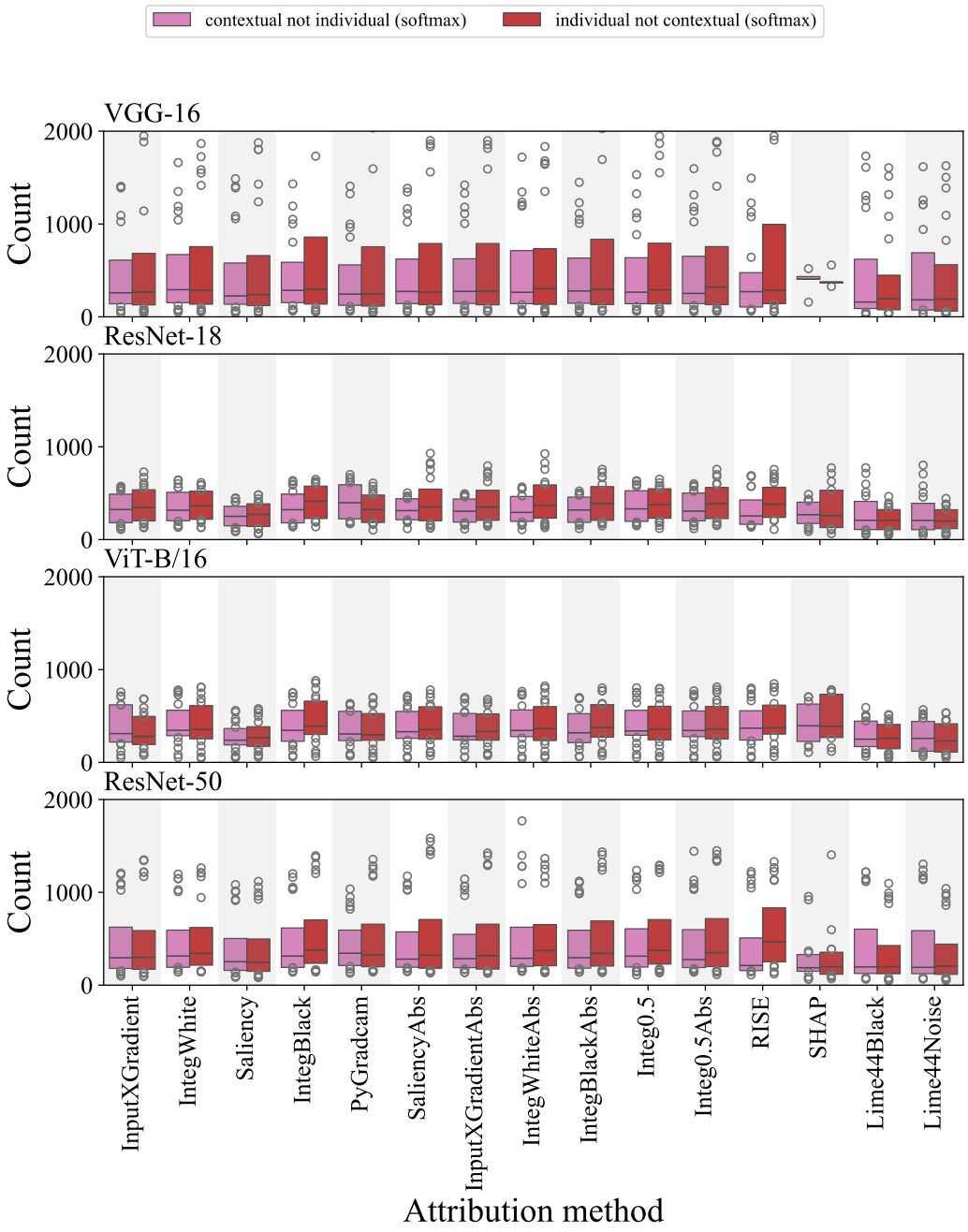


Figure F.1: Attribution results in the cases where the highest contributor in an individual sense is different from the highest contributor in a contextual sense. We count the number of times attributions are correct when evaluated against the highest contributor in each sense. We found no statistically significant difference in attribution quality when evaluated against each of these complementary definitions.

F.1 Model Statistics

Table F.1: Training and validation accuracies for models trained on The Squares Problem. Results reported are as means and standard deviations over 10 different initial seeds of the model weights.

Architecture	$p_{S_{1..k}}$	Training accuracy	Validation accuracy
ResNet-18	[0.7, 0.1, 0.9]	$94.47_{\pm 0.14}$	$96.84_{\pm 0.08}$
ResNet-18	[0.7, 0.5, 0.9]	$96.33_{\pm 0.07}$	$98.03_{\pm 0.17}$
ResNet-18	[0.9, 0.5, 0.9]	$98.00_{\pm 0.07}$	$99.23_{\pm 0.05}$
ResNet-18	[1.0, 1.0, 1.0]	$99.91_{\pm 0.01}$	$100.00_{\pm 0.00}$
ResNet-50	[0.7, 0.1, 0.9]	$94.25_{\pm 0.12}$	$96.84_{\pm 0.17}$
ResNet-50	[0.7, 0.5, 0.9]	$96.25_{\pm 0.10}$	$98.12_{\pm 0.11}$
ResNet-50	[0.9, 0.5, 0.9]	$97.99_{\pm 0.07}$	$99.23_{\pm 0.04}$
ResNet-50	[1.0, 1.0, 1.0]	$99.88_{\pm 0.05}$	$100.00_{\pm 0.00}$
VGG-16	[0.7, 0.1, 0.9]	$93.69_{\pm 0.11}$	$96.68_{\pm 0.14}$
VGG-16	[0.7, 0.5, 0.9]	$95.79_{\pm 0.12}$	$97.91_{\pm 0.18}$
VGG-16	[0.9, 0.5, 0.9]	$97.57_{\pm 0.25}$	$98.94_{\pm 0.14}$
VGG-16	[1.0, 1.0, 1.0]	$99.67_{\pm 0.67}$	$100.00_{\pm 0.00}$
ViT-B/16	[0.7, 0.1, 0.9]	$93.93_{\pm 0.20}$	$96.13_{\pm 0.26}$
ViT-B/16	[0.7, 0.5, 0.9]	$96.07_{\pm 0.15}$	$97.23_{\pm 0.29}$
ViT-B/16	[0.9, 0.5, 0.9]	$97.89_{\pm 0.11}$	$98.57_{\pm 0.11}$
ViT-B/16	[1.0, 1.0, 1.0]	$99.91_{\pm 0.01}$	$100.00_{\pm 0.00}$

Table F.2: The number of samples eliminated by our validity checks and lack of consensus per model and setting. Note the variability in the number of sample-model instances filtered out by the validity checks. A notable example is ViT-B/16, where the checks can eliminate over 70% of the evaluation data.

Architecture	$p_{S_{1..k}}$	Number eliminated by validity checks	Number eliminated by lack of consensus
ResNet-18	[0.7,0.1,0.9]	$581_{\pm 258}$	$415_{\pm 55}$
ResNet-18	[0.7,0.5,0.9]	$239_{\pm 126}$	$646_{\pm 47}$
ResNet-18	[0.9,0.5,0.9]	$132_{\pm 76}$	$1298_{\pm 95}$
ResNet-18	[1.0,1.0,1.0]	$0_{\pm 0}$	$1741_{\pm 231}$
ResNet-50	[0.7,0.1,0.9]	$98_{\pm 37}$	$417_{\pm 27}$
ResNet-50	[0.7,0.5,0.9]	$52_{\pm 14}$	$643_{\pm 37}$
ResNet-50	[0.9,0.5,0.9]	$18_{\pm 8}$	$1248_{\pm 75}$
ResNet-50	[1.0,1.0,1.0]	$0_{\pm 1}$	$3609_{\pm 252}$
VGG-16	[0.7,0.1,0.9]	$1_{\pm 1}$	$242_{\pm 71}$
VGG-16	[0.7,0.5,0.9]	$1_{\pm 1}$	$445_{\pm 86}$
VGG-16	[0.9,0.5,0.9]	$1_{\pm 1}$	$1268_{\pm 159}$
VGG-16	[1.0,1.0,1.0]	$24_{\pm 66}$	$5114_{\pm 556}$
ViT-B/16	[0.7,0.1,0.9]	$1574_{\pm 696}$	$720_{\pm 225}$
ViT-B/16	[0.7,0.5,0.9]	$1570_{\pm 784}$	$1152_{\pm 190}$
ViT-B/16	[0.9,0.5,0.9]	$1308_{\pm 777}$	$2094_{\pm 208}$
ViT-B/16	[1.0,1.0,1.0]	$7782_{\pm 1345}$	$889_{\pm 857}$

Table F.3: Percentage of samples where different notions of importance identify the same square as the highest contributor. We refer to the case where all definition agree on the square that has the highest contribution as cases of "consensus".

Architecture	$p_{S_{1..k}}$	Importance agreement (softmax)	Importance agreement (logits)	Highest contribution consensus
ResNet-18	[0.7,0.1,0.9]	$96.37_{\pm 0.47}$	$96.51_{\pm 0.51}$	$95.59_{\pm 0.66}$
ResNet-18	[0.7,0.5,0.9]	$94.38_{\pm 0.45}$	$94.81_{\pm 0.46}$	$93.38_{\pm 0.53}$
ResNet-18	[0.9,0.5,0.9]	$88.93_{\pm 0.75}$	$89.34_{\pm 0.83}$	$86.85_{\pm 0.99}$
ResNet-18	[1.0,1.0,1.0]	$87.08_{\pm 1.38}$	$85.72_{\pm 1.73}$	$82.59_{\pm 2.31}$
ResNet-50	[0.7,0.1,0.9]	$96.80_{\pm 0.32}$	$97.08_{\pm 0.28}$	$95.79_{\pm 0.27}$
ResNet-50	[0.7,0.5,0.9]	$95.14_{\pm 0.46}$	$95.41_{\pm 0.31}$	$93.53_{\pm 0.38}$
ResNet-50	[0.9,0.5,0.9]	$90.45_{\pm 0.57}$	$90.75_{\pm 0.46}$	$87.49_{\pm 0.75}$
ResNet-50	[1.0,1.0,1.0]	$74.41_{\pm 2.41}$	$79.06_{\pm 2.55}$	$63.91_{\pm 2.52}$
VGG-16	[0.7,0.1,0.9]	$98.42_{\pm 0.47}$	$98.63_{\pm 0.44}$	$97.58_{\pm 0.71}$
VGG-16	[0.7,0.5,0.9]	$96.53_{\pm 0.66}$	$97.02_{\pm 0.54}$	$95.55_{\pm 0.86}$
VGG-16	[0.9,0.5,0.9]	$90.69_{\pm 1.66}$	$91.93_{\pm 1.43}$	$87.32_{\pm 1.59}$
VGG-16	[1.0,1.0,1.0]	$76.22_{\pm 5.33}$	$72.80_{\pm 4.01}$	$48.74_{\pm 5.51}$
ViT-B/16	[0.7,0.1,0.9]	$93.46_{\pm 2.05}$	$92.91_{\pm 2.33}$	$91.48_{\pm 2.58}$
ViT-B/16	[0.7,0.5,0.9]	$88.95_{\pm 2.08}$	$88.60_{\pm 3.06}$	$86.17_{\pm 3.13}$
ViT-B/16	[0.9,0.5,0.9]	$80.42_{\pm 2.38}$	$80.30_{\pm 2.62}$	$75.69_{\pm 3.62}$
ViT-B/16	[1.0,1.0,1.0]	$79.26_{\pm 6.77}$	$88.38_{\pm 10.44}$	$66.31_{\pm 13.55}$

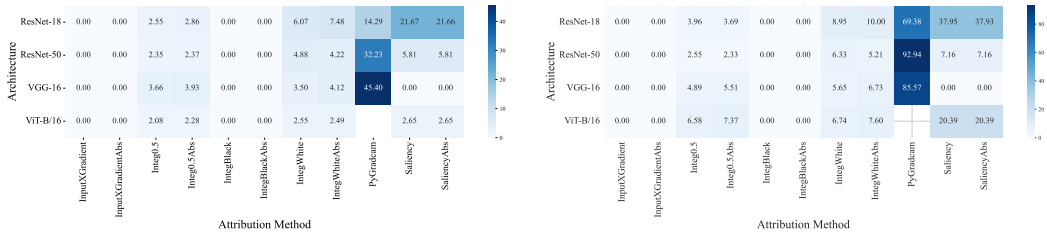
Table F.4: The percentage of samples for which a given square was the highest contributor.

Architecture	$p_{S_{1..k}}$	samples highest contributor		
		1	2	3
ResNet-18	[0.7,0.1,0.9]	$14.42_{\pm 1.75}$	$0.34_{\pm 0.08}$	$85.25_{\pm 1.73}$
ResNet-18	[0.7,0.5,0.9]	$13.85_{\pm 1.32}$	$3.94_{\pm 0.32}$	$82.22_{\pm 1.44}$
ResNet-18	[0.9,0.5,0.9]	$56.73_{\pm 3.72}$	$1.33_{\pm 0.26}$	$41.94_{\pm 3.62}$
ResNet-18	[1.0,1.0,1.0]	$3.08_{\pm 1.01}$	$95.91_{\pm 1.30}$	$1.01_{\pm 0.42}$
ResNet-50	[0.7,0.1,0.9]	$14.85_{\pm 1.25}$	$0.77_{\pm 0.12}$	$84.38_{\pm 1.33}$
ResNet-50	[0.7,0.5,0.9]	$14.11_{\pm 0.64}$	$4.98_{\pm 0.52}$	$80.91_{\pm 0.96}$
ResNet-50	[0.9,0.5,0.9]	$55.01_{\pm 2.40}$	$2.04_{\pm 0.36}$	$42.95_{\pm 2.44}$
ResNet-50	[1.0,1.0,1.0]	$24.86_{\pm 1.96}$	$73.50_{\pm 2.29}$	$1.64_{\pm 0.50}$
VGG-16	[0.7,0.1,0.9]	$13.51_{\pm 1.98}$	$0.80_{\pm 0.11}$	$85.69_{\pm 2.05}$
VGG-16	[0.7,0.5,0.9]	$12.27_{\pm 1.37}$	$2.98_{\pm 0.38}$	$84.76_{\pm 1.49}$
VGG-16	[0.9,0.5,0.9]	$49.71_{\pm 6.55}$	$0.95_{\pm 0.34}$	$49.34_{\pm 6.68}$
VGG-16	[1.0,1.0,1.0]	$50.34_{\pm 10.38}$	$34.52_{\pm 8.18}$	$15.14_{\pm 6.91}$
ViT-B/16	[0.7,0.1,0.9]	$14.68_{\pm 2.18}$	$0.54_{\pm 0.14}$	$84.78_{\pm 2.24}$
ViT-B/16	[0.7,0.5,0.9]	$12.25_{\pm 1.04}$	$3.92_{\pm 0.50}$	$83.83_{\pm 1.41}$
ViT-B/16	[0.9,0.5,0.9]	$48.44_{\pm 2.46}$	$1.91_{\pm 0.29}$	$49.65_{\pm 2.39}$
ViT-B/16	[1.0,1.0,1.0]	$37.34_{\pm 3.09}$	$33.74_{\pm 1.12}$	$28.92_{\pm 2.76}$

F.1.1 Increasing the number of training samples

In Section 3.2 we note that over 75% of the models have samples on which the background contributes non-negligibly. We set out to verify if adding more samples to the training set would impact this. We chose ViT-B/16, as this was the model with the highest VC elimination rate (as per Table F.2). We increased the number of training samples from 60K to 80K. The number of eliminated samples in the two cases are within each other's standard deviation (computed across 5 seeds). In this particular instance, adding more samples did not reduce the number of VC eliminations.

G Additional Attribution Results



(a) (b)

Figure G.1: Percentage of times that the most attributed pixel is outside of: a. any of the possible areas and b. any of the components (using the relaxed 'around' definition; i.e., considering a 10px buffer around the square).

G.1 Alternative Cases for Attribution Success

In the main paper, we consider an attribution to be successful by verifying the attribution against a generously large area. Here we present alternative results, obtained for more restrictive settings. Figure G.2 provides results for the pixel and square case, in the "around area", as well as the "within area" setting. That is, compared to our main results, which consider a 10-pixel buffer around the square, the "within area" only considers an attribution to be correct if it highly attributes the square itself.

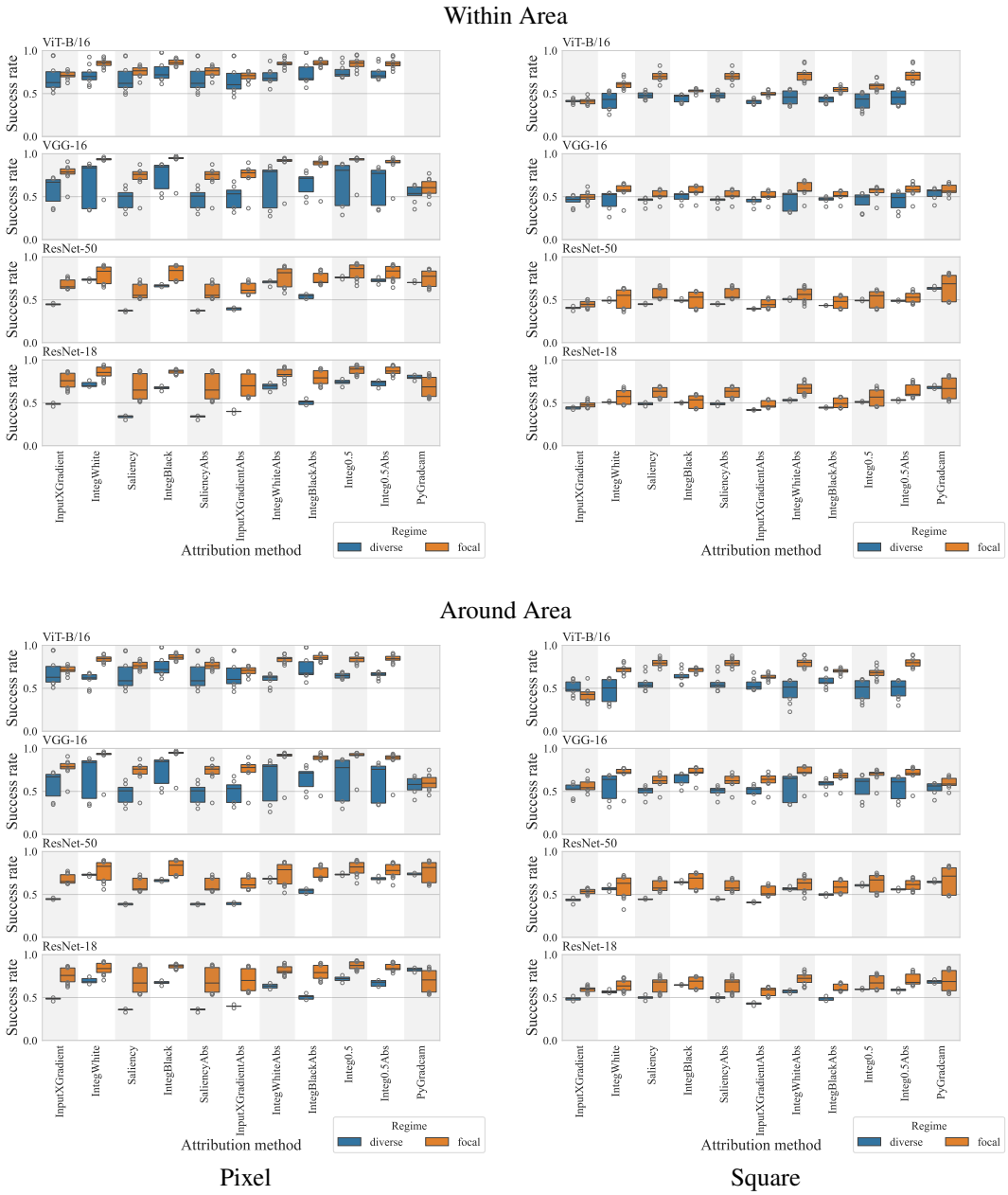


Figure G.2: Per-regime (focal versus diverse) results when evaluating attributions according to different criteria (pixel and square; around area and within area). Results are aggregated over all $p_{S_{1..k}}$ settings.

G.2 Results for varying the colour deviation

Table G.1: Training and validation accuracies for ResNet-18 models trained on The Squares Problem with differing Colour variance. Results reported are as means and standard deviations over 10 different initial seeds of the model weights.

Architecture	$p_{S_{1..k}}$	Colour variance	Training accuracy	Validation accuracy
ResNet-18	[0.9, 0.5, 0.9]	[1.0, 5.0, 1.0]	$98.14_{\pm 0.06}$	$99.34_{\pm 0.08}$
ResNet-18	[0.9, 0.5, 0.9]	[5.0, 1.0, 5.0]	$98.02_{\pm 0.07}$	$99.24_{\pm 0.08}$
ResNet-18	[0.9, 0.5, 0.9]	[5.0, 5.0, 5.0]	$98.00_{\pm 0.07}$	$99.23_{\pm 0.05}$

Table G.2: Percentage of samples where different notions of importance identify the same square as the highest contributor **when changing the colour variance**. We refer to the case where all definition agree on the square that has the highest contribution as cases of "consensus".

Architecture	$p_{S_{1..k}}$	Colour variance	Importance agreement (softmax)	Importance agreement (logits)	Highest contribution consensus
ResNet-18	[0.9,0.5,0.9]	[1.0,5.0,1.0]	$86.42_{\pm 1.01}$	$86.63_{\pm 1.12}$	$83.79_{\pm 1.15}$
ResNet-18	[0.9,0.5,0.9]	[5.0,1.0,5.0]	$89.49_{\pm 0.82}$	$89.72_{\pm 0.75}$	$87.28_{\pm 0.94}$
ResNet-18	[0.9,0.5,0.9]	[5.0,5.0,5.0]	$88.93_{\pm 0.75}$	$89.34_{\pm 0.83}$	$86.85_{\pm 0.99}$

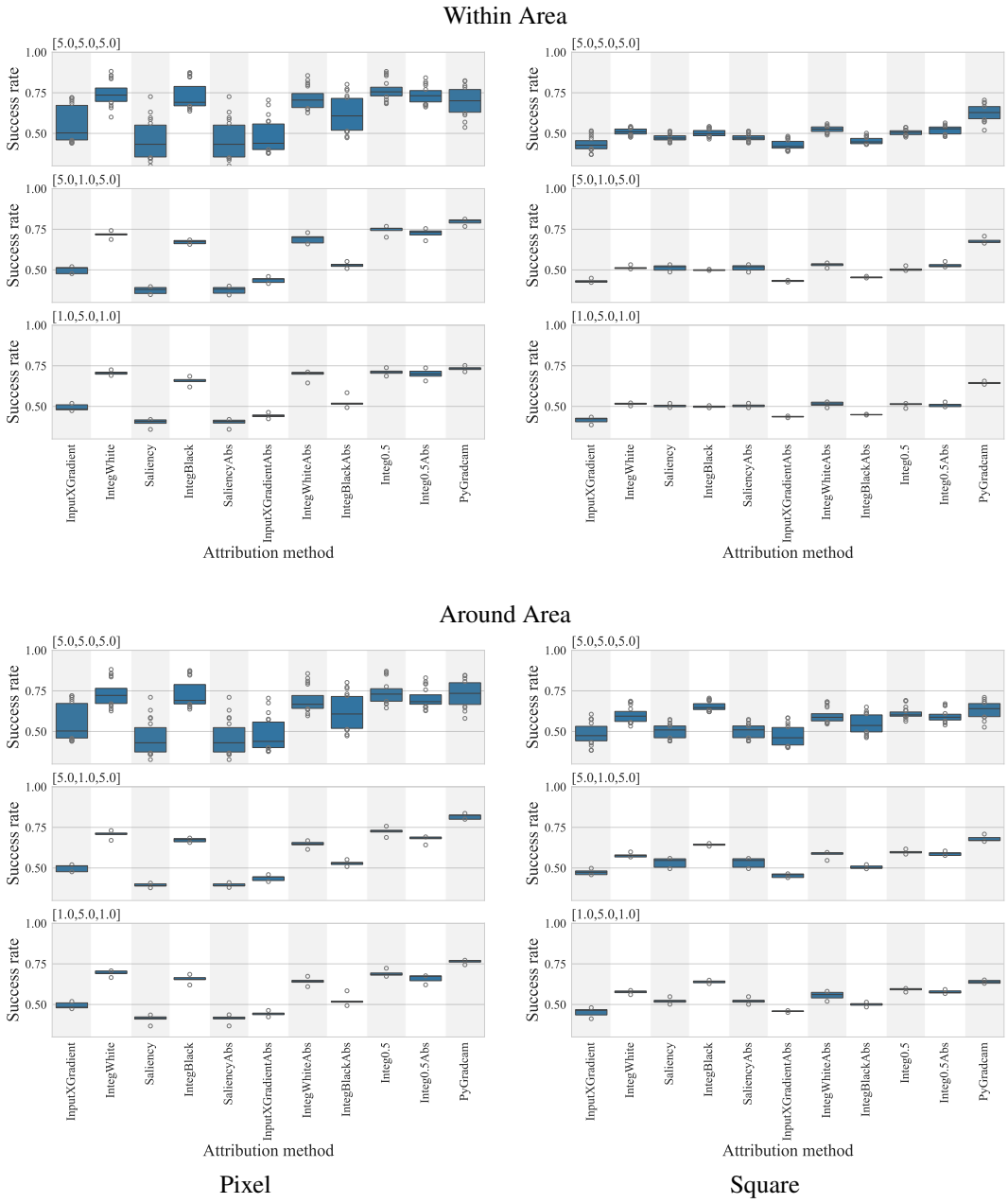


Figure G.3: Per-regime (focal versus diverse) results when evaluating attributions according to different criteria (pixel and square; around area and within area) **when changing the colour variance** for $p_{S_{1..k}}$ = [0.9,0.5,0.9].

G.3 Relative contributions

For each sample, we take the contextual and individual contribution of each of the 3 components in our data set and we normalise them between 0 and 1. That is, after normalisation, each sample will have a square with a 0 contribution, another square with a contribution of 1 (the correct square), and the last square will have a contribution that is relative to the two. Then, in the case of incorrect attributions, we compute the average relative contribution of the attributed square. Relative contributions with values close to 1 signify that the attributed and the correct square had highly similar contributions. We find that the average relative contextual are well below 0.5, and therefore closer to the least important component rather than the most important one. In fact, the fraction of cases where the relative contribution is above 0.9 is, on average, less than 2% for contextual. In the individual sense, we observe a higher number of cases of close attribution (an average around 20%). This could simply be because of the inherent distribution of individual compared to contextual contributions, or it might be an indication that attributions tend to implicitly capture the individual rather than the contextual contributions. However, in Table G.3 we explicitly investigate this and find that attributions do not, in fact, tend to capture one type of contribution over the other.

Table G.3: Investigation of incorrect predictions. We consider a "close" incorrect attribution to be one where the contrbution of the attributed square is higher than 90% of that of the highest contributor. We compute the number of close attributions for the individual and contextual cases, as well as the average relative contribution of the incorrectly predicted square.

Attribution Method	#close attributions individual	#close attributions contextual	average relative contextual contribution	average relative individual contribution
InputXGradient	$16.98_{\pm 5.94}$	$1.25_{\pm 0.76}$	$0.18_{\pm 0.07}$	$0.50_{\pm 0.10}$
InputXGradientAbs	$19.16_{\pm 8.13}$	$1.28_{\pm 0.76}$	$0.20_{\pm 0.05}$	$0.58_{\pm 0.12}$
Integ0.5	$17.19_{\pm 6.03}$	$1.32_{\pm 0.78}$	$0.23_{\pm 0.07}$	$0.65_{\pm 0.13}$
Integ0.5Abs	$16.68_{\pm 6.48}$	$1.34_{\pm 0.79}$	$0.25_{\pm 0.06}$	$0.65_{\pm 0.11}$
IntegBlack	$16.71_{\pm 5.48}$	$1.35_{\pm 0.79}$	$0.25_{\pm 0.06}$	$0.70_{\pm 0.11}$
IntegBlackAbs	$17.85_{\pm 7.16}$	$1.33_{\pm 0.77}$	$0.22_{\pm 0.05}$	$0.61_{\pm 0.10}$
IntegWhite	$17.75_{\pm 6.41}$	$1.32_{\pm 0.76}$	$0.24_{\pm 0.06}$	$0.66_{\pm 0.13}$
IntegWhiteAbs	$16.74_{\pm 6.76}$	$1.34_{\pm 0.79}$	$0.25_{\pm 0.06}$	$0.66_{\pm 0.12}$
Lime44Black	$12.72_{\pm 4.53}$	$0.87_{\pm 0.67}$	$0.09_{\pm 0.05}$	$0.29_{\pm 0.10}$
Lime44Noise	$12.72_{\pm 4.74}$	$0.88_{\pm 0.68}$	$0.09_{\pm 0.05}$	$0.29_{\pm 0.10}$
PyGradcam	$18.54_{\pm 6.28}$	$1.29_{\pm 0.77}$	$0.23_{\pm 0.10}$	$0.65_{\pm 0.19}$
RISE	$18.21_{\pm 5.69}$	$1.32_{\pm 0.76}$	$0.30_{\pm 0.07}$	$0.78_{\pm 0.14}$
SHAP	$20.34_{\pm 6.05}$	$1.19_{\pm 0.70}$	$0.15_{\pm 0.08}$	$0.43_{\pm 0.12}$
Saliency	$15.17_{\pm 5.89}$	$1.00_{\pm 0.65}$	$0.12_{\pm 0.05}$	$0.36_{\pm 0.10}$
SaliencyAbs	$18.96_{\pm 8.63}$	$1.32_{\pm 0.79}$	$0.23_{\pm 0.05}$	$0.64_{\pm 0.13}$

G.4 More choice: 4-squares

We trained a ResNet-18, ResNet-50, VGG-16, and ViT-B/16 on a 4-squares version of the dataset, using the same size (20px) of squares as in 3-squares. Over all four models and all attributions[3], we observe an attribution success rate of just $14.1\%_{\pm 6.1}$ (see Table G.4 for per model results) [4].

[3]For this experiment, only these were evaluated: InputXGradient, InputXGradientAbs, Integ0.5, Integ0.5Abs, IntegBlack,IntegBlackAbs, IntegWhite, IntegWhiteAbs, PyGradcam, Saliency, SaliencyAbs.

[4]Success is higher at $22.7\%_{\pm 16.6}$ when importance is measured across classes. In the 3-squares experiment, the equivalent measure was yielding worse results than predicted-class importance. This further supports our inconsistency claims.

Table G.4: Mean success rates for each Architecture on four-squares data

Architecture	Success rate
ResNet-18	$14.9_{\pm 8.5}$
ResNet-50	$12.8_{\pm 4.9}$
VGG-16	$12.2_{\pm 4.6}$
ViT-B/16	$16.3_{\pm 4.3}$

Table G.5: Training and validation accuracies for models trained on Four Squares (Only one correlations setting). Results reported are as means and standard deviations over 5 different initial seeds of the model weights.

Architecture	$p_{S_{1..k}}$	Training accuracy	Validation accuracy
ResNet-18	[0.9, 0.5, 0.9, 0.9]	$99.48_{\pm 0.05}$	$99.81_{\pm 0.03}$
ResNet-50	[0.9, 0.5, 0.9, 0.9]	$99.45_{\pm 0.05}$	$99.84_{\pm 0.04}$
VGG-16	[0.9, 0.5, 0.9, 0.9]	$98.68_{\pm 0.66}$	$99.46_{\pm 0.14}$
ViT-B/16	[0.9, 0.5, 0.9, 0.9]	$99.45_{\pm 0.09}$	$99.28_{\pm 0.27}$

Table G.6: The number of samples eliminated by our validity checks and lack of consensus per model on the Four Squares problem. Only one setting evaluated. Note the variability in the number of sample-model instances filtered out by the validity checks.

Architecture	$p_{S_{1..k}}$	Number eliminated by validity checks	Number eliminated by lack of consensus
ResNet-18	[0.9,0.5,0.9,0.9]	$157_{\pm 137}$	$4205_{\pm 480}$
ResNet-50	[0.9,0.5,0.9,0.9]	$5_{\pm 2}$	$3149_{\pm 224}$
VGG-16	[0.9,0.5,0.9,0.9]	$0_{\pm 1}$	$3278_{\pm 448}$
ViT-B/16	[0.9,0.5,0.9,0.9]	$2364_{\pm 1083}$	$3571_{\pm 381}$

Table G.7: Percentage of samples where different notions of importance identify the same square as the highest contributor on the Four Squares problem. We refer to the case where all definition agree on the square that has the highest contribution as cases of "consensus".

Architecture	$p_{S_{1..k}}$	Importance agreement (softmax)	Importance agreement (logits)	Highest contribution consensus
ResNet-18	[0.9,0.5,0.9,0.9]	$66.26_{\pm 3.59}$	$60.93_{\pm 4.72}$	$57.29_{\pm 4.64}$
ResNet-50	[0.9,0.5,0.9,0.9]	$79.65_{\pm 1.98}$	$74.12_{\pm 2.74}$	$68.50_{\pm 2.24}$
VGG-16	[0.9,0.5,0.9,0.9]	$76.88_{\pm 4.03}$	$78.81_{\pm 2.46}$	$67.21_{\pm 4.48}$
ViT-B/16	[0.9,0.5,0.9,0.9]	$61.88_{\pm 1.96}$	$62.24_{\pm 2.61}$	$52.96_{\pm 2.95}$

Table G.8: The percentage of samples for which a given square was the highest contributor on the Four Squares problem.

Architecture	$p_{S_{1..k}}$	#samples highest contributor			
		1	2	3	4
ResNet-18	[0.9,0.5,0.9,0.9]	$30.74_{\pm 6.84}$	$0.67_{\pm 0.12}$	$47.12_{\pm 8.48}$	$21.47_{\pm 7.06}$
ResNet-50	[0.9,0.5,0.9,0.9]	$38.62_{\pm 1.95}$	$2.90_{\pm 0.71}$	$31.71_{\pm 1.67}$	$26.77_{\pm 3.20}$
VGG-16	[0.9,0.5,0.9,0.9]	$36.18_{\pm 9.66}$	$0.05_{\pm 0.03}$	$37.05_{\pm 4.25}$	$26.71_{\pm 6.82}$
ViT-B/16	[0.9,0.5,0.9,0.9]	$31.98_{\pm 2.45}$	$1.27_{\pm 0.39}$	$33.96_{\pm 0.61}$	$32.79_{\pm 2.41}$

G.5 More complex patterns: ImageNet squares



Figure G.4: Example of samples we create for ImageNet squares. The squares are selected from classes that have a high BagNet17 accuracy.

In this section we explore more complex patterns for the three squares compared to colour. We aim to "cut out" squares from ImageNet [15] classes, and use these as the patterns depicted by our three squares (See Figure G.4).

The backgrounds are replaced by 224x224 images from ImageNet while the foregrounds are replaced by squares sampled from another set of images from ImageNet.

To increase the learnability of the squares, we select ImageNet classes that a pretrained BagNet17 [10] classifies well. Since BagNet17 has a maximum receptive field of 17x17, squares of roughly this size should provide learnable information to the model, depending on which part of the image the square is sampled from. We chose 45 images from each of BagNet17's 10 most accurate classes to make up the sets for the 10 Squares classes. These are yellow lady's slipper, bee eater, zebra, jackfruit, European fire salamander, echidna, limpkin, tiger beetle, sulphur, butterfly and rapeseed. For the background, we mix the images from the 10 least accurate classes to make the set to sample background images. The classes used for this are hook, overskirt, soap dispenser, can opener, screen, space bar, spotlight, purse, water jug and power drill.

The ImageNet Squares dataset is generated from these sets of images. When instantiating the data set, one background image and one foreground image are chosen for each item of the data set. Both of these images are transformed with standard ImageNet transforms. Additionally, the positions of all squares are generated on dataset instantiation using the same procedure as for The Squares Problem. Then, for each image, these squares are copied from the foreground image and inserted into the background image.

No noise is added to the dataset since ImageNet is already noisy. Similarly, the square classes are always correlated to the label to account for the possibility that the squares do not capture informative parts of the image. However, to increase the accuracy of our models, we train them with a square size of 40 for this dataset. For computational reasons, we restrict these additional results to the attributions that had been computed before the rebuttal.

Table G.9: Training and validation accuracies for models trained on The ImageNet Squares Problem. Results reported are as means and standard deviations over 5 different initial seeds of the model weights.

Architecture	$p_{S_{1..k}}$	Training accuracy	Validation accuracy
ResNet-18	[1.0, 1.0, 1.0]	$98.92_{\pm 0.17}$	$89.26_{\pm 1.15}$
ResNet-50	[1.0, 1.0, 1.0]	$99.24_{\pm 0.15}$	$94.90_{\pm 0.57}$

Table G.10: Number of samples eliminated by validity checks for the ImageNet Squares problem. Up to 90% of models fail the checks even with a significantly lower threshold.

Architecture	$p_{S_{1..k}}$	Number eliminated by validity checks	Number eliminated by lack of consensus
ResNet-18	[1.0,1.0,1.0]	$8941.20_{\pm 145.86}$	$178.00_{\pm 33.12}$
ResNet-50	[1.0,1.0,1.0]	$7631.00_{\pm 280.44}$	$622.40_{\pm 133.04}$

Table G.11: Agreement results for models for the ImageNet Squares problem

Architecture	$p_{S_{1..k}}$	Importance agreement (softmax)	Importance agreement (logits)	Highest contribution consensus
ResNet-18	[1.0,1.0,1.0]	$90.26_{\pm 0.86}$	$90.91_{\pm 1.19}$	$83.26_{\pm 1.18}$
ResNet-50	[1.0,1.0,1.0]	$87.29_{\pm 1.24}$	$85.62_{\pm 2.22}$	$73.94_{\pm 2.46}$

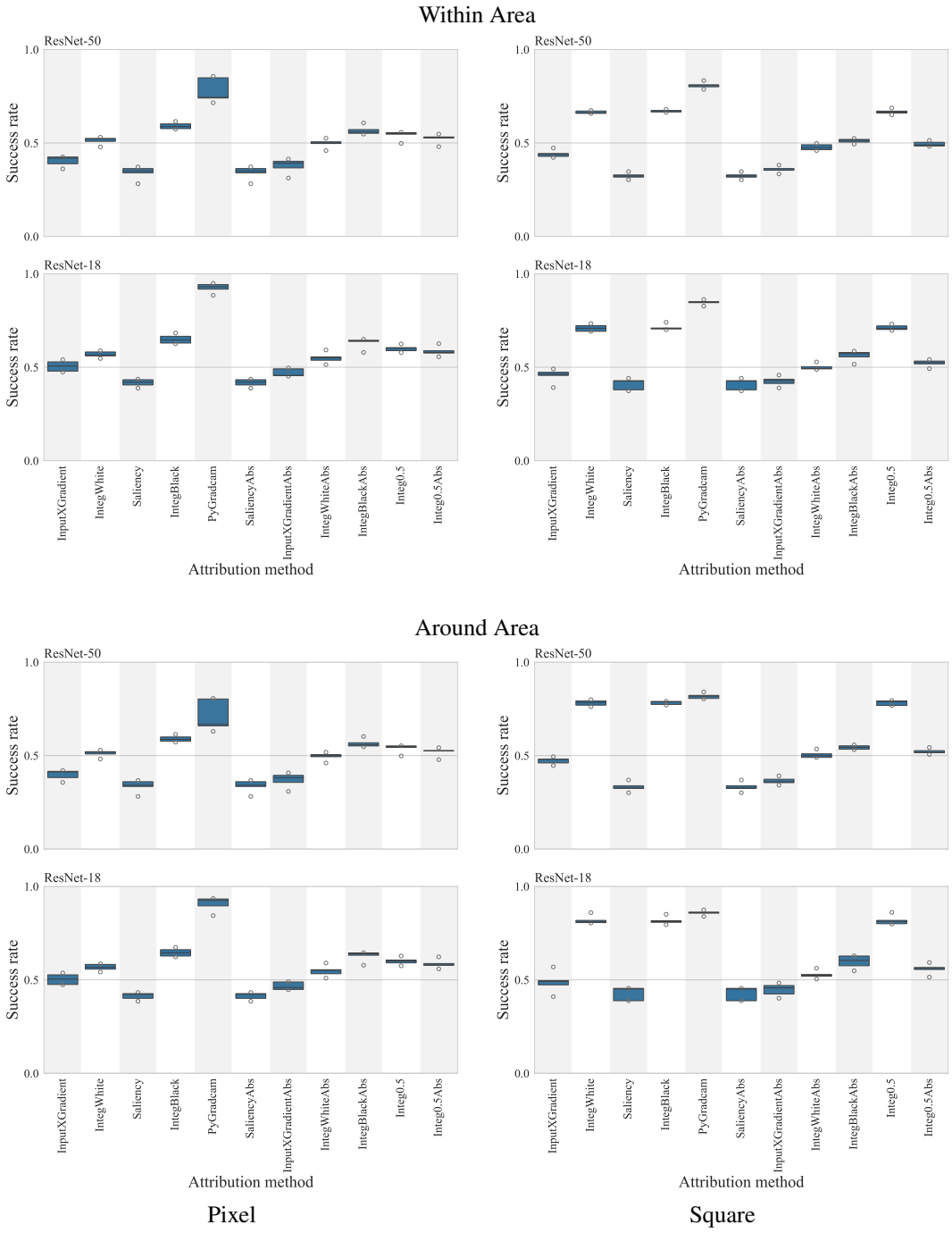


Figure G.5: Success rate plots for ImageNet Squares for all the attribution methods we consider across the different evaluation criteria (square versus pixel, within area versus around area).

G.6 The observed gap is not a result of post-selection

Different settings resulted in varying numbers of samples being eliminated, by both validity checks and consensus (see Table F.2). To consider whether differences in attribution success are due to mismatches in the number of excluded samples, we look at ViT-B/16 on [0.7,0.5,0.9] and [0.9,0.5,0.9] (which we observe to result in a range of focal and diverse model instances) as this is the closest case to having comparable elimination rate for both checks. To be precise focal models lose on average $\sim 16\%$ to the VC1&2 checks, while diverse lose $\sim 13\%$. On consensus checks, focal models lose

$\sim 12\%$, while diverse ones lose $\sim 21\%$. The performance of these models contributes to the plots in Figure 3, but we present them numerically in Table G.12. Note that a gap in attribution success between focal and diverse models still exists for many attribution methods.

We also observe that our conclusions hold for models that passed all the validity checks, as much as for models where more than 70% of the samples are eliminated by the checks. A gap in performance is, therefore, noted both when nothing is filtered, as well as when significant filtering is happening (i.e. both when something is post-selected, as well as when nothing is post-selected).

Table G.12: Attribution success for ViT-B/16 across several attribution methods and 5 seeds for [0.7,0.5,0.9] (focal) and [0.9,0.5,0.9] (diverse) settings. Many of the attributions show a gap in performance between focal and diverse models.

	InputX Gradient	Integ White	SaliencyAbs	Integ Black	Integ WhiteAbs	Integ BlackAbs	Integ 0.5	Integ 0.5Abs
Focal	$39_{\pm 6}$	$70_{\pm 6}$	$78_{\pm 5}$	$69_{\pm 3}$	$78_{\pm 6}$	$68_{\pm 3}$	$65_{\pm 8}$	$78_{\pm 6}$
Diverse	$46_{\pm 7}$	$59_{\pm 2}$	$53_{\pm 2}$	$63_{\pm 1}$	$57_{\pm 1}$	$57_{\pm 2}$	$58_{\pm 1}$	$58_{\pm 0}$

H Choice of baseline for focal and diverse regimes

Based on raw attribution success, we find that many attributions tend to perform better in focal as opposed to diverse scenarios. As discussed in Section 4, when benchmarking attribution performance in artificial setting, one might establish potential baselines. The choice of baseline impacts the **relative** attribution performance.

It is worth noting that the baselines we discussed (i.e. 1. uniformly pick between the three squares and 2. always pick the same square which happens, by chance, to be the highest contributor in most cases) are inherently linked to our specific relaxations of the evaluation process. If, for example, we would consider attributions that fall outside of the squares (which are framework-informed areas), then a chance baseline would set a much lower baseline, while a "always pick the same area" strategy would be less likely to align with one of the squares by chance. We once again note that we typically do not have a reliable way of determining if a real-world model is focal or diverse. Nor do we have access to details about the data-generating process that could inform a "lucky" baseline. Therefore, the choice of baseline needs to be carefully considered if the objective is to establish a comprehensive evaluation of attributions.

Irrespective of the choice of baseline, our main Section 4 observations relate to the variable performance of attributions and the need to account for different model behaviours when designing and evaluating attributions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the proposed framework and other contributions. Furthermore, our claims are evidenced by results in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly acknowledges limitations and assumptions of the framework throughout the text. An explicit limitations paragraph is included in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper presents an empirical evaluation framework rather than a theoretical one.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We fully describe the experimental settings and datasets used with sufficient detail that they could be independently reproduced. Code will be released upon publication of the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code for all aspects of the paper (including artificial dataset generation and the evaluation framework) will be released upon publication. We will also release the weights of sets of models that we trained as a resource for the community to further experiment and build upon.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Design of the datasets is described in Section 3.1 and Appendix G.5. Details of model training is described in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experimental results were computed across 5 runs from models trained on different seeds. We report the means and standard deviations for all results, and in graphical representations using enhanced boxplots we illustrate the distribution through exponentially decreasing quantiles (e.g., 25%, 12.5%, 6.25%, etc.).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Full details of compute resources used are included in Appendix D.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper complies with NeurIPS Code of Ethics. We believe there are no harmful consequences of this research.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No such data or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

13. Answer: [Yes]

    Justification: All code and assets are properly referenced.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

14. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will provide datasheets for both The Squares Problem and the ImageNet derived datasets, as well as the meta-dataset of trained model weights upon publication. Instructions for use will be included.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

15. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not use crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

16. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not include research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

17. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development, scientific approach and original research does not involve use of any LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.