TACKLING THE XAI DISAGREEMENT PROBLEM WITH ADAPTIVE FEATURE GROUPING

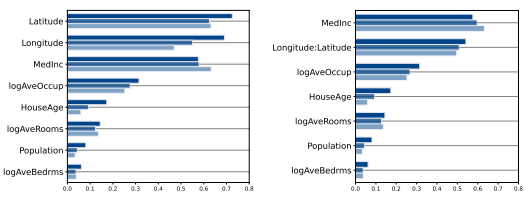
Anonymous authors
Paper under double-blind review

ABSTRACT

Post-hoc explanations aim at understanding which input features (or groups thereof) are the most impactful toward certain model decisions. Many such methods have been proposed (ArchAttribute, Occlusion, SHAP, RISE, LIME, Integrated Gradient) and it is hard for practitioners to understand the differences between them. Even worse, faithfulness metrics, often used to quantitatively compare explanation methods, also exhibit inconsistencies. To address these issues, recent work has unified explanation methods through the lens of Functional Decomposition. We extend such work to scenarios where input features are partitioned into groups (e.g. pixel patches) and prove that disagreements between explanation methods and faithfulness metrics are caused by between-group interactions. Crucially, getting rid of between-group interactions leads to a single explanation that is optimal according to all faithfulness metrics. We finally show how to reduce the disagreements by grouping features on tabular/image data.

1 INTRODUCTION

With the rise in complexity of Machine Learning models, there has also been a rise in concerns regarding the black-box nature of the most performant models. As a result, the field of eXplainable Artificial Intelligence (XAI) has rapidly grown and now proposes a myriad of techniques to "explain" model predictions (Molnar, 2025).



(a) (b)

Figure 1: California Housing. (a) PFI, SHAP, and PDP disagree on the importance of `Longitude`. (b) By grouping features, we increase agreement between the techniques.

One of the main roadblocks to XAI is the so-called *Disagreement Problem* (DP) (Krishna et al., 2022), which refers to inconsistencies between explanation methods. Due to lack of ground truth in XAI, practitioners cannot decide which explanation, if any, is the correct one when they disagree. To address this issue, methods like Shapley Values (Lundberg & Lee, 2017) and the Integrated Gradient (Sundararajan et al., 2017) have been motivated as the *unique* explanations satisfying a set of theoretical properties. As such, they are advertised as a form of ground-truth. Still, in the case of Shapley Values, it was demonstrated that their "Dummy" property can be violated in practice (Sundararajan & Najmi, 2020). Regarding the Integrated Gradient, its properties were proven to be insufficient at specifying a unique explanation (Lerma & Lucas, 2021).

Alternatively, some works benchmark explainability methods using *unfaithfulness metrics* : for example the F-score (Tomsett et al., 2020), μ-Fidelity (Bhatt et al., 2020), INFD (Dai et al., 2022), and Shapley-Weighted Fidelity (SWF) (Muschalik et al., 2025). Unfortunately, unfaithfulness metrics were previously shown to be inconsistent : an explanation can be ranked first by a metric and ranked last by another (Tomsett et al., 2020).

Recent work has unified the various explanation techniques through the lens of Functional Decomposition (Fumagalli et al., 2025; Deng et al., 2024). We extend these efforts to scenarios where features are partitioned into groups (*e.g.* pixel patches for images), we identify the root cause of disagreements between explainers and faithfulness metrics: *between-group interactions*, and we

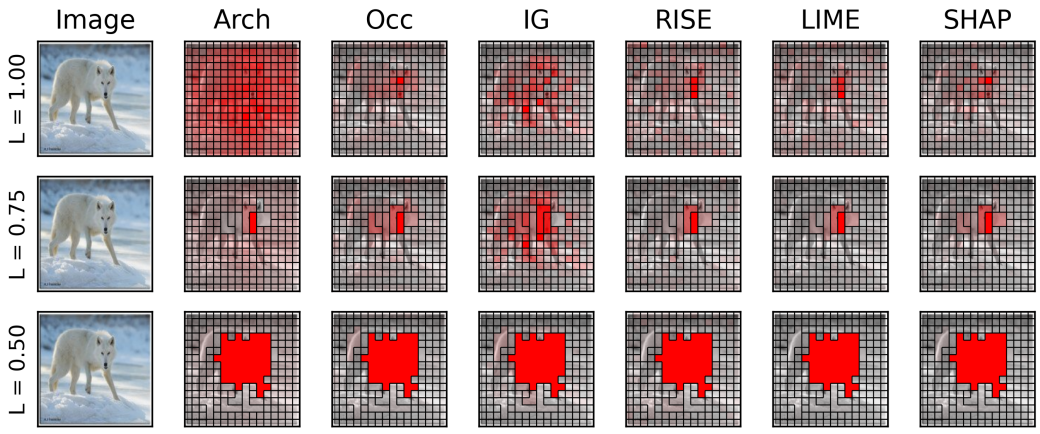


Figure 2: Explaining the "White Wolf" prediction of a ResNet18 using several saliency map methods. (Top), When considering 14×14 patches, the saliency maps highlight different parts of the image. (Middle) To minimize a disagreement objective function L, AGREED fuses the various patches. For example, the initial patches covering the wolf's eyes and nose are fused. (Bottom), Eventually, AGREED leads to large patches where agreement among saliency map techniques is increased.

minimize it by adaptively grouping features. Our framework, called Adaptive Grouping to REduce Explanation Disagreements (AGREED), is empirically assessed on tabular and image datasets. Figures 1 and 2 illustrate AGREED on the California and ImageNet datasets respectively. To resume, our contributions are

1. Unifying group importance methods and faithfulness metrics through Functional Decomposition and demonstrate that disagreements are caused by between-groups interactions.
2. Proposing the AGREED algorithm to discover feature groups and assessing its performance on tabular and image datasets.

2 BACKGROUND

2.1 FUNCTIONAL DECOMPOSITIONS

All notation used throughout the paper is enumerated in Appendix B.1. We let $[d] := \{1, \dots, d\}$ be a set of d features, $\mathcal{X} \subseteq \mathbb{R}^d$ be the input domain, $\boldsymbol{x} \in \mathcal{X}$ be an arbitrary input, $f : \mathcal{X} \to \mathbb{R}$ be a model, and $\mathcal{D}$ be the data distribution of inputs ($\boldsymbol{x} \sim \mathcal{D}$). Given a feature subset $u \subseteq [d]$, we denote its cardinality by $|u|$. Functional Decomposition aims to represent f as a sum of 2^d sub-functions

$$f(\boldsymbol{x}) = \sum_{u \subseteq [d]} f_u(\boldsymbol{x}), \tag{1}$$

where f_u only depends on $(x_j)_{j \in u}$. The term $f_\emptyset$ is a constant, the terms f_u for $|u| = 1$ are called *main-effect* while the terms $|u| \geq 2$ are referred to as $|u|$*-way interactions*. Functional Decompositions are not unique and their definition is often based on a heuristic for removing feature x_j. One heuristic consists of freezing x_j at a baseline value b_j using the *replace-function* $\boldsymbol{r}_u : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$ defined as

$$r_u(\boldsymbol{b}, \boldsymbol{x})_j = x_j \text{ if } j \in u \text{ otherwise } b_j. \tag{2}$$

Treating the baseline $\boldsymbol{b} \sim \mathcal{B}$ as random leads to the *Marginal Decomposition* (Fumagalli et al., 2025).

Definition 2.1 (Marginal Decomposition)**.** *Given a distribution $\mathcal{B}$, the Marginal Decomposition is*

$$f_{u,\mathcal{B}}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{b} \sim \mathcal{B}}[f(\boldsymbol{r}_u(\boldsymbol{b}, \boldsymbol{x}))] - \sum_{v \subset u} f_{v,\mathcal{B}}(\boldsymbol{x}). \tag{3}$$

When $\mathcal{B} = \delta_{\boldsymbol{b}}$ is a Dirac delta centered at $\boldsymbol{b}$, the decomposition falls back to the so-called Anchored-Decomposition (Kuo et al., 2010). If $\mathcal{B} = \mathcal{B}_{\text{ind}} := \prod_{j=1}^{d} \mathcal{B}_i$ (*i.e.* input features are independent), the Marginal Decomposition becomes the *ANOVA Decomposition* (Hooker, 2004) (see Appendix B.2). Since the Marginal Decomposition is relative to a distribution $\mathcal{B}$, it cannot explain model prediction $f(\boldsymbol{x})$ in isolation. Rather, the main effects and interactions explain the *Gap* $f(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{b}\sim\mathcal{B}}[f(\boldsymbol{b})]$ between a specific prediction and the average prediction. For Tabular data, it is common to let $\mathcal{B} = \mathcal{D}$ be the data distribution (Lundberg & Lee, 2017), while for Images it is common to use a single baseline image $\boldsymbol{b}$: the average color (Ribeiro et al., 2016) or $\mathbf{0}$ (Petsiuk et al., 2018).

2.2 Explaining Feature Groups

Many explainability methods explain the joint effect of feature groups, instead of their individual effects. This is the case for saliency maps that first group pixels into super-pixels (Ribeiro et al., 2016; Tsang et al., 2020) or square patches (Zeiler & Fergus, 2014; Petsiuk et al., 2018). Assuming d features are fed to the model, consider a partition of $[d]$ into D disjoint groups. This partition can be described with a function $\mathcal{P} : [d] \to [D]$ that associates each feature $j \in [d]$ to its group index $\mathcal{P}(j) \in [D]$. We will employ the mapping $\mathcal{P}(u) := \{\mathcal{P}(j) : j \in u\}$ for $u \subseteq [d]$ to enumerate all groups indices within a $|u|$-way interaction and the inverse map $\mathcal{P}^{-1}(U) := \{j \in [d] : \mathcal{P}(j) \in U\}$ for $U \subseteq [D]$ to list all features that are part of certain groups. Finally, $\mathcal{P}'$ is a super-partition of $\mathcal{P}$ if $\mathcal{P}(j) = \mathcal{P}(k) \Rightarrow \mathcal{P}'(j) = \mathcal{P}'(k)$.

When investigating the effect of feature groups on the gap $f(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{b}\sim\mathcal{B}}[f(\boldsymbol{b})]$, the ideal scenario is that of a Groupwise Additive Model.

Definition 2.2 (Groupwise Additive Model (Sivill & Flach, 2023))**.** *Let $R \subseteq \mathcal{X}$ be a hyperrectangle region. A model $f : \mathcal{X} \to \mathbb{R}$ is called Groupwise additive in R w.r.t $\mathcal{P}$ if there exists D functions $g_{\mathcal{P}^{-1}(\{i\})}$ that each only depend on features in group i and such that*

$$f(\boldsymbol{x}) = \omega_0 + \sum_{i=1}^{D} g_{\mathcal{P}^{-1}(\{i\})}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in R. \tag{4}$$

In this ideal scenario, the contribution of group i toward the gap $f(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{b}\sim\mathcal{B}}[f(\boldsymbol{b})]$ is unambiguous: $g_{\mathcal{P}^{-1}(\{i\})}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{b}\sim\mathcal{B}}[g_{\mathcal{P}^{-1}(\{i\})}(\boldsymbol{b})]$. If Equation 4 does not hold however, there is no longer a unique group attribution. This has caused the development of a myriad of post-hoc explainers: $\boldsymbol{\phi}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) \in \mathbb{R}^D$ that estimate the contribution to each group $i \in [D]$. Many methods are conveniently defined in terms of a coalitional game.

Definition 2.3 (Grouped Coalitional Game)**.** *Define the coalitional game $\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}$*

$$\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) := \mathbb{E}_{\boldsymbol{b}\sim\mathcal{B}}[f(\boldsymbol{r}_{\mathcal{P}^{-1}(U)}(\boldsymbol{b}, \boldsymbol{x}))] \quad \forall\, U \subseteq [D]. \tag{5}$$

that applies the replace-function simultaneously to all features within the same group.

Various explainability methods can be expressed as a weighted sum of marginal contributions for group i

$$\phi_i^{\mu}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) = \sum_{U \subseteq [D]\setminus\{i\}} \mu(U)[\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U \cup \{i\}) - \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U)], \tag{6}$$

with $\mu(U) \in \mathbb{R}^+$ such that $\sum_{U\subseteq[D]\setminus\{i\}} \mu(U) = 1$. The joint-PDP (Friedman, 2001) and ArchAttribute (Tsang et al., 2020) employs $\mu(\emptyset) = 1$ and zero otherwise. The joint-PFI (Au et al., 2022) and Patch-Occlusion (Zeiler & Fergus, 2014) consider $\mu([D] \setminus \{i\}) = 1$ and zero otherwise. Finally, the SHAP(Lundberg & Lee, 2017) uses $\mu(U) = \binom{D-1}{|U|}^{-1}/D$, while RISE (Petsiuk et al., 2018) defines $\mu(U) = 2^{D-1}$.

The Integrated Gradient (Sundararajan et al., 2017) is an alternative feature importance that is not easily expressed in terms of a coallitional game. Nevertheless, it is naturally extended to feature groups

$$\phi_i^{\text{IG}}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) = \sum_{j \in \mathcal{P}^{-1}(\{i\})} \mathbb{E}_{\substack{\boldsymbol{b}\sim\mathcal{B} \\ t\sim\text{Uniform}(0,1)}} \left[(x_j - b_j)\, \frac{\partial f}{\partial x_j}((1-t)\boldsymbol{b} + t\boldsymbol{x}) \right], \tag{7}$$

by summing over all features within group i (Tsang et al., 2020). The calculation of post hoc explanations can require evaluating expectations involving $\mathcal{B}$. Throughout our experiments, all expectations were estimated via Monte Carlo (MC) with N samples.

2.3 Disagreement Problem

Many post-hoc explainers have been proposed, each with different motivations. So, it is not surprising that they often disagree in practice (Krishna et al., 2022). To characterize said disagreements, we use the L_2 metric (Laberge et al., 2024)

$$D_{L_2}(\boldsymbol{\phi}, \boldsymbol{\phi}') := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\|\boldsymbol{\phi}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) - \boldsymbol{\phi}'(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P})\|^2]. \tag{8}$$

The current literature tackles disagreements between explanation methods by benchmarking them using *unfaithfullness metrics*. Many such metrics take the form

$$\overline{F}(\boldsymbol{\phi}, f, w) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\sum_{U\subseteq[D]} w(U)\left(\sum_{i\in U}\phi_i(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) - [\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}([D]) - \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}([D]\setminus U)]\right)^2\right] \tag{9}$$

or

$$\underline{F}(\boldsymbol{\phi}, f, w) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\sum_{U\subseteq[D]} w(U)\left(\sum_{i\in U}\phi_i(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) - [\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) - \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\emptyset)]\right)^2\right] \tag{10}$$

using a specific weight $w(U) \in \mathbb{R}^+$ for any $U \in [D]$ *e.g.* Sensitivity-n (Ancona et al., 2017), INFD (Yeh et al., 2019), μ-Fidelity (Bhatt et al., 2020), Shapley-Weighted Fidelity (SWF) (Muschalik et al., 2025). We refer to Appendix B.3 for the complete definition of each metric. Note that the choice of weighing scheme impacts which explanation is considered optimal: minimizing $\underline{F}$ with $w(U) = 2^{-D}$ as done in LIME (Ribeiro et al., 2016) leads to the Banzhaf index (Tsai et al., 2023), minimizing $\underline{F}$ with $w(U) = \frac{d-1}{\binom{D}{|U|}|U|(D-|U|)}$ leads to SHAP(Lundberg & Lee, 2017).

3 METHODOLOGY

Because of the aforementioned inconsistencies, practitioners do not have a mean of determining which unfaithfulness metric, and by extend which explanation method, is correct. Each method employs a different weighting scheme μ to aggregate marginal contributions (cf Equation 6) or a different weight w to aggregate errors of additive reconstructions (cf. Equations 9 & 10). Our next objective is to lessen the impact of the choices for μ and w.

Theorem 3.1. *The Arch/Occ/LIME/SHAP/RISE Attributions can be expressed via the Marginal Decomposition*

$$\phi_i^{\mu}(\boldsymbol{x}, f, \mathcal{B}, \mathcal{P}) = \sum_{u\subseteq[d]:\{i\}=\mathcal{P}(u)} f_{u,\mathcal{B}}(\boldsymbol{x}) + \sum_{u\subseteq[d]:i\subsetneq\mathcal{P}(u)} h(|\mathcal{P}(u)|) f_{u,\mathcal{B}}(\boldsymbol{x}), \tag{11}$$

where h is different for each μ. Different explainability methods disagree on how to redistribute ***between-group interactions*** *($f_{u,\mathcal{B}}(\boldsymbol{x})$ with $|\mathcal{P}(u)| \geq 2$) among the groups involved.*

Theorem 3.2. *Let $R \subseteq \mathcal{X}$ be a hyperrectangle region such that supp$(\mathcal{D}) \subseteq R$ and supp$(\mathcal{B}) \subseteq R$. Let $\mathcal{P}$ be a feature partition. Whenever the model f is groupwise additive in R w.r.t $\mathcal{P}$, any unfaithfulness metrics that follow Equations 9 and 10 are all simultaneously optimized*

$$F(\boldsymbol{\phi}^{\mu}, f, w) = 0 \tag{12}$$

for any weight function w and attribution $\boldsymbol{\phi}^{\mu}$.

The proofs are presented in Appendix C.2. Both theorems imply that if there were no between-group interactions (*i.e.* $f_{u,\mathcal{B}} = 0$ whenever $|\mathcal{P}(u)| \geq 2$), then **all explanation method would agree** on the group importance and **all unfaithfulness metrics** be minimized. Therefore, the Disagreement Problem can potentially be tackled by searching for a partition $\mathcal{P}$ with respect to which f is group-wise additive, or "almost" group-wise additive. This is trivially achieved by considering a single group containing all features, so we must trade-off explanation agreement with group sizes. Similarly to recent work, we frame this search as an optimization w.r.t $\mathcal{P}$ using a special class of loss functions.

Definition 3.1. *A partition loss function $\mathcal{L}_f(\mathcal{D}, \mathcal{B}, \mathcal{P}) \in \mathbb{R}^+$ should respect:*

1. *If f is groupwise additive in R w.r.t $\mathcal{P}$, such that supp$(\mathcal{D}) \subseteq R$ and supp$(\mathcal{B}) \subseteq \mathcal{X}$, then $\mathcal{L}_f(\mathcal{D}, \mathcal{B}, \mathcal{P}) = 0$.*
2. *If f is additive w.r.t to group i (i.e. $f(\boldsymbol{x}) = g_{\mathcal{P}^{-1}(\{i\})}(\boldsymbol{x}) + g_{\mathcal{P}^{-1}([D]\setminus\{i\})}(\boldsymbol{x})$), fusing group i with another one does not impact the loss.*
3. *If $\mathcal{P}'$ is a super-partition of $\mathcal{P}$, $\mathcal{L}_f(\mathcal{B}_{ind}, \mathcal{B}_{ind}, \mathcal{P}') \leq \mathcal{L}_f(\mathcal{B}_{ind}, \mathcal{B}_{ind}, \mathcal{P})$.*

These properties ensure that $\mathcal{L}_f$ is a sensible objective to minimize w.r.t $\mathcal{P}$. Property 1 guarantees convergence once the optimal $\mathcal{P}$ is found. Property 2 encourages the minimization algorithm to only fuse groups that interact with some other. Property 3 suggests that, on tabular data where $\mathcal{B} = \mathcal{D}$, an iterative algorithm monotonically decreases its loss if features are independent.

Theorem 3.3. *The L_2 disagreements $D_{L_2}(\boldsymbol{\phi}^{Occ}, \boldsymbol{\phi}')$ between Occlusion and the Arch/LIME/RISE/SHAP explainers respect Definition 3.1. Proof in Appendix C.3*

Some pairings of explainers (*e.g.* LIME vs. SHAP) are not considered since they might break Property 3. Theorem 3.3 suggests leveraging the L_2 disagreements between explanation methods to infer feature groups. The following disagreements among the PDP/ArchAttribute and PFI/Occlusion

$$\mathcal{L}_f^{\text{AGREED}}(\mathcal{D}, \mathcal{B}, \mathcal{P}) := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\sum_{i=1}^{D}(\phi_i^{\text{PDP/Arch}}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) - \phi_i^{\text{PFI/Occ}}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}))^2\right] = \sum_{i=1}^{D}\Psi(i) \tag{13}$$

were considered since these explanations are the cheapest to compute. Our minimization of equation 13 w.r.t $\mathcal{P}$ follows: 1) start with a granular partition $\{\{1\}, \ldots, \{d\}\}$; 2) select the group i with highest potential $\Psi(i)$; 3) compute its pair-wise interactions with a set of candidate groups i'; 4) fuse group i with the group i' that yields the maximal pairwise interaction; 5) repeat until the objective (cf. Equation 13) falls below ϵ. We refer to Appendix D for the technical details.

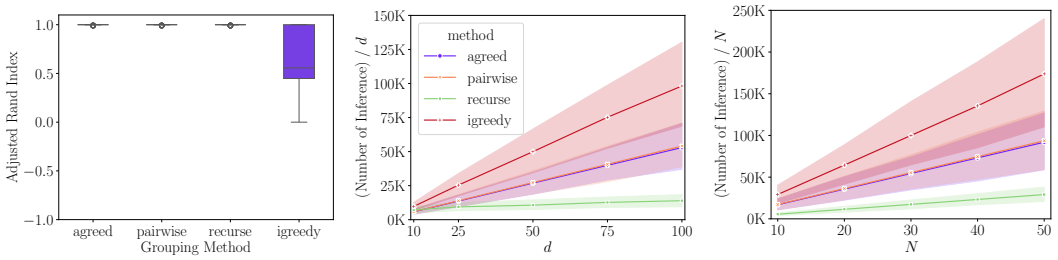
Solving $\min_{\mathcal{P}} \mathcal{L}_f^{\text{AGREED}}(\mathcal{D}, \mathcal{B}, \mathcal{P})$ yields a feature partition that is valid over the support of the distributions $\mathcal{D}$ and $\mathcal{B}$ passed as parameters. In the tabular setting, we set $\mathcal{D}$ and $\mathcal{B}$ to the data distribution so solving $\min_{\mathcal{P}} \mathcal{L}_f^{\text{AGREED}}(\mathcal{D}, \mathcal{D}, \mathcal{P})$ leads to a partition useful to explain any data point. For image data, we set $\mathcal{D} = \delta_{\boldsymbol{x}}$ and $\mathcal{B} = \delta_{\boldsymbol{b}}$ to Dirac measures over an image of interest and the baseline. Accordingly, solving $\min_{\mathcal{P}} \mathcal{L}_f^{\text{AGREED}}(\delta_{\boldsymbol{x}}, \delta_{\boldsymbol{b}}, \mathcal{P})$ leads to a partition $\mathcal{P}$ that is only valid for this single image $\boldsymbol{x}$. The partitioning algorithm must be run separately on each image.

4 Related Work

There are multiple existing methods for grouping input features. The Pairwise algorithm advocated by Tsang et al. (2020) groups features in three-steps: 1) computes all $d(d-1)/2$ pairwise interactions between features; 2) Retain only the interaction whose strength is above a threshold ϵ and organize them in a graph; 3) define groups as the cliques of said graph. The Recursive algorithm proposed by Sivill & Flach (2023) was demonstrated to scale as $\mathcal{O}(d \log d)$ on tabular data $[d]$ into two sets $\{1\}$ and $\{2 \ldots, d\}$. Finally, the iGreedy algorithm introduced by Xu et al. (2024) starts from a granular partition $\{\{1\}, \ldots, \{d\}\}$, and progressively fuses pairs of groups until convergence.

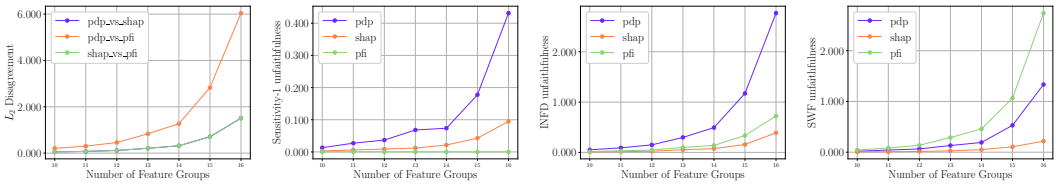
Laberge et al. (2024) recently identified feature interactions as the root-cause of disagreement between PDP/SHAP/PFI and minimized them by restricting the baseline distribution $\mathcal{B}$ to rule-based regions. Although promising for tabular data, rule-based regions do not work on pixels. AGREED, in contrast, minimizes disagreements by partitioning features into disjoint groups, a methodology that works for tabular data **and** images.

Prior work has already unified most explainability methods through Functional Decomposition. (Deng et al., 2024) have previously unified 14 saliency maps while (Fumagalli et al., 2025) developed a categorization of explanations along three axes: conditional-marginal-anchored decompositions, pure-partial-full explanations, and individual-joint-interactions effects. Unlike ours, these existing frameworks do not consider disjoint feature groups. Moreover, while prior frameworks also highlight interactions as the root-cause of disagreements, ours is the first to propose a practical methodology to minimize said disagreements.



(a) (b) (c)

Figure 3: Synthetic tabular data. (a) RandIndex comparing ground-truth partitions to each grouping method. (b) Scalability of each method w.r.t the number of features d. (c) Scalability of each method w.r.t the number of Monte Carlo samples N. Confidence bands range from 5^{th} percentile to the 95^{th}.



(a) (b) (c) (d)

Figure 4: Tabular data: EBM fitted on Marketing. Explanation L_2 Disagreement and unfaithfulness metrics as a function of the number of feature groups.

5 EXPERIMENTS

5.1 TABULAR DATASETS

5.1.1 KNOWN GROUND-TRUTHS ON SYNTHETIC DATA

We compared AGREED with iGreedy, Recursive, and Pairwise on synthetic data where the model f is known to be groupwise additive over $\mathcal{X}$ w.r.t some ground-truth partition $\mathcal{P}^\star$. The exact and estimated partitions were compared using the RandIndex (Hubert & Arabie, 1985), a score between 0 and 1 that is maximized for identical partitions. We set $\mathcal{B} = \mathcal{D}$ as the data distribution. For various numbers of dimension d, MC samples N, and five random seeds, we generated random data/models with correlated features (see Appendix E.1). Since an optimal partition exists, we set $\epsilon = 10^{-10}$. All tabular experiments were run on a laptop with an 11th Gen Intel(R) Core(TM) i7-11850H CPU, 16 threads, and 32 GiB of RAM.

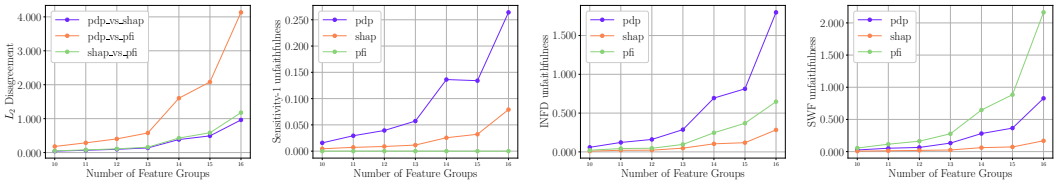
Figure 3 presents the results aggregated over five seeds. According to Figure 3 (a), only iGreedy fails to consistently find the optimal partition. We suspect this is because its termination criterion assumes its objective function monotonically decreases. From Figures 3 (b) and (c), AGREED and Pairwise have a similar performance, but it is the recursive method that scales best with d and N.

5.1.2 REAL DATASETS

When studying real-world black-boxes, we no longer know the optimal partitions. We instead evaluate AGREED ability to reduce PDP/SHAP/PFI disagreements and Sensitivity-1, INFD, and SWF unfaithfulness scores.

The Marketing[1] ($d = 16$), Default-Credit[2] ($d = 23$), SPAM[3] ($d = 57$), and NOMAO[4] ($d = 118$) datasets were investigated as they contain a large number of features, which renders computing all

[1] https://archive.ics.uci.edu/dataset/222/bank+marketing
[2] https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients
[3] https://archive.ics.uci.edu/dataset/94/spambase
[4] https://archive.ics.uci.edu/dataset/227/nomao



(a) (b) (c) (d)

Figure 5: Tabular data: GBT fitted on Marketing. explanation L_2 disagreement and unfaithfulness metrics as a function of the number of feature groups.

pairwise interactions intractable. We split each dataset into train/test sets with ratio 0.9:0.1, and we trained Explainable Boosting Machines (EBM) (Nori et al., 2019) from the `InterpretML` library and Histogram Gradient Boosted Trees (HGBT) from the `ScikitLearn` package (Pedregosa et al., 2011). Hyperparameters were tuned using 5-Fold CV and random search.

The AGREED partitioning algorithms requires a random subset of N data samples to approximate expectations w.r.t $\mathcal{D}$. For the Marketing and Default-Credit, we considered $N = 100$ samples and for the other two, we employed $N = 50$ samples. To account for the randomness that arises from subsampling large datasets, we repeated the subsampling of N points five times using different random seeds, ran the partitioning algorithms, and reported the L_2 disagreements and Sensitivity-1, INFD, SWF infidelity scores. The runtimes for AGREED ranged from 10 to 100 seconds. Note that the partitioning only needs to be done once and can used to compute local explanation on any data point.

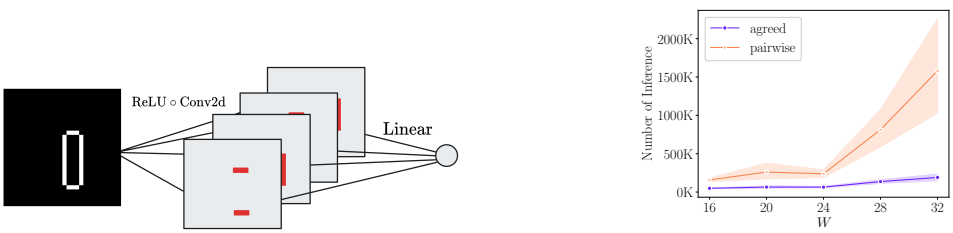
Figures 4 & 5 present the tradeoffs between disagreements/unfaithfulness and number of feature groups on the Marketing dataset. Other datasets exhibit how a similar trend, see Appendix E.2.1. From both Figures (left), we observe that AGREED is able to reduce disagreements between any pairing of explainers although the algorithm is designed to minimize differences between PDP and PFI. This highlights the role of between-group interactions in the disagreements between explainability methods. Moreover, looking at the three unfaithfulness scores, we note inconsistencies with their ranking. For instance, Sensitivity-1 claims that PFI is the most faithful explanation, while the other two metrics claim that SHAP is most faithful. Also note that INFD and SWF disagree on the ranking between PDP and PFI. However, by reducing the number of groups (*i.e.* grouping more features together), all unfaithfulness metrics collectively converge to zero for either PDP/SHAP/PFI.

Grouping interacting features reduces the inconsistencies between the explainability methods and unfaithfulness scores. Nevertheless, the resulting explanation must be interpreted with care since the joint-attribution of a group is a *multivariate* function of all features involved. Appendix E.2.2 presents practical examples of how to interpret the joint influence of groups containing at most three features.

5.2 Images

Convolutional Neural Networks (CNN) remain a strong baseline across a wide range of image domain tasks. Since these models involve the composition on multiple non-linear spatial filters, it is unrealistic to find a single partition $\mathcal{P}$ w.r.t which f is groupwise additive across all images from the dataset. Instead, by fixing an image of interest $\boldsymbol{x}$, a baseline $\boldsymbol{b}$, and setting $\mathcal{D} = \delta_{\boldsymbol{x}}$ and $\mathcal{B} = \delta_{\boldsymbol{b}}$, AGREED will search for a partition w.r.t which the model is group-wise additive in the region $\prod_{j=1}^{d}[b_j, x_j] \subset \mathcal{X}$. While this could leads to explanations with reduced disagreements and increased faithfulness, this also implies that AGREED must be rerun for each individual image $\boldsymbol{x}$ leading to a partition that is only valid for this image.

Moreover, since pixels have an inherent sense of proximity, we further restrict the partition to describe D *path-connected* patches: any two pixels within a patch must be connected via a path spanning said patch. To produce such patches, group fusion is only performed in AGREED if two patches share a boundary. Image experiments were run using a NVIDIA GeForce RTX 3090 GPU with 25GiB of RAM.



(a) (b)

Figure 6: Synthetic image example. (a) A toy CNN can detect vertical and horizontal edges to perfectly classify rectangles while being locally group-wise additive. (b) Comparing the scalability of AGREED and Pairwise as the image size W increases.

5.2.1 Known Ground-Truths on Synthetic Data

We first experimented on a synthetic image dataset for which ground-truth partitions are known. The dataset consisted of $W \times W$ images with a random rectangle drawn into them. Each image is labeled as $y = 1$ if the rectangle is tall and $y = 0$ if it is wide. This classification problem can be solved exactly with a CNN that composes a convolution layer (with 2×2 filters that detect left, right, up, and down edges), a ReLU non-linearity, and a linear layer (see Figure 6 (a)). Although this model is not groupwise additive over its whole domain, it is groupwise additive in the region $\prod_{j=1}^{d}[0, x_j]$ and the optimal partition $\mathcal{P}^\star$ that considers each of the four rectangle edges as a separate group. Fixing the baseline $\boldsymbol{b} = \mathbf{0}$, we ran the AGREED and Pairwise methods (iGreedy and Recurse were not developed for images), and confirmed that they each converge to the optimal partition (RandIndex is systematically 1.0).

However, the Pairwise approach is much more computationally expensive than AGREED, as evidenced by Figure 6 (b). Indeed, Pairwise scales poorly w.r.t W compared to AGREED. This is because the number of pairwise interactions to consider is $\mathcal{O}(W^4)$. AGREED avoids this complexity by making the assumption that pixel interactions in a CNN model are local. This assumption leads to efficient convergence on this synthetic data, see Figure 7 for a qualitative example.

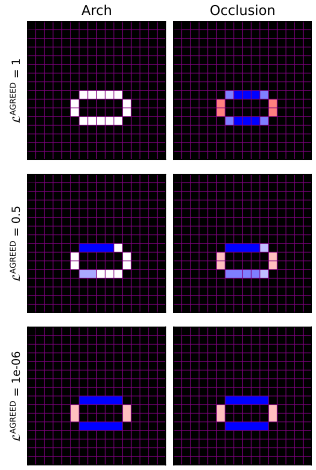
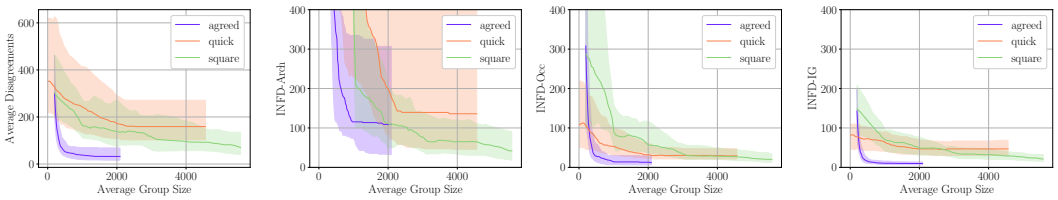


Figure 7: AGREED finds four patches to increase agreement between Arch and Occ on toy image data.

5.2.2 Saliency Maps on MiniImageNet

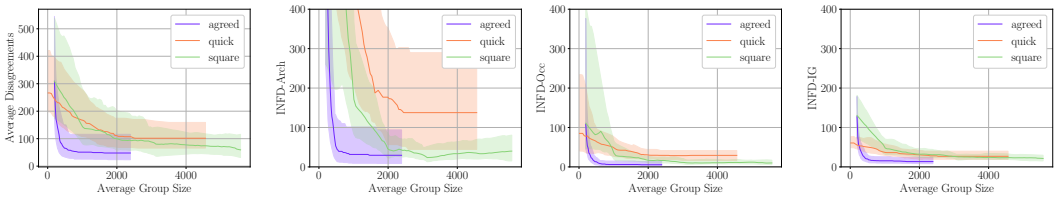
We studied VGG16 and ResNet18 pre-trained on ImageNet and explained their predictions on the MiniImageNet subset containing 100 classes and 600 images per class (Ravi & Larochelle, 2017). Given the lack of ground-truth partitions, we compared partitioning algorithms tradeoffs between disagreement/unfaithfulness and patch size. The Pairwise grouping algorithm was not investigated because, on a realistic ConvNet, the resulting patches are no longer guaranteed to be path-connected. Instead, we compared AGREED with two algorithms previously used on ImageNet: the Quickshift image segmentation algorithm used by LIME and ArchAttribute, and $W \times W$ squares patches implicit to Occlusion and RISE. To accelerate AGREED, we started with a partition of small 14×14 patches. AGREED took on average 3-15 seconds per image to generate a partition.

Figures 8 & 9 present the tradeoffs between disagreement/unfaithfulness and patch sizes for the AGREED, Quickshift, Square partitioning algorithms methods using the zero baseline $\boldsymbol{b} = \mathbf{0}$. To compute a single disagreement score, we averaged disagreements between 100 randomly chosen test images $\boldsymbol{x}$ and also averaged disagreements between all pairings of the Arch/Occ/IG/LIME/SHAP explainers. RISE was excluded because it is equivalent to LIME. Unfaithfulness metrics were also estimated using the same 100 test set images. From both figures, we see that AGREED offers the



(a) (b) (c) (d)

Figure 8: ResNet18 pretrained on ImageNet. We report the average Arch/Occ/LIME/SHAP/IG Disagreement and the INFD unfaithfulness metric of Arch, Occ, and IG.



(a) (b) (c) (d)

Figure 9: VGG16 pretrained on ImageNet. We report the average Arch/Occ/LIME/SHAP/IG Disagreement and the INFD unfaithfulness metric of Arch, Occ, and IG.

most competitive tradeoffs. We also conclude that, the more the different saliency map agree with each other, the more faithful they are to the model.

Appendix E.3 presents saliency maps yielded by AGREED, which tend to generate one large important patch that covers the object being classified, although there are exceptions. This observation suggests that, when explaining the function f that maps pixels to class logit, it might be unrealistic to expect saliency maps that are simultaneously unambiguous (all saliency maps agree), faithful (unfaithfulness metrics are minimized) and granular (the patches highlight distinct object parts). In future work, we envision using AGREED to explain the function f that maps input pixels to concept activations in hidden layers (Fel et al., 2023) instead of logits in the hopes of obtaining more granular and semantically meaningful saliency maps.

6 CONCLUSION

We unified feature groups explanation methods through Functional Decomposition. We identified the culprit that prohibits agreement among the methods and unfaithfulness metrics: between-group features interactions, and minimized it using an algorithm that iteratively fuses feature groups when they strongly interact. On two data modalities, our algorithm named AGREED was demonstrated to efficiently reduce inconsistencies between explanation techniques and unfaithfulness metrics. AGREED is broadly applicable to Tabular and Image data although both structures are treated differently in the algorithm. Future efforts could extend it to other modalities where groups are natural *e.g.* time-series, text.

For tabular data, the remaining challenge is to provide an automatic methodology for visualizing the joint attribution of large feature groups. We envision combining AGREED with regional based explanations (Laberge et al., 2024) to interpret these high-dimensional functions regionally. For images, we managed to make the ArchAttribute/Occlusion/IG/RISE/LIME/SHAP saliency maps agree on "where" the network is looking. Yet, it is not clear "what" the network is seeing. Combining AGREED along a concept-based technique (Fel et al., 2023) could address this issue as well as making the saliency maps more granular. Finally, future work should apply AGREED on image-based Transformers.

7 Reproducibility Statement

To ensure the reproducibility of our empirical results, we are engaged to make the source code public upon acceptance. All Machine Learning packages employed are open-source (*e.g.* Scikit-Lean, InterpretML, Pytorch), with access to the code could rerun our scripts.

References

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

Quay Au, Julia Herbinger, Clemens Stachl, Bernd Bischl, and Giuseppe Casalicchio. Grouped feature importance and combined features effect plot. *Data Mining and Knowledge Discovery*, 36(4):1401–1450, 2022.

Umang Bhatt, Adrian Weller, and José MF Moura. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H Bach, and Himabindu Lakkaraju. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 203–214, 2022.

Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, Ziwei Yang, Zheyang Li, and Quanshi Zhang. Unifying fourteen post-hoc attribution methods with taylor interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4625–4640, 2024.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721, 2023.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Fabian Fumagalli, Maximilian Muschalik, Eyke Hüllermeier, Barbara Hammer, and Julia Herbinger. Unifying feature-based explanations with functional anova and cooperative game theory. In *International Conference on Artificial Intelligence and Statistics*, pp. 5140–5148. PMLR, 2025.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

John C Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963.

Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 575–580, 2004.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

F Kuo, I Sloan, Grzegorz Wasilkowski, and Henryk Woźniakowski. On decompositions of multivariate functions. *Mathematics of computation*, 79(270):953–966, 2010.

Gabriel Laberge, Yann Batiste Pequignot, Mario Marchand, and Foutse Khomh. Tackling the xai disagreement problem with regional explanations. In *International Conference on Artificial Intelligence and Statistics*, pp. 2017–2025. PMLR, 2024.

Miguel Lerma and Mirtha Lucas. Symmetry-preserving paths in integrated gradients. *arXiv preprint arXiv:2103.13533*, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777, 2017.

Jean-Luc Marichal, Ivan Kojadinovic, and Katsushige Fujimoto. Axiomatic characterizations of generalized values. *Discrete Applied Mathematics*, 155(1):26–43, 2007.

Christoph Molnar. *Interpretable Machine Learning*. 3 edition, 2025. ISBN 978-3-911578-03-5. URL `https://christophm.github.io/interpretable-ml-book`.

Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. shapiq: Shapley interactions for machine learning. *Advances in Neural Information Processing Systems*, 37:130324–130357, 2025.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, pp. 307–317, 1953.

Torty Sivill and Peter Flach. Shapley sets: Feature attribution via recursive function decomposition. *arXiv preprint arXiv:2307.01777*, 2023.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. Sanity checks for saliency metrics. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6021–6029, 2020.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-shap: The faithful shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? interpretable attribution for feature interactions. *Advances in neural information processing systems*, 33:6147–6159, 2020.

Sascha Xu, Joscha CÃžppers, and Jilles Vreeken. Succinct interaction-aware explanations. *arXiv preprint arXiv:2402.05566*, 2024.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019.

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

A Broader Impacts

The penultimate goal of explainability is to provide confidence (or lack thereof when necessary) into a model before its deployment. For instance, we want to make sure that the model has learned the intended patterns to solve the task and is not relying on "shortcuts" *e.g.* using the background or image artifacts to predict a given class.

However, explainability techniques come with their own issues which inhibits their use in practice. Notably the Disagreement Problem, which refers to inconsistencies between the existing saliency maps, makes it hard to derive insights from explanations. Looking at Figure 2, on the top row, the Occ/LIME/RISE/SHAP method highlight the wolf face, suggesting the model is relying on facial features to classify this wolf. An engineer who is only shown these saliency maps would be tempted to trust the model. However, other techniques like Arch and IG highlight the animal and the snowy background. What should an engineer conclude from these contradictory claims? Should the model be deployed? Given the current state of affairs, we encourage practitionners to always report a variety of saliency map techniques to avoid falling into confirmation bias.

Our work aims at reducing ambiguities between saliency map techniques by finding patches w.r.t which the model is almost additive so that Arch/Occ/IG/RISE/LIME/SHAP all agree on the importance of each patch. This can potentially increase the use of explainability techniques in real-world settings. Nevertheless, since a ConvNet is never perfectly patch-wise additive, the explanations yielded by AGREED are still imperfect and should still be treated with skepticism.

B Extended Background

B.1 Notation Table

Notation	Definition
Sets and Partitions	
d	Number of input features.
$[d] := \{1, 2, \dots, d\}$	Set of all features indices.
$j \in [d]$	j is an input feature index.
$u \subseteq [d]$	u is a subset of input features indices.
D	Number of features groups.
$[D] := \{1, 2, \dots, D\}$	Set of all groups indices.
$i \in [D]$	i is a group index.
$U \subseteq [D]$	U is a subset of groups indices.
$\mathcal{P} : [d] \to [D]$	Partition of $[d]$ into D disjoint groups.
$\mathcal{P}(u) := \{\mathcal{P}(j) : j \in u\}$	Partition mapping of feature subset u.
$\mathcal{P}^{-1}(U) := \{j \in [d] : \mathcal{P}(j) \in U\}$	Partition inverse map of the group subset U.
Explanations and Cooperative Game	
$\mathcal{X} \subseteq \mathbb{R}^d$	Input domain.
$\boldsymbol{x} \in \mathcal{X}$	Input to explain.
$\mathcal{D}$	Probability distribution for $\boldsymbol{x} \sim \mathcal{D}$.
$\boldsymbol{b} \in \mathcal{X}$	Baseline used as a reference.
$\mathcal{B}$	Probability distribution for $\boldsymbol{b} \sim \mathcal{B}$. Sometimes equal to $\mathcal{D}$.
$f : \mathcal{X} \to \mathbb{R}$	Model to explain.
$\boldsymbol{r}_u : \mathcal{X} \times \mathcal{X} \to \mathcal{X}$	Replace-function.
$f(\boldsymbol{x}) = \sum_{u \subseteq [d]} f_u(\boldsymbol{x})$	Functional Decomposition.
$f_{u,\mathcal{B}}(\boldsymbol{x})$	Marginal Decomposition.
$\boldsymbol{\phi}(h, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) \in \mathbb{R}^D$	Group Importance toward the gap $f(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{b} \sim \mathcal{B}}[f(\boldsymbol{b})]$.
$\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}$	Coalitional game.
$\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}$	Harsanyi Dividend.
$\mathcal{L}_f(\mathcal{D}, \mathcal{B}, \mathcal{P}) \in \mathbb{R}_+$	Grouped Lack of Additivity of f w.r.t $\mathcal{D}$, $\mathcal{B}$, and $\mathcal{P}$.

Table 1: All notation used throughout the paper.

B.2 ANOVA Decomposition

When the baseline distribution represents independent features (*i.e.* $\mathcal{B} = \mathcal{B}_{\text{ind}} := \prod_{i=1}^{d} \mathcal{B}_i$), the Marginal Decomposition (cf. Def 2.1) falls back to the ANOVA decomposition (Hooker, 2004). This functional decomposition enjoys additional theoretical properties that do not necessarily hold for the Marginal Decomposition. Notably, the components $f_{u,\mathcal{B}_{\text{ind}}}$ are zero mean and uncorrelated

$$u \neq \emptyset \Rightarrow \mathbb{E}_{\boldsymbol{x} \sim \mathcal{B}_{\text{ind}}}[f_{u,\mathcal{B}_{\text{ind}}}(\boldsymbol{x})] = 0. \tag{14}$$

$$u \neq v \Rightarrow \mathbb{E}_{\boldsymbol{x} \sim \mathcal{B}_{\text{ind}}}[f_{u,\mathcal{B}_{\text{ind}}}(\boldsymbol{x})\, f_{v,\mathcal{B}_{\text{ind}}}(\boldsymbol{x})] = 0. \tag{15}$$

Letting $\sigma_u^2 := \mathbb{E}_{\boldsymbol{x} \sim \mathcal{B}_{\text{ind}}}[f_{u,\mathcal{B}_{\text{ind}}}(\boldsymbol{x})^2]$, the total variance of the model can be decomposed

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{B}_{\text{ind}}}[\,(f(\boldsymbol{x}) - f_{\emptyset,\mathcal{B}_{\text{ind}}})^2] = \sum_{\substack{u \subseteq [d] \\ |u| \geq 1}} \sigma_u^2. \tag{16}$$

This property is where the terminology ANalysis Of VAriance comes from. Note that this elegant decomposition of model variance is not guaranteed to holds if features are non-independent. In that case, Equation 16 can potentially involve negative/positive correlation terms between f_u and f_v. Since feature independence is unlikely to hold on realistic Machine Learning datasets, the ANOVA decomposition will be of solely theoretical interest : it is used to derive Property 3 of Definition 3.1.

B.3 UNFAITHFULNESS METRICS

Many unfaithfulness metrics proposed in the literature can be framed as Equations 9 & 10 using a distinct weight $w(U)$ for every coalition $U \subseteq [D]$.

The Sensitivity-n metric proposed by Ancona et al. (2017) and the μ-Fidelity of Bhatt et al. (2020) report the $\overline{F}$ metric using the weight $w(U) = \binom{D}{|U|}^{-1}$ for $|U| = n$ and $w(U) = 0$ otherwise. The INFD metric of Yeh et al. (2019) also employs the $\overline{F}$ metric but it weights the various coalition uniformly: $w(U) = 1/2^D$.

The Shapley-weighted Fidelity (Muschalik et al., 2025) reports the $\underline{F}$ score using the weights $w(\emptyset) = w(D) = \infty$ and $w(U) \propto \binom{D-2}{|U|-1}^{-1}$. The infinite weights can be seen as a hard constraint that the attribution should respect the *efficiency axiom*: they should sum up to the model prediction (Tsai et al., 2023). Nevertheless, for the purpose of comparing different explanations quality, we will be ignoring the edge cases $U = \emptyset, D$ when computing this faithfulness score.

C PROOFS

C.1 HARSANYI DIVIDENDS

Before unifying the various post-hoc explainers proposed in the litterature, we must introduce an additional definition.

Definition C.1 (Harsanyi Dividends (Harsanyi, 1963))**.** *Given a coallitional game $\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}$ involving D players, its Harsanyi Dividend $\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}$ is defined recursively*

$$\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) := \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) - \sum_{V \subset U} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(V), \tag{17}$$

for any coallition $U \subseteq [D]$.

The base cases are $\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\emptyset) = \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\emptyset)$ and $\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\{i\}) = \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\{i\}) - \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\emptyset)$.

Harsanyi dividends $\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U)$ can be interpreted as the *excess gain* of a coallition $\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) - \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\emptyset)$ that cannot be explained by any cooperation between strict non-empty subsets $V \subset U$. We can reorganise Equation 17 to express any coalitional game in terms of its dividends

$$\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) = \sum_{V \subseteq U} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(V) \quad \text{for any } U \subseteq [D]. \tag{18}$$

Similarily, we can reexpress the definition of the Marginal Decomposition (cf. Definition 2.1)

$$\mathbb{E}_{\boldsymbol{b}\sim\mathcal{B}}[f(\boldsymbol{r}_u(\boldsymbol{b}, \boldsymbol{x}))] = \sum_{v \subseteq u} f_{v,\mathcal{B}}(\boldsymbol{x}) \quad \text{for any } u \subseteq [d]. \tag{19}$$

The ressemblance between Equations 18 and 19 is striking and suggests a deeper connection between the Marginal Decomposition and the Harsanyi Dividend. Our goal with this subsection is to highlight the link between these two related (but different) concepts.

Lemma C.1. *Given a partition $\mathcal{P} : [d] \to [D]$, the following holds*

$$u \subseteq \mathcal{P}^{-1}(U) \iff \mathcal{P}(u) \subseteq U, \tag{20}$$

where $u \subseteq [d]$ and $U \subseteq [D]$ are subsets of features and feature groups respectivelly.

Proof. We start from this simple consequence of the definitions of $\mathcal{P}$

$$j \in \mathcal{P}^{-1}(U) \iff \mathcal{P}(j) \in U. \tag{21}$$

The goal of the lemma is to translate this equivalence to feature subsets u and not just a single feature j. Letting $u \subseteq [d]$, we have

$$u \subseteq \mathcal{P}^{-1}(U) \iff \forall i \in u : i \in \mathcal{P}^{-1}(U) \iff \forall i \in u : \mathcal{P}(i) \in U \iff \mathcal{P}(u) \subseteq U. \tag{22}$$

□

Lemma C.2. *The Harsanyi Dividend $\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}$ can be expressed in terms of the Marginal Decomposition*

$$\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) = \sum_{u \subseteq [d] : \mathcal{P}(u) = U} f_{u,\mathcal{B}}(\boldsymbol{x}), \tag{23}$$

under the convention that $\mathcal{P}(\emptyset) = \emptyset$.

Proof. The proof proceeds by induction. The base case covers all dividends $\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U)$ such that $|U| \leq 1$. Indeed, we have:

$$\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\emptyset) := \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\emptyset) := \mathbb{E}_{\boldsymbol{b}\sim\mathcal{B}}[f(\boldsymbol{b})] = f_{\emptyset,\mathcal{B}}(\boldsymbol{x}). \tag{24}$$

and

$$\begin{aligned}
\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\{i\}) :=& \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\{i\}) - \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\emptyset) \\
:=& \mathbb{E}_{\boldsymbol{b}\sim\mathcal{B}}[f(\boldsymbol{r}_{\mathcal{P}^{-1}(\{i\})}(\boldsymbol{b},\boldsymbol{x}))] - \mathbb{E}_{\boldsymbol{b}\sim\mathcal{B}}[f(\boldsymbol{b})] && \text{(cf. Definition 2.3)} \\
=& \sum_{u\subseteq\mathcal{P}^{-1}(\{i\})} f_{u,\mathcal{B}}(\boldsymbol{x}) - f_{\emptyset,\mathcal{B}}(\boldsymbol{x}) && \text{(cf. Equation 19)} \\
=& \sum_{u\subseteq[d]:\mathcal{P}(u)\subseteq\{i\}} f_{u,\mathcal{B}}(\boldsymbol{x}) - f_{\emptyset,\mathcal{B}}(\boldsymbol{x}) && \text{(cf. Lemma C.1)} \\
=& \sum_{u\subseteq[d]:\mathcal{P}(u)=\{i\}} f_{u,\mathcal{B}}(\boldsymbol{x}).
\end{aligned}$$

Now fixing $U \subseteq [D]$ and assuming the premise holds for all $V \subset U$, we have

$$\begin{aligned}
\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) &= \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) - \sum_{V\subset U} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(V) && \text{(cf. Definition C.1)} \\
&= \mathbb{E}_{\boldsymbol{b}\sim\mathcal{B}}[f(\boldsymbol{r}_{\mathcal{P}^{-1}(U)}(\boldsymbol{b},\boldsymbol{x}))] - \sum_{V\subset U} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(V) && \text{(cf. Definition 2.3)} \\
&= \sum_{u\subseteq\mathcal{P}^{-1}(U)} f_{u,\mathcal{B}}(\boldsymbol{x}) - \sum_{V\subset U} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(V) && \text{(cf. Equation 19)} \\
&= \sum_{u\subseteq[d]:\mathcal{P}(u)\subseteq U} f_{u,\mathcal{B}}(\boldsymbol{x}) - \sum_{V\subset U} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(V) && \text{(cf. Lemma C.1)} \\
&= \sum_{u\subseteq[d]:\mathcal{P}(u)\subseteq U} f_{u,\mathcal{B}}(\boldsymbol{x}) - \sum_{V\subset U} \sum_{u\subseteq[d]:\mathcal{P}(u)=V} f_{u,\mathcal{B}}(\boldsymbol{x}) \\
& && \text{(By the recursion Assumption)} \\
&= \sum_{u\subseteq[d]:\mathcal{P}(u)\subseteq U} f_{u,\mathcal{B}}(\boldsymbol{x}) - \sum_{u\subseteq[d]:\mathcal{P}(u)\subset U} f_{u,\mathcal{B}}(\boldsymbol{x}) \\
&= \sum_{u\subseteq[d]:\mathcal{P}(u)=U} f_{u,\mathcal{B}}(\boldsymbol{x})
\end{aligned}$$

concluding the proof.

□

C.2 UNIFICATION

Theorem C.1 (Theorem 3.1). *The Arch/Occ/LIME/SHAP/RISE Attributions following can be expressed via the Marginal Decomposition*

$$\phi_i^{\mu}(\boldsymbol{x}, f, \mathcal{B}, \mathcal{P}) = \sum_{u \subseteq [d]: \{i\} = \mathcal{P}(u)} f_{u,\mathcal{B}}(\boldsymbol{x}). + \sum_{u \subseteq [d]: \{i\} \subsetneq \mathcal{P}(u)} h(|\mathcal{P}(u)|) f_{u,\mathcal{B}}(\boldsymbol{x}), \tag{25}$$

*where h is different for each μ. Different explainability method result disagree on how to redistribute **between-group interactions** $f_{u,\mathcal{B}}(\boldsymbol{x})$ with $\mathcal{P}(u) \geq 2$ among the groups involved.*

Proof. The proof of the theorem consists of expressing the attributions in terms of the Harsanyi Dividend $\Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}$ (cf. Definition C.1), 2) Use Lemma C.2 to express the Dividend in terms of the Marginal Decomposition.

ArchAttribute/Joint-PDP can be expressed in terms of the Harsanyi Dividend

$$\phi_i^{\text{Arch}}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) = \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\{i\}). \tag{26}$$

Following Lemma C.2 its holds that

$$\phi_i^{\text{Arch}}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) = \sum_{u \subseteq [d]: \mathcal{P}(u) = \{i\}} f_{u,\mathcal{B}}(\boldsymbol{x}). \tag{27}$$

Patch-Occlusion/joint-PFI is also easily expressed in terms of Harsanyi Dividends

$$\begin{aligned}
\phi_i^{\text{Occ}}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) &= \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}([D]) - \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}([D] \setminus \{i\}) \\
&= \sum_{U \subseteq [D]} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) - \sum_{U \subseteq [D] \setminus \{i\}} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) && \text{(cf. Equation 18)} \\
&= \sum_{U \subseteq [D]: i \in U} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) \\
&= \sum_{U \subseteq [D]: i \in U} \sum_{u \subseteq [d]: \mathcal{P}(u) = U} f_{u,\mathcal{B}}(\boldsymbol{x}) && \text{(cf. Lemma C.2)} \\
&= \sum_{u \subseteq [d]: i \in \mathcal{P}(u)} f_{u,\mathcal{B}}(\boldsymbol{x}).
\end{aligned}$$

Shapley Values are known to redistribute the Harsanyi Dividends evenly between all players involved (Shapley, 1953)

$$\begin{aligned}
\phi_i^{\text{SHAP}}(f, x, \mathcal{B}, \mathcal{P}) &= \sum_{U \subseteq [D]: i \in U} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U)/|U| && \text{((Shapley, 1953))} \\
&= \sum_{U \subseteq [D]: i \in U} |U|^{-1} \sum_{u \subseteq [d]: \mathcal{P}(u) = U} f_{u,\mathcal{B}}(\boldsymbol{x}) && \text{(cf. Lemma C.2)} \\
&= \sum_{U \subseteq [D]: i \in U} \sum_{u \subseteq [d]: \mathcal{P}(u) = U} f_{u,\mathcal{B}}(\boldsymbol{x})/|\mathcal{P}(u)| && \text{(Since } |U| = |\mathcal{P}(u)|\text{)} \\
&= \sum_{u \subseteq [d]: i \in \mathcal{P}(u)} f_{u,\mathcal{B}}(\boldsymbol{x})/|\mathcal{P}(u)|.
\end{aligned}$$

RISE computes

$$\phi_i^{\text{RISE}}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) := \frac{1}{2^{D-1}} \sum_{U \subseteq [D] \setminus \{i\}} \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U \cup \{i\}) - \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U), \tag{28}$$

which is actually the Banzhaf Index (Marichal et al., 2007). It is well-established that this attribution method assigns a score to each player by sharing Harsanyi Dividends using a power-of-two rule

$$\begin{aligned}
\phi_i^{\text{RISE}}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) &= \sum_{U \subseteq [D]: i \in U} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U)/2^{|U|-1} && \text{(See Page 8 from (Marichal et al., 2007))} \\
&= \sum_{U \subseteq [D]: i \in U} 1/2^{|U|-1} \sum_{u \subseteq [d]: \mathcal{P}(u) = U} f_{u,\mathcal{B}}(\boldsymbol{x}) && \text{(cf. Lemma C.2)} \\
&= \sum_{U \subseteq [D]: i \in U} \sum_{u \subseteq [d]: \mathcal{P}(u) = U} f_{u,\mathcal{B}}(\boldsymbol{x})/2^{|\mathcal{P}(u)|-1} && \text{(Since } |U| = |\mathcal{P}(u)|) \\
&= \sum_{u \subseteq [d]: i \in \mathcal{P}(u)} f_{u,\mathcal{B}}(\boldsymbol{x})/2^{|\mathcal{P}(u)|-1}.
\end{aligned}$$

LIME advocates fitting a linear model on the function output evaluated on masked inputs

$$(\omega_0, \omega_1, \omega_2, \ldots, \omega_D) = \underset{\boldsymbol{\omega} \in \mathbb{R}^{D+1}}{\arg\min} \frac{1}{2^D} \sum_{U \subseteq [D]} \left(\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) - \omega_0 - \sum_{i=1}^{D} \omega_i \mathbb{1}[i \in U]\right)^2. \tag{29}$$

and reporting the coefficients $(w_1, w_2, \ldots, w_D)$ as the local feature-groups attributions. This minimization problem is an alternative formulation of the Banzhaf Index (Tsai et al., 2023) so LIME is equivalent to RISE.

We have thus proven that Arch/Occ/LIME/RISE/SHAP can be expressed as

$$\phi_i(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) = \sum_{u \subseteq [d]: \mathcal{P}(u) = \{i\}} f_{u,\mathcal{B}}(\boldsymbol{x}) + \sum_{u \subseteq [d]: \{i\} \subsetneq \mathcal{P}(u)} h(|\mathcal{P}(u)|) f_{u,\mathcal{B}}(\boldsymbol{x}), \tag{30}$$

where $h(|\mathcal{P}(u)|) = 0$ for ArchAttribute, $h(|\mathcal{P}(u)|) = 1$ for Occlusion, $h(|\mathcal{P}(u)|) = 1/|\mathcal{P}(u)|$ for SHAP, and $h(|\mathcal{P}(u)|) = 1/2^{|\mathcal{P}(u)|-1}$ for RISE/LIME.

□

Before highlighting the critical role of between-group interactions in the unfaithfulness metrics, we first recall the Minimality property of the Marginal Decomposition.

Corollary C.2 (**Corollary A.1** from (Laberge et al., 2024))**.** *Let $R \subseteq \mathbb{R}^d$ be a hyperrectangle and let $f : \mathbb{R}^d \to \mathbb{R}$ be a function that can be written $f(\boldsymbol{x}) = \sum_{u \subseteq [d]} g_u(\boldsymbol{x})$ $\forall \boldsymbol{x} \in R$, where g_u only depends on $\boldsymbol{x}_u$. Also, assume that a subset $v \subset [d]$ exists such that*

$$u \supseteq v \Rightarrow \forall \boldsymbol{x} \in R \;\; g_u(\boldsymbol{x}) = 0.$$

Then, for any probability distribution $\mathcal{B}$ such that supp$(\mathcal{B}) \subseteq R$, the Marginal Decomposition respects

$$u \supseteq v \Rightarrow \forall \boldsymbol{x} \in R \;\; f_{u,\mathcal{B}}(\boldsymbol{x}) = 0.$$

Minimality implies that the Marginal Decomposition will not contain interactions that are not present in the model in the first place. This theorem induces an important corollary.

Lemma C.3. *Let R be a hyperrectangle region such that supp$(\mathcal{B}) \subseteq R$. If f is groupwise additive in R w.r.t partition $\mathcal{P}$, then it holds that*

$$|\mathcal{P}(v)| \geq 2, \boldsymbol{x} \in R \Rightarrow f_{v,\mathcal{B}}(\boldsymbol{x}) = 0.$$

Proof. Let f be groupwise additive in R w.r.t $\mathcal{P}$. Thus there exists a rectangular region R such that supp$(\mathcal{B}) \subseteq R$, and f can be written as

$$f(\boldsymbol{x}) = \omega_0 + \sum_{i=1}^{D} g_{\mathcal{P}^{-1}(\{i\})}(\boldsymbol{x}) \;\; \forall \boldsymbol{x} \in R.$$

Thus, f can be written in the form $\sum_{u \subseteq [d]} g_u$ where $|\mathcal{P}(u)| \geq 2 \Rightarrow g_u = 0$.

Now, letting $\boldsymbol{x} \in R$ be an input and $v \subseteq [d]$ be any feature subset such that $|\mathcal{P}(v)| \geq 2$. Any superset $u \supseteq v$ respects $|\mathcal{P}(u)| \geq |\mathcal{P}(v)| \geq 2$ and so $u \supseteq v \Rightarrow g_u(\boldsymbol{x}) = 0$. By minimality (cf. Theorem C.2), this implies that $f_{v,\mathcal{B}}(\boldsymbol{x}) = 0$ and so

$$|\mathcal{P}(v)| \geq 2, \boldsymbol{x} \in R \Rightarrow f_{v,\mathcal{B}}(\boldsymbol{x}) = 0.$$

□

With this Corollary now proven, we can demonstrate how group-wise additive models have faithful explanations according to the various unfaithfulness metrics.

Theorem C.3 (Theorem 3.2). *Let $R \subseteq \mathcal{X}$ be a hyperrectangle region such that supp$(\mathcal{D}) \subseteq R$ and supp$(\mathcal{B}) \subseteq R$. Let $\mathcal{P}$ be a feature partition. Whenever the model f is groupwise additive in R w.r.t $\mathcal{P}$, any unfaithfulness metrics that follow Equations 9 and 10 are all simultaneously minimized*

$$F(\boldsymbol{\phi}^{\mu}, f, w) = 0 \tag{31}$$

for any weight function w and attribution $\boldsymbol{\phi}^{\mu}$.

Proof. Fix the set $U \subseteq [D]$. Unfaithfulness metrics $\overline{F}$ and $\underline{F}$ are simply weighted aggregates of the difference between $\sum_{i \in U} \phi_i^{\mu}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P})$ and either $\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) - \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\emptyset)$ or $\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}([D]) - \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}([D] \setminus U)$. By Theorem 3.1 and Lemma C.3, the first term is equal to

$$\begin{aligned}
\sum_{i \in U} \phi_i^{\mu}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) &= \sum_{i \in U} \sum_{u \subseteq [d]: \{i\} = \mathcal{P}(u)} f_{u,\mathcal{B}}(\boldsymbol{x}) \\
&= \sum_{u \subseteq [d]: \mathcal{P}(u) \subseteq U, |\mathcal{P}(u)| = 1} f_{u,\mathcal{B}}(\boldsymbol{x}).
\end{aligned}$$

By Corollary C.2, the other two terms are also equal to this quantity

$$\begin{aligned}
\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(U) - \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\emptyset) &= \sum_{V \subseteq U} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(V) - \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(\emptyset) && \text{(cf. Eq 18)} \\
&= \sum_{V \subseteq U: V \neq \emptyset} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(V) \\
&= \sum_{V \subseteq U: V \neq \emptyset} \sum_{u \subseteq [d]: \mathcal{P}(u) = V} f_{u,\mathcal{B}}(\boldsymbol{x}) && \text{(cf. Lemma C.2)} \\
&= \sum_{u \subseteq [d]: \mathcal{P}(u) \subseteq U, \mathcal{P}(u) \neq \emptyset} f_{u,\mathcal{B}}(\boldsymbol{x}) \\
&= \sum_{u \subseteq [d]: \mathcal{P}(u) \subseteq U, |\mathcal{P}(u)| = 1} f_{u,\mathcal{B}}(\boldsymbol{x}) && \text{(cf. Lemma C.3)} \\
\nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}([D]) - \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}([D] \setminus U) &= \sum_{V \subseteq [D]} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(V) - \sum_{V \subseteq [D] \setminus U} \nu_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(V) && \text{(cf. Eq 18)} \\
&= \sum_{V \subseteq U: V \cap U \neq \emptyset} \Delta_{f,\boldsymbol{x},\mathcal{B},\mathcal{P}}(V) \\
&= \sum_{V \subseteq U: V \cap U \neq \emptyset} \sum_{u \subseteq [d]: \mathcal{P}(u) = V} f_{u,\mathcal{B}}(\boldsymbol{x}) && \text{(cf. Lemma C.2)} \\
&= \sum_{u \subseteq [d]: \mathcal{P}(u) \cap U \neq \emptyset} f_{u,\mathcal{B}}(\boldsymbol{x}) \\
&= \sum_{u \subseteq [d]: \mathcal{P}(u) \cap U \neq \emptyset, |\mathcal{P}(u)| = 1} f_{u,\mathcal{B}}(\boldsymbol{x}) && \text{(cf. Lemma C.2)} \\
&= \sum_{u \subseteq [d]: \mathcal{P}(u) \subseteq U, |\mathcal{P}(u)| = 1} f_{u,\mathcal{B}}(\boldsymbol{x}).
\end{aligned}$$

□

C.3 GROUPED LACK OF ADDITIVITY

In this section, we demonstrate the various properties of GLoA loss functions, which will be used to infer feature partitions. We start by presentng a link between the cardinalities $|\mathcal{P}(u)|$ and $|\mathcal{P}'(u)|$ when $\mathcal{P}'$ is a superpartition of $\mathcal{P}$

Lemma C.4. *Let $\mathcal{P}'$ be a superpartition of $\mathcal{P}$, then for any subset $u \subseteq [d]$ is holds that*

$$|\mathcal{P}(u)| \geq |\mathcal{P}'(u)|. \tag{32}$$

Proof. Recall the definition of superpartition: $\mathcal{P}'$ is a superpartition of $\mathcal{P}$ if $\mathcal{P}(i) = \mathcal{P}(j) \Rightarrow \mathcal{P}'(i) = \mathcal{P}'(j)$. Conversly, $\mathcal{P}'(i) \neq \mathcal{P}'(j) \Rightarrow \mathcal{P}(i) \neq \mathcal{P}(j)$ must hold. This implies that $|\mathcal{P}(u)| \geq |\mathcal{P}'(u)|$ for any $u \subseteq [d]$. To prove it, assume the opposite holds: $|\mathcal{P}(u)| < |\mathcal{P}'(u)|$. This implies the existence of $|\mathcal{P}'(u)|$ points $i, j \in u$ such that $\mathcal{P}'(i) \neq \mathcal{P}'(j)$. However, by definition of superpartition, $\mathcal{P}(i) \neq \mathcal{P}(j)$ must also hold for these $|\mathcal{P}'(u)|$ points which contradicts the assumption that $\mathcal{P}(u)$ has smaller cardinality than $\mathcal{P}'(u)$. □

Lemma C.5. *Define the objective*

$$\mathcal{L}_f(\mathcal{D}, \mathcal{B}, \mathcal{P}) := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\Bigg[\sum_{\substack{u,v\subseteq[d] \\ |\mathcal{P}(u)|\geq 2, |\mathcal{P}(v)|\geq 2}} w(\mathcal{P}(u), \mathcal{P}(v)) f_{u,\mathcal{B}}(\boldsymbol{x}) f_{v,\mathcal{B}}(\boldsymbol{x}) \Bigg] \tag{33}$$

for some function w such that $w(\mathcal{P}(u), \mathcal{P}(u)) \geq w(\mathcal{P}'(u), \mathcal{P}'(u)) \geq 0$ for any interaction u. Then $\mathcal{L}_f$ respects Definition 3.1.

Proof. We prove the function $\mathcal{L}_f$ from Equation 33 respects the three properties of **Definition 3.1**.

Property 1 Let f be group-wise additive w.r.t $\mathcal{P}$, $\mathcal{D}$, and $\mathcal{B}$. By Lemma C.3, there exists a rectangular region R such that $\text{supp}(\mathcal{B}) \subseteq R$, $\text{supp}(\mathcal{D}) \subseteq R$, and f can be written as

$$|\mathcal{P}(v)| \geq 2, \boldsymbol{x} \in R \Rightarrow f_{v,\mathcal{B}}(\boldsymbol{x}) = 0.$$

Sampled inputs $\boldsymbol{x} \sim \mathcal{D}$ are guaranteed to land in R (since $\text{supp}(\mathcal{D}) \subseteq R$) and so

$$\mathcal{L}_f(\mathcal{D}, \mathcal{B}, \mathcal{P}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\Bigg[\sum_{\substack{u,v\subseteq[d] \\ |\mathcal{P}(u)|\geq 2, |\mathcal{P}(v)|\geq 2}} w(\mathcal{P}(u), \mathcal{P}(v)) f_{u,\mathcal{B}}(\boldsymbol{x}) f_{v,\mathcal{B}}(\boldsymbol{x}) \Bigg] = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[0] = 0.$$

Property 2 Given a partition $\mathcal{P}$ of $[d]$ into D groups, assume w.l.o.g that f is additive w.r.t group D. Also, assume we wish to fuse group D with group $D-1$, which will lead to a super partition $\mathcal{P}'$ such that

$$\mathcal{P}'(i) = \mathcal{P}(i) \quad \forall i \in \mathcal{P}^{-1}(\{1, 2, \ldots, D-1\}), \tag{34}$$

but $\mathcal{P}'(i) = D - 1 \quad \forall i \in \mathcal{P}^{-1}(\{D\})$. The GLoA under partition $\mathcal{P}$ can be written

$$\begin{aligned}
\mathcal{L}_f(\mathcal{D}, \mathcal{B}, \mathcal{P}) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\Bigg[\sum_{\substack{u,v\subseteq[d] \\ |\mathcal{P}(u)|\geq 2, |\mathcal{P}(v)|\geq 2}} w(\mathcal{P}(u), \mathcal{P}(v)) f_{u,\mathcal{B}}(\boldsymbol{x}) f_{v,\mathcal{B}}(\boldsymbol{x}) \Bigg] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\Bigg[\sum_{\substack{u,v\subseteq[d] \\ |\mathcal{P}(u)|\geq 2, |\mathcal{P}(v)|\geq 2 \\ D\in\mathcal{P}(u) \text{ or } D\in\mathcal{P}(v)}} w(\mathcal{P}(u), \mathcal{P}(v)) f_{u,\mathcal{B}}(\boldsymbol{x}) f_{v,\mathcal{B}}(\boldsymbol{x}) \\
&\qquad + \sum_{\substack{u,v\subseteq[d] \\ |\mathcal{P}(u)|\geq 2, |\mathcal{P}(v)|\geq 2 \\ D\notin\mathcal{P}(u) \text{ and } D\notin\mathcal{P}(v)}} w(\mathcal{P}(u), \mathcal{P}(v)) f_{u,\mathcal{B}}(\boldsymbol{x}) f_{v,\mathcal{B}}(\boldsymbol{x}) \Bigg].
\end{aligned} \tag{35}$$

We prove that $\mathcal{L}_f(\mathcal{D}, \mathcal{B}, \mathcal{P}) = \mathcal{L}_f(\mathcal{D}, \mathcal{B}, \mathcal{P}')$ by rewritting both summation terms of Equation 35. For the first term, we exploit the fact that f is additive w.r.t group D, implying the existence of a rectangular region R such that supp$(\mathcal{B}) \subseteq R$, supp$(\mathcal{D}) \subseteq R$, and

$$f(\boldsymbol{x}) = g_{\mathcal{P}^{-1}(\{D\})}(\boldsymbol{x}) + g_{\mathcal{P}^{-1}(\{1,2\ldots,D-1\})}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in R.$$

Hence, f can be written in the form $\sum_{u \subseteq [d]} g_u$ where $[|\mathcal{P}(u)| \geq 2 \text{ and } D \in \mathcal{P}(u)] \Rightarrow g_u = 0$.

Assuming $\boldsymbol{x} \in R$ and $v \subseteq [d]$ is some feature subset such that $|\mathcal{P}(v)| \geq 2$ and $D \in \mathcal{P}(v)$. Any superset $u \supseteq v$ respects $|\mathcal{P}(u)| \geq |\mathcal{P}(v)| \geq 2$ and $D \in \mathcal{P}(u)$, thus $u \supseteq v \Rightarrow g_u(\boldsymbol{x}) = 0$. By minimality (cf. Theorem C.2), the component $f_{v,\mathcal{B}}(\boldsymbol{x}) = 0$ is null and so

$$|\mathcal{P}(v)| \geq 2, D \in \mathcal{P}(v), \boldsymbol{x} \in R \Rightarrow f_{v,\mathcal{B}}(\boldsymbol{x}) = 0. \tag{36}$$

Accordingly, the first summation term of Equation 35 is null

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}}\Bigg[\sum_{\substack{u,v \subseteq [d] \\ |\mathcal{P}(u)| \geq 2, |\mathcal{P}(v)| \geq 2 \\ D \in \mathcal{P}(u) \text{ or } D \in \mathcal{P}(v)}} w(\mathcal{P}(u), \mathcal{P}(v)) f_{u,\mathcal{B}}(\boldsymbol{x}) f_{v,\mathcal{B}}(\boldsymbol{x}) \Bigg] = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}}[0] = 0. \tag{37}$$

By Lemma C.4, we also have that $|\mathcal{P}'(u)| \geq 2$ implies $|\mathcal{P}(u)| \geq 2$ and so

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}}\Bigg[\sum_{\substack{u,v \subseteq [d] \\ |\mathcal{P}'(u)| \geq 2, |\mathcal{P}'(v)| \geq 2 \\ D \in \mathcal{P}(u) \text{ or } D \in \mathcal{P}(v)}} w(\mathcal{P}'(u), \mathcal{P}'(v)) f_{u,\mathcal{B}}(\boldsymbol{x}) f_{v,\mathcal{B}}(\boldsymbol{x}) \Bigg] = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}}[0] = 0. \tag{38}$$

As a result, the left-most terms of Equations 37 and 38 are equal.

Now tackling the second term of Equation 35. By Equation 34, for any $u \subseteq [d]$ such that $D \notin \mathcal{P}(u)$ it holds that $\mathcal{P}(u) = \mathcal{P}'(u)$. So for any $\boldsymbol{x}$ we have

$$\sum_{\substack{u,v \subseteq [d] \\ |\mathcal{P}(u)| \geq 2, |\mathcal{P}(v)| \geq 2 \\ D \notin \mathcal{P}(u) \text{ and } D \notin \mathcal{P}(v)}} w(\mathcal{P}(u), \mathcal{P}(v)) f_{u,\mathcal{B}}(\boldsymbol{x}) f_{v,\mathcal{B}}(\boldsymbol{x}) = \sum_{\substack{u,v \subseteq [d] \\ |\mathcal{P}'(u)| \geq 2, |\mathcal{P}'(v)| \geq 2 \\ D \notin \mathcal{P}(u) \text{ and } D \notin \mathcal{P}(v)}} w(\mathcal{P}'(u), \mathcal{P}'(v)) f_{u,\mathcal{B}}(\boldsymbol{x}) f_{v,\mathcal{B}}(\boldsymbol{x}). \tag{39}$$

Thus we have proven that $\mathcal{L}_f(\mathcal{D}, \mathcal{B}, \mathcal{P}) = \mathcal{L}_f(\mathcal{D}, \mathcal{B}, \mathcal{P}')$.

Property 3 When features are independent, the Marginal Decomposition falls back to the ANOVA decomposition (see Appendix B.2). In this case, the functional components become zero-mean, uncorrelated, and have variance σ_u^2. Using Equation 15 and setting $\mathcal{B}_{\text{ind}} = \mathcal{D}$

$$\mathcal{L}_f(\mathcal{B}_{\text{ind}}, \mathcal{B}_{\text{ind}}, \mathcal{P}) = \sum_{u \subseteq [d]: |\mathcal{P}(u)| \geq 2} w(\mathcal{P}(u), \mathcal{P}(u))\, \sigma_u^2 \tag{40}$$

and

$$\mathcal{L}_f(\mathcal{B}_{\text{ind}}, \mathcal{B}_{\text{ind}}, \mathcal{P}') = \sum_{u \subseteq [d]: |\mathcal{P}'(u)| \geq 2} w(\mathcal{P}'(u), \mathcal{P}'(u))\, \sigma_u^2.$$

We define the sets

$$S_{\mathcal{P}} = \{u \subseteq [d] : |\mathcal{P}(u)| \geq 2\} \text{ and } S_{\mathcal{P}'} = \{u \subseteq [d] : |\mathcal{P}'(u)| \geq 2\}.$$

By Lemma C.4, we have that $S_{\mathcal{P}'} \subseteq S_{\mathcal{P}}$ and so

$$\begin{aligned}
\mathcal{L}_f(\mathcal{B}_{\text{ind}}, \mathcal{B}_{\text{ind}}, \mathcal{P}) - \mathcal{L}_f(\mathcal{B}_{\text{ind}}, \mathcal{B}_{\text{ind}}, \mathcal{P}') &= \sum_{u \in S_{\mathcal{P}}} w(\mathcal{P}(u), \mathcal{P}(u))\, \sigma_u^2 - \sum_{u \in S_{\mathcal{P}'}} w(\mathcal{P}'(u), \mathcal{P}'(u))\, \sigma_u^2 \\
&= \sum_{u \in S_{\mathcal{P}} \setminus S_{\mathcal{P}'}} w(\mathcal{P}(u), \mathcal{P}(u))\, \sigma_u^2 \\
&\quad + \sum_{u \in S_{\mathcal{P}'}} \left[w(\mathcal{P}(u), \mathcal{P}(u)) - w(\mathcal{P}'(u), \mathcal{P}'(u)) \right] \sigma_u^2 \\
&\geq 0 \qquad \text{(Since } w(\mathcal{P}(u), \mathcal{P}(u)) \geq w(\mathcal{P}'(u), \mathcal{P}'(u)) \geq 0)
\end{aligned}$$

□

Theorem C.4 (Theorem 3.3). *The L_2 disagreements $D_{L_2}(\boldsymbol{\phi}^{Occ}, \boldsymbol{\phi}')$ between Occlusion and the Arch/LIME/RISE/SHAP explainers respect Definition 3.1. Proof in Appendix C.3*

Proof. We must prove that the L_2 disagreements between the post-hoc explanation methods respect the premise of Lemma C.5.

$$\begin{aligned}
D_{L_2}(\boldsymbol{\phi}, \boldsymbol{\phi}') &= \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \left[\sum_{i=1}^{D} \left(\sum_{u \subseteq [d]: i \in \mathcal{P}(u), |\mathcal{P}(u)| \geq 2} \underbrace{\left(h(|\mathcal{P}(u)|) - h'(|\mathcal{P}(u)|) \right)}_{b(|\mathcal{P}(u)|)} f_{u,\mathcal{B}}(\boldsymbol{x}) \right)^2 \right] \\
&= \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \left[\sum_{i=1}^{D} \sum_{\substack{u,v \subseteq [d] \\ i \in \mathcal{P}(u), |\mathcal{P}(u)| \geq 2 \\ i \in \mathcal{P}(v), |\mathcal{P}(v)| \geq 2}} b(|\mathcal{P}(u)|) b(|\mathcal{P}(v)|) f_{u,\mathcal{B}}(\boldsymbol{x}) f_{v,\mathcal{B}}(\boldsymbol{x}) \right] \\
&= \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \left[\sum_{u,v \subseteq [d]: |\mathcal{P}(u)| \geq 2, |\mathcal{P}(v)| \geq 2} |\mathcal{P}(u) \cap \mathcal{P}(v)|\, b(|\mathcal{P}(u)|) b(|\mathcal{P}(v)|) f_{u,\mathcal{B}}(\boldsymbol{x})\, f_{v,\mathcal{B}}(\boldsymbol{x}) \right].
\end{aligned}$$

The corresponding interaction penalization is $w(\mathcal{P}(u), \mathcal{P}(v)) := |\mathcal{P}(u) \cap \mathcal{P}(v)|\, b(|\mathcal{P}(u)|) b(|\mathcal{P}(v)|)$. Now, obviously $w(\mathcal{P}(u), \mathcal{P}(u)) = |\mathcal{P}(u)| b(|\mathcal{P}(u)|)^2 \geq 0$ but $w(\mathcal{P}(u), \mathcal{P}(u)) \geq w(\mathcal{P}'(u), \mathcal{P}'(u))$ only holds for any superpartition $\mathcal{P}'$ if we compare certain pairs of explainers : Occ-Arch, Occ-SHAP, Occ-LIME, Occ-RISE. □

D AGREED

The AGREED algorithm aims at minimizing the L_2 disagreements between the joint-PDP/ArchAttribute and joint-PFI/Occlusion explanations

$$\mathcal{L}_f^{\text{AGREED}}(\mathcal{D}, \mathcal{B}, \mathcal{P}) := \sum_{i=1}^{D} \Psi(i) \quad \text{with} \quad \Psi(i) := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\big[(\phi_i^{\text{Arch}}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}) - \phi_i^{\text{Occ}}(f, \boldsymbol{x}, \mathcal{B}, \mathcal{P}))^2\big]. \tag{41}$$

Starting from a granular partition $\{\{1\}, \ldots, \{d\}\}$, we greedily minimize Equation 41

1. Select the group i with highest potential $\Psi(i)$.
2. Compute its pair-wise interaction with other groups j.
3. Fuse group i with the group j of maximal pairwise interaction.
4. Repeat until the disagreements fall below ϵ.

The implementation details of the algorithm depend on whether $\mathcal{B} = \mathcal{D}$ or not.

D.1 Cases where $\mathcal{B} = \mathcal{D}$

It is common to set $\mathcal{B} = \mathcal{D}$ on Tabular data so that $\boldsymbol{x}$ and $\boldsymbol{b}$ can both be interpreted as a random sample from the dataset. This assumption introduces a symmetry between $\boldsymbol{x}$ and $\boldsymbol{b}$ that makes statistical estimates more efficient.

Step 1. Computing the Group Potential The building blocks of AGREED are the following $D \times N \times N$ tensors

Definition D.1. *Let $\mathcal{P}$ be a partition of $[d]$ into D disjoint groups, $\mathcal{D}$ be the data distribution and $\{\boldsymbol{x}^{(k)}\}_{k=1}^N$ be N points sampled from it. We define the $D \times N \times N$ tensor $\boldsymbol{G}$*

$$G_{i,k,\ell} = f(\boldsymbol{r}_{\mathcal{P}^{-1}(\{i\})}(\boldsymbol{x}^{(\ell)}, \boldsymbol{x}^{(k)})) - f(\boldsymbol{x}^{(\ell)}). \tag{42}$$

These tensors are of interest because averaging them along their second and third axis leads to consistent estimate of the joint-PDP/ArchAttribute and joint-PFI/Occlusion respectively

$$\frac{1}{N}\sum_{\ell=1}^{N} G_{i,k,\ell} \xrightarrow{p} \phi_i^{\text{Arch}}(f, \boldsymbol{x}^{(k)}, \mathcal{D}, \mathcal{P}) \tag{43}$$

$$-\frac{1}{N}\sum_{k=1}^{N} G_{i,k,\ell} \xrightarrow{p} \phi_i^{\text{Occ}}(f, \boldsymbol{x}^{(\ell)}, \mathcal{D}, \mathcal{P}). \tag{44}$$

Given $\boldsymbol{G}$, we can efficiently compute the potential of group i

$$\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{N}\sum_{\ell=1}^{N}(\, G_{i,k,\ell} + G_{i,\ell,k}\,)\right)^2 \xrightarrow{p} \Psi(i). \tag{45}$$

2. Computing Interaction Between Groups At any point in the iterative algorithm, we will have access to the $\boldsymbol{G}$ tensor of the current partition. Thus, a good between-group interaction score should leverage this precomputed tensor to avoid unnecessary model inferences.

Definition D.2. *Let $\mathcal{P}$ be a partition of $[d]$ into D disjoint groups, $i \in [D]$ be a group that we want to extend, $\mathcal{D}$ be the data distribution and $\{\boldsymbol{x}^{(k)}\}_{k=1}^N$ be N points sampled from it. We define the $(D-1) \times N \times N$ matrix $\boldsymbol{I}$ such that*

$$I_{j,k,\ell} := f(\boldsymbol{r}_{\mathcal{P}^{-1}(\{i,j\})}(\boldsymbol{x}^{(\ell)}, \boldsymbol{x}^{(k)})) - f(\boldsymbol{r}_{\mathcal{P}^{-1}(\{i\})}(\boldsymbol{x}^{(\ell)}, \boldsymbol{x}^{(k)})) - f(\boldsymbol{r}_{\mathcal{P}^{-1}(\{j\})}(\boldsymbol{x}^{(\ell)}, \boldsymbol{x}^{(k)})) + f(\boldsymbol{x}^{(\ell)}). \tag{46}$$

Crucially, $\boldsymbol{I}$ can be computed efficiently by querying $\boldsymbol{G}$

$$I_{j,k,\ell} = f(\boldsymbol{r}_{\mathcal{P}^{-1}(\{i,j\})}(\boldsymbol{x}^{(\ell)}, \boldsymbol{x}^{(k)})) - f(\boldsymbol{x}^{(\ell)}) - G_{i,k,\ell} - G_{j,k,\ell}. \tag{47}$$

Averaging this tensor along the third axis yields

$$\frac{1}{N}\sum_{\ell=1}^{N} I_{j,k,\ell} \xrightarrow{p} \sum_{u\subseteq[d]:\mathcal{P}(u)=\{i,j\}} f_{u,\mathcal{D}}(\boldsymbol{x}^{(k)}), \tag{48}$$

a pure measure of interactions only involving features from groups i and j. Averaging the tensor along its second axis leads to

$$\frac{1}{N}\sum_{k=1}^{N} I_{j,k,\ell} \xrightarrow{p} \sum_{u\subseteq[d]:\{i,j\}\subseteq\mathcal{P}(u)} f_{u,\mathcal{D}}(\boldsymbol{x}^{(\ell)}), \tag{49}$$

a full measure of interactions involving features from groups i, j, and possibly other groups. To compute the strength of the interaction between two groups i and j, we could report the pure interaction, the full interaction, or a weighted average. Like (Tsang et al., 2020), we average the pure and full interactions with weights 0.5.

3. Fusing Groups After having identified two groups i and j that interact, we define a superpartition $\mathcal{P}'$ of size $D-1$ where i, j are fused into a single group. The $D-2$ groups $k \neq i, j$ are simply re-indexed from 1 to $D-2$, while i and j are considered the $(D-1)^{\text{th}}$ group. Since the $\boldsymbol{G}$ tensor is relative to the current partition, it must be updated to $\boldsymbol{G}'$ when performing group fusion. For the $D-2$ groups that were not fused, we copy their $G_{k,:,:}$ values. For the 2 groups that were fused, we store the joint effect

$$G'_{D-1,k,\ell} := f(\boldsymbol{r}_{\mathcal{P}^{-1}(\{i,j\})}(\boldsymbol{x}^{(\ell)}, \boldsymbol{x}^{(k)})) - f(\boldsymbol{x}^{(\ell)}) = I_{j,k,\ell} + G_{i,k,\ell} + G_{j,k,\ell}, \tag{50}$$

that is computed without additional model inference. Here is the pseudocode for updating the partition.

Algorithm 1 Update Partition $\mathcal{P}$ by fusing $i, j \in [D]$ into a new group

procedure UPDATE_PARTITION($\mathcal{P}, \boldsymbol{G}, \boldsymbol{I}, i, j$)
  Initialize new partition map $\mathcal{P}'$;
  $G' \leftarrow \texttt{zeros}(D-1, N, N)$;
  % Counter Variable
  $c \leftarrow 1$
  for $k \in [D]$ such that $k \neq i, j$ **do**
    Define $\mathcal{P}'(\ell) := c, \quad \forall \ell \in \mathcal{P}^{-1}(\{k\})$;
    $G'_{c,:,:} \leftarrow G_{k,:,:}$;
    $c \leftarrow c + 1$;
  % Groups i and j are fused
  Define $\mathcal{P}'(\ell) := D-1, \quad \forall \ell \in \mathcal{P}^{-1}(\{i, j\})$;
  $G'_{D-1,:,:} \leftarrow I_{j,:,:} + G_{i,:,:} + G_{j,:,:}$;
  return $\mathcal{P}'$, $\boldsymbol{G}'$

This partition update requires no model inference since all relevant computations are queried from the $\boldsymbol{G}$ and $\boldsymbol{I}$ tensors. Putting everything togheter, we end up with Algorithm 2. AGREED requires $\mathcal{O}(d^2N^2)$ model inferences since line 5 calls f dN^2 times, then for each iteration of the while loop (of which there are at most $d-1$), $\boldsymbol{I}$ is computed which does not call f more than dN^2 times.

D.2 Cases where $\mathcal{B} \neq \mathcal{D}$

When computing saliency maps for Image Classification, it is common to use a single baseline $\mathcal{B} = \delta_{\boldsymbol{b}}$ (typically an image with no information). Also, there is no need to find a partition $\mathcal{P}$ that works across all images $\boldsymbol{x}$ from the dataset. Finding a partition that works on a single image $\boldsymbol{x}$ is a more realistic goal so we set $\mathcal{D} = \delta_{\boldsymbol{x}}$.

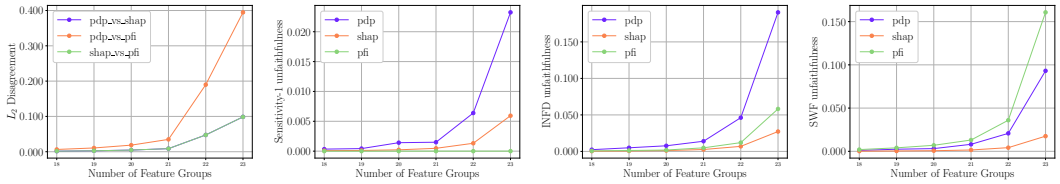
Algorithm 2 Adaptive Grouping to ReducE Explanation Disagreements (requires $\mathcal{D} = \mathcal{B}$).

procedure AGREED($f, \{\boldsymbol{x}^{(k)}\}_{k=1}^{N} \sim \mathcal{D}^N, \epsilon$)
  % Initialization
  Initialize partition $\mathcal{P}$ such that $\mathcal{P}(j) := j \ \ \forall\, j \in [d]$;
  $D \leftarrow d$;
  Compute the $D \times N \times N$ tensor $\boldsymbol{G}$ (cf. Equation 42);
  Compute the potentials $\Psi(i),\ i = 1, 2, \ldots, D$ from $\boldsymbol{G}$ (cf. Equation 45);
  Obj $\leftarrow \sum_{i=1}^{D} \Psi(i)$;
  while Obj $> \epsilon$ **do**
    % Which group to extend
    $i \leftarrow \arg\max_{i=1,2,\ldots,D} \Psi(i)$;
    $\boldsymbol{I} \leftarrow \texttt{zeros}(D-1, N, N)$
    % Find the fuse candidate
    for $j \in [D]$ such that $j \neq i$ **do**
      Compute Between-Group Interaction $I_{j,:,:}$ (cf. Equation 47);
    $j = \arg\max_{j=1,2,\ldots,D-1} \frac{1}{2N} \sum_{k=1}^{N} \left[(\frac{1}{N} \sum_{\ell=1}^{N} I_{j,k,\ell})^2 + (\frac{1}{N} \sum_{\ell=1}^{N} I_{j,\ell,k})^2 \right]$;
    % Fuse groups i and j leading to a superpartition
    $\mathcal{P}, \boldsymbol{G} \leftarrow$ UPDATE_PARTITION($\mathcal{P}, \boldsymbol{G}, \boldsymbol{I}, i, j$)
    $D \leftarrow D - 1$
    Compute the potentials $\Psi(i),\ i = 1, 2, \ldots, D$ from $\boldsymbol{G}$ (cf. Equation 45);
    Obj $\leftarrow \sum_{i=1}^{D} \Psi(i)$;
  return $\mathcal{P}$, $\boldsymbol{G}$;

Since AGREED expects $\mathcal{B} = \mathcal{D}$, we introduce a mixture distribution $\mathcal{Q} = \frac{1}{2}(\delta_{\boldsymbol{x}} + \delta_{\boldsymbol{b}})$ and note that

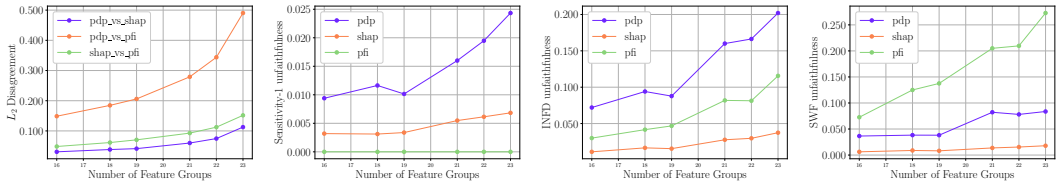
$$\mathcal{L}_f^{\text{AGREED}}(\delta_{\boldsymbol{x}}, \delta_{\boldsymbol{b}}, \mathcal{P}) = 4 \times \mathcal{L}_f^{\text{AGREED}}(\mathcal{Q}, \mathcal{Q}, \mathcal{P}). \tag{51}$$

Thus, we can apply AGREED on Images by feeding Algorithm 2 with two samples ($\boldsymbol{x}$ and $\boldsymbol{b}$) from a fictional data distribution $\mathcal{Q}$. This will lead to a $D \times 2 \times 2$ $\boldsymbol{G}$ tensor and a $D - 1 \times 2 \times 2$ $\boldsymbol{I}$ tensor.



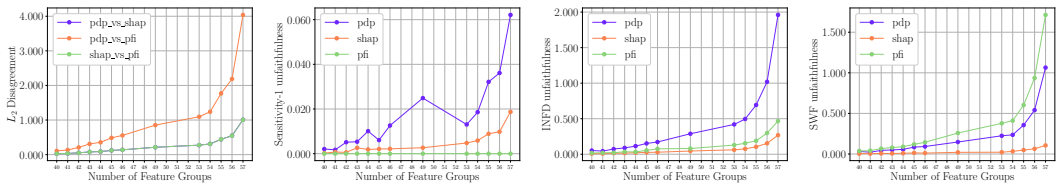
(a) (b) (c) (d)

Figure 10: Tabular Data: EBM fitted on default credit. Explanation L_2 Disagreement and unfaithfulness metrics as a function of the number of feature groups for each partitioning algorithm.



(a) (b) (c) (d)

Figure 11: Tabular Data: GBT fitted on default credit. Explanation L_2 Disagreement and unfaithfulness metrics as a function of the number of feature groups for each partitioning algorithm.



(a) (b) (c) (d)

Figure 12: Tabular Data: EBM fitted on SPAM. Explanation L_2 Disagreement and unfaithfulness metrics as a function of the number of feature groups for each partitioning algorithm.

E Extended Experiments

E.1 Tabular Toy Data

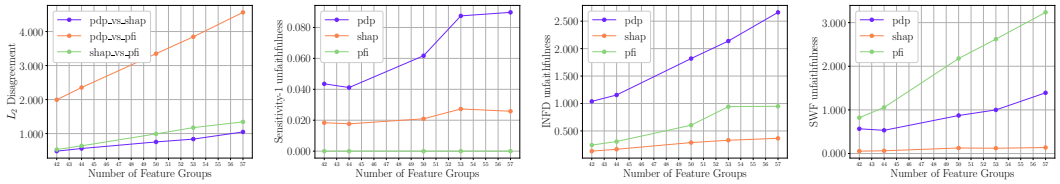
Tabular synthetic datasets are generated by first sampling N samples of d features $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{B})$ with a block-diagonal covariance matrix. The feature groups forming each block are chosen randomly. Then, another random feature partition $\mathcal{P}$ is generated to defined a group-wise additive model. Each component $g_{\mathcal{P}^{-1}(\{i\})}$ of the model is generated from the following list

1. $g_{\mathcal{P}^{-1}(\{i\})}(\boldsymbol{x}) = \prod_{j \in \mathcal{P}^{-1}(\{i\})} \prod_{k \in \mathcal{P}^{-1}(\{i\})} x_j x_k$
2. $g_{\mathcal{P}^{-1}(\{i\})}(\boldsymbol{x}) = \exp\left[-0.5 \sum_{j \in \mathcal{P}^{-1}(\{i\})} x_j^2 \right]$
3. $g_{\mathcal{P}^{-1}(\{i\})}(\boldsymbol{x}) = \sigma(\sum_{j \in \mathcal{P}^{-1}(\{i\})} \omega_j x_j^2)$ where σ is a Sine, Cosine, Tanh, ReLU.

E.2 Tabular Data

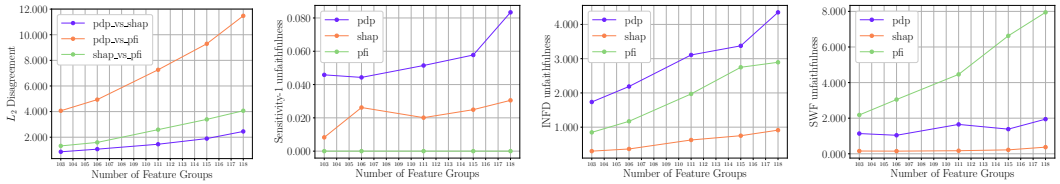
E.2.1 Additional Quantitative Results

Figures 10-14 show the trade-offs between explanation disagreement/unfaithfulness and and feature group sizes for EBM and GBT models fitted on Default-Credit, Spam, and NOMAO. The insights identical to those discussed in the main manuscript: as we group features together, disagreements between PDP/SHAP/PFI are reduced and the Sensitivity-1, INFD, SWF unfaithfulness metrics also increase in agreement.



(a) (b) (c) (d)

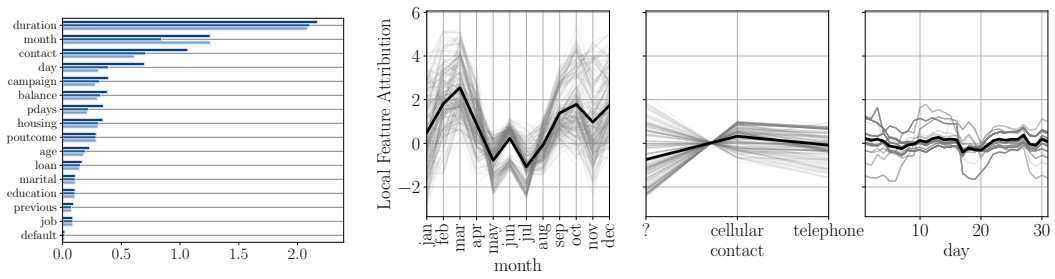
Figure 13: Tabular Data: GBT fitted on Spam. Explanation L_2 Disagreement and unfaithfulness metrics as a function of the number of feature groups for each partitioning algorithm.



(a) (b) (c) (d)

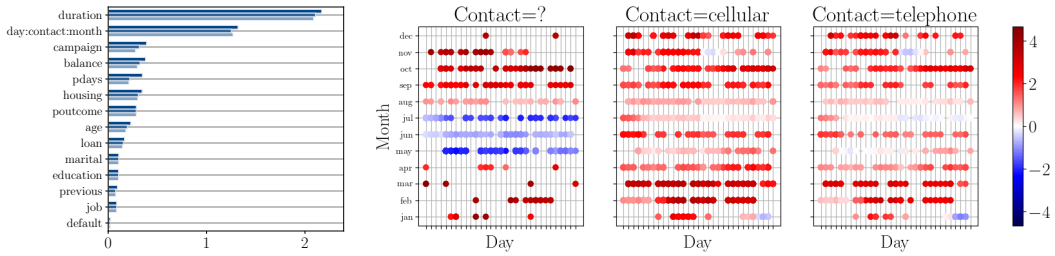
Figure 14: Tabular Data: GBT fitted on Nomao. Explanation L_2 Disagreement and unfaithfulness metrics as a function of the number of feature groups for each partitioning algorithm.

The positive effects of feature grouping are most prominent on EBMs compared to GBTs. This is because EBM are restricted to only model interaction of order 2, while GBTs with depth-T trees can model interaction whose order at-most T. Apparently, the GBTs trained on the two largest datasets (Figures 13 & 14) have learned very high-order interactions that are extremely hard to minimize. Although AGREED fails to find agreement in those two settings, the algorithm is still useful to warn the user that the model might be too complicated to be explained with feature-based explanations. Hence, it might be best to rely on a EBM if faithful and unambiguous explanations are desired.



(a) (b)

Figure 15: Marketing with no grouping. a) The global feature importance according to the PFI (opaque), SHAP (semi-transparent), and PDP (transparent) explainers. b) The PDPs of three disagreeing features (thick black line) along with the ICE (thin lines).



(a) (b)

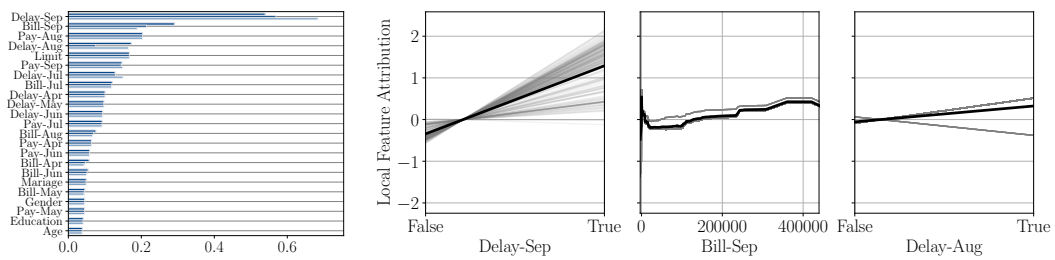
Figure 16: Marketing with grouping. a) The global feature importance according to the PFI (opaque), SHAP (semi-transparent), and PDP (transparent) explainers. b) The joint-PDP of the `month:day:contact` feature group.

E.2.2 Qualitative Results

The challenge in interpreting joint-importance scores is that, when features j, k and ℓ are treated as a group i, their joint-PDP $\phi_i^{\text{PDP}}(h, \boldsymbol{x}, \mathcal{B}, \mathcal{P})$ becomes a multivariate function of x_j, x_k and x_ℓ. Humans are notoriously bad at visualizing high-dimensional functions, so we advocate selecting 1-2 features to plot while *conditioning* the remaining ones. When features are binary or categorical, condition on their unique values. For numerical features, condition on their quantiles.

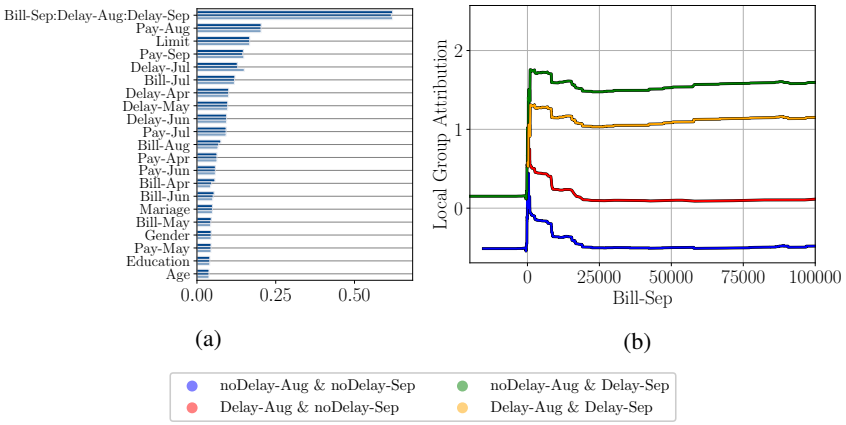
Marketing The Marketing dataset describes the marketing campaign of a Portuguese banking institution. Each instance corresponds to a distinct phone call and the binary label encodes whether the client subscribed to a term deposit. We explain an EBM fitted on this dataset using the PDP/SHAP/PFI explainers. From Figure 15 (a), the three techniques attribute very different global importances to the features `month`, `day`, and `contact`. Figure 15 (b) shows the PDP of these three features along side their Individual Conditional Expectation (ICE) (Goldstein et al., 2015). The ICE curves can be interpreted as the local trend when varying x_j while the PDP is the average trend. It is apparent that the average trend is very different from local ones, especially at the value `contact=?`. These disagreements are induced by strong feature interactions within the model.

To reduce the disagreements caused by feature interactions, we ran AGREED and obtained a group `month:day:contact`, see Figure 16. According to Figure 16 (a), there are now almost no disagreements between the global group importance reported by PDP/SHAP/PFI. The joint-PDP of `month:day:contact` can be visualized using a scatter plot along `month:day` while conditioning on different values of `contact` $\in$ [`?`, `cellular`, `telephone`]. This is presented in Figure 16 (b). When `contact=?`, there is a significant drop in model output during June and July compared to other values of `contact`. Moreover, the trends along `month` and `day` hardly appear to be additive : there is a sharp drop in late January that does not occur in other months. Therefore, it is better to interpret them jointly as a single "date" feature group.



(a) (b)

Figure 17: Default-Credit with no grouping. a) The global feature importance according to the PFI (opaque), SHAP (semi-transparent), and PDP (transparent) explainers. b) The PDPs of three important features (thick black line) along with the ICE (thin lines).



(a) (b)

Figure 18: Default-Credit with grouping. a) The global feature importance according to the PFI (opaque), SHAP (semi-transparent), and PDP (transparent) explainers. b) The joint-PDP of the `Delay-Aug:Delay-Sep:Bill-Sep` feature group.

Default-Credit The Default-Credit dataset aims at predicting if clients of a Taiwanese bank will default on their credit. The data contains records of 30K individuals and 23 features related to past payments/bills/delays and demographic characteristics. We explain an EBM fitted on this dataset using the PDP/SHAP/PFI explainers. Figure 17 (a) demonstrates that the three explainers provide different global importances to the features `Delay-Sep`, `Delay-Aug`, and `Bill-Sep`. Figure 17 (b) shows their PDP and ICE local attribution. By comparing the PDP and ICE, it clear that the PDP of `Delay-Aug` is misleading since having a delayed payment in August sometimes increases the model output and sometimes decreases it. This suggests that the effects of `Delay-Aug` cannot be faithfully described using single feature importance score.

To faithfully explain the effects of interacting features, we ran AGREED and obtained a group `Delay-Aug:Delay-Sep:Bill-Sep`. According to Figure 18 (a), PDP/SHAP/PFI now agree on the global importance of each feature group. The joint-PDP of `Delay-Aug:Delay-Sep:Bill-Sep` is a multivariate function involving two binary features and a numerical one. This function can be visualized by plotting four line charts w.r.t `Bill-Sep` (one line for each configuration of the remaining two binary variables). See Figure 18 (b) for the results. Interestingly, the impact of August delays depends on whether there was a September delay. Comparing the yellow curve to the green one, and the red/blue curves, the effect of `Delay-Aug` is completely reversed depending on `Delay-Sep`. The effect of `Bill-Sep` also depends on `Delay-Sep`. We are not sure why these trends are happening in the data, but at least AGREED warns us that trends involving `Delay-Aug:Delay-Sep:Bill-Sep` are inherently high-dimensional and should be visualized accordingly.

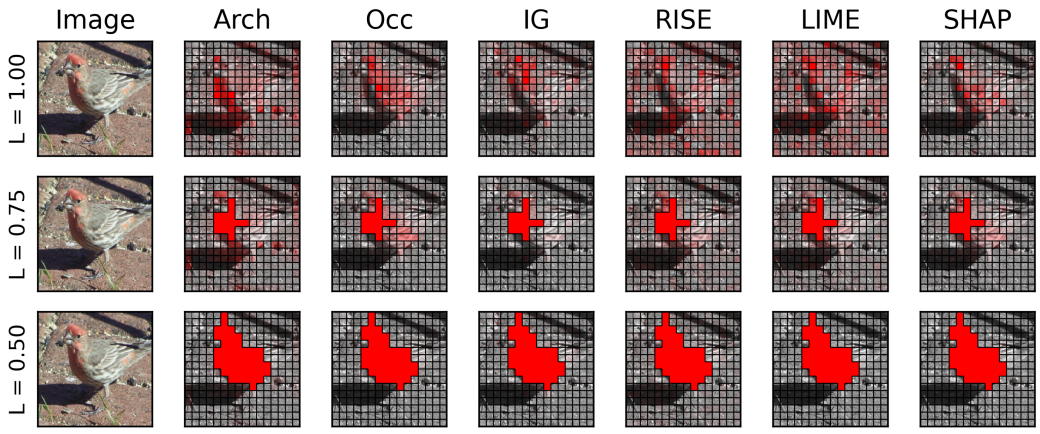


Figure 19: Explaining the "House Finch" prediction of a ResNet18. AGREED yields a partition with increased agreement between the various saliency map methods.

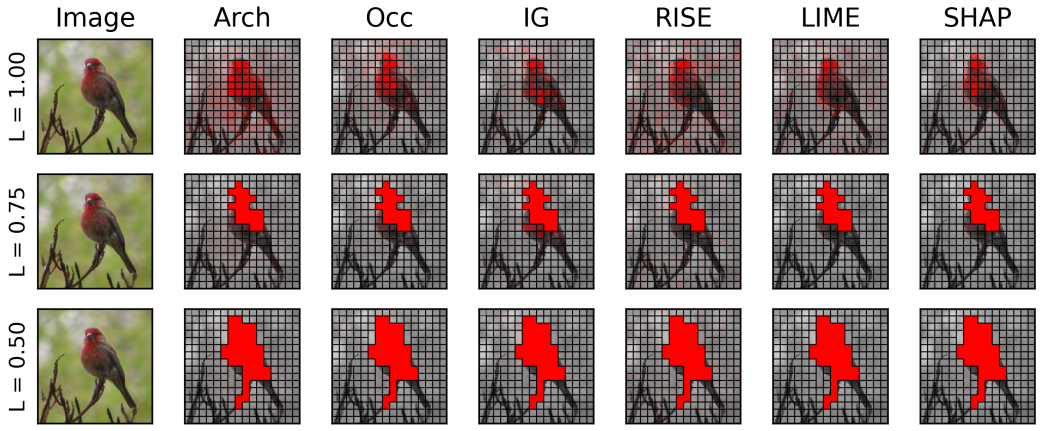


Figure 20: Explaining the "House Finch" prediction of a ResNet18. AGREED yields a partition with increased agreement between the various saliency map methods.

E.3 MiniImageNet

Figures 19 to 29 present the saliency maps resulting from the AGREED partitions. The model under study is a ResNet18 pre-trained on ImageNet. We see that, in general, AGREED identifies a large patch of great importance that covers the animal. However, there are exceptions: in Figures 21, 24, 26, 29 there are multiple patches that cover specific parts of the animal.

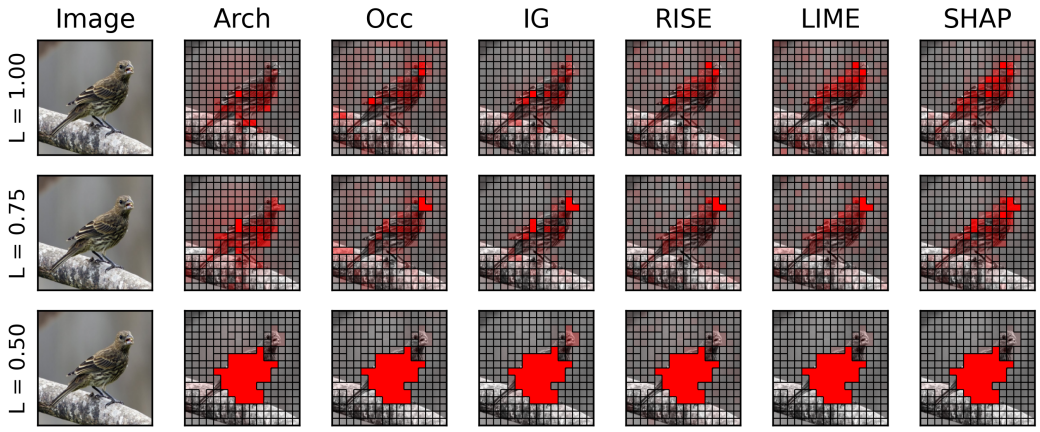


Figure 21: Explaining the "House Finch" prediction of a ResNet18. AGREED yields a partition with increased agreement between the various saliency map methods.

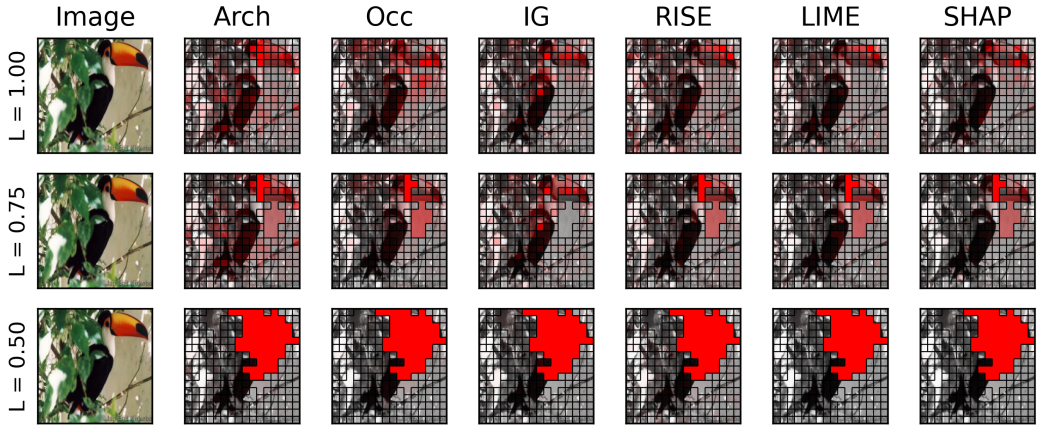


Figure 22: Explaining the "Toucan" prediction of a ResNet18. AGREED yields a partition with increased agreement between the various saliency map methods.

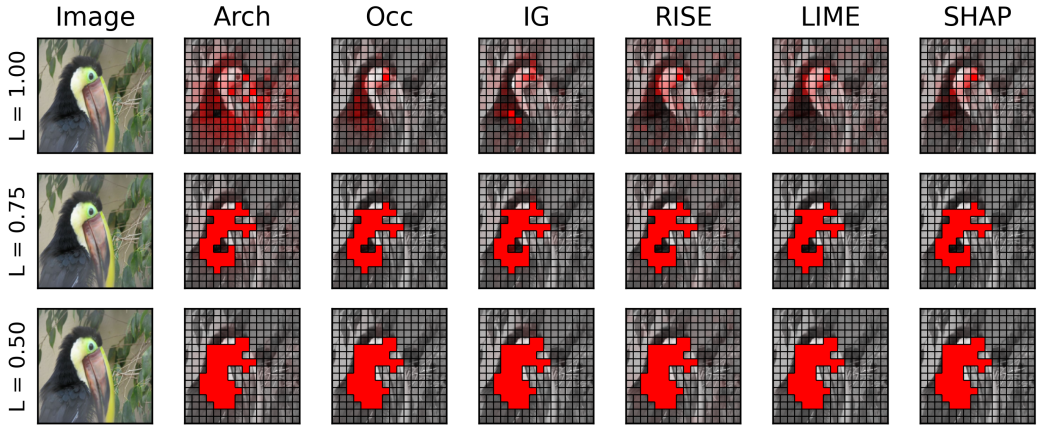


Figure 23: Explaining the "Toucan" prediction of a ResNet18. AGREED yields a partition with increased agreement between the various saliency map methods.

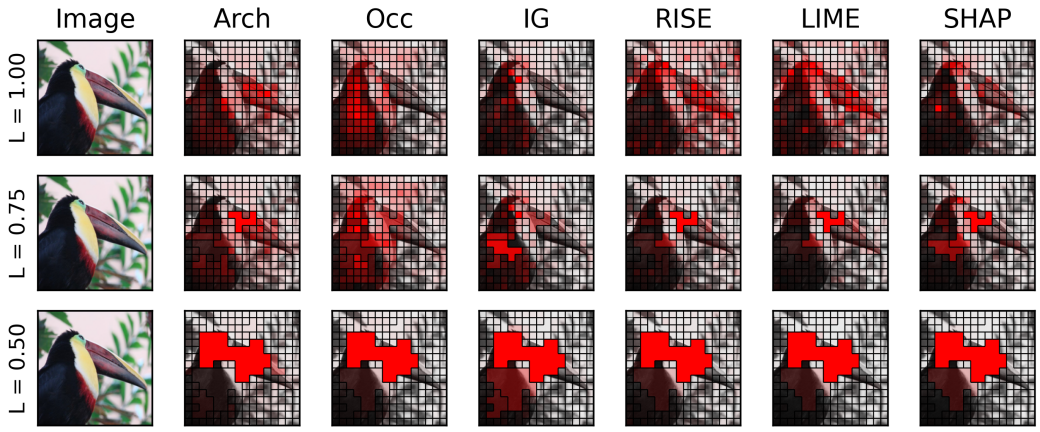


Figure 24: Explaining the "Toucan" prediction of a ResNet18. AGREED yields a partition with increased agreement between the various saliency map methods.

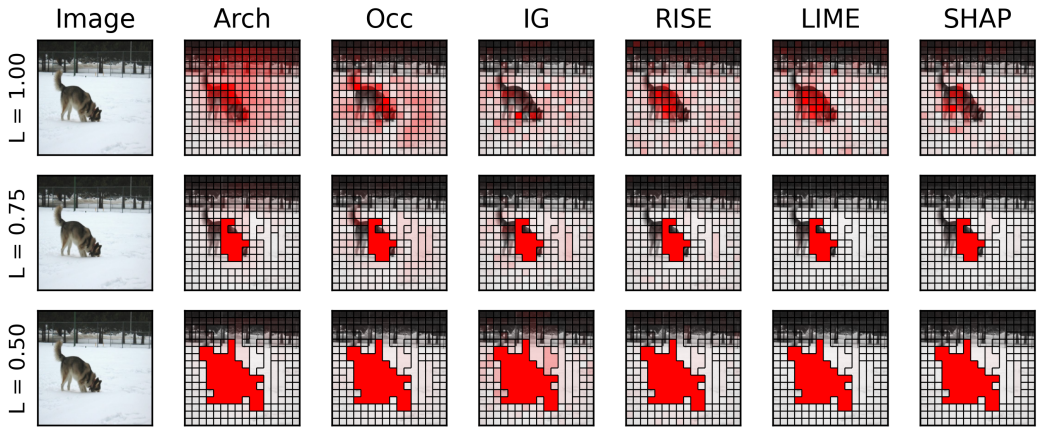


Figure 25: Explaining the "Malamute" prediction of a ResNet18. AGREED yields a partition with increased agreement between the various saliency map methods.

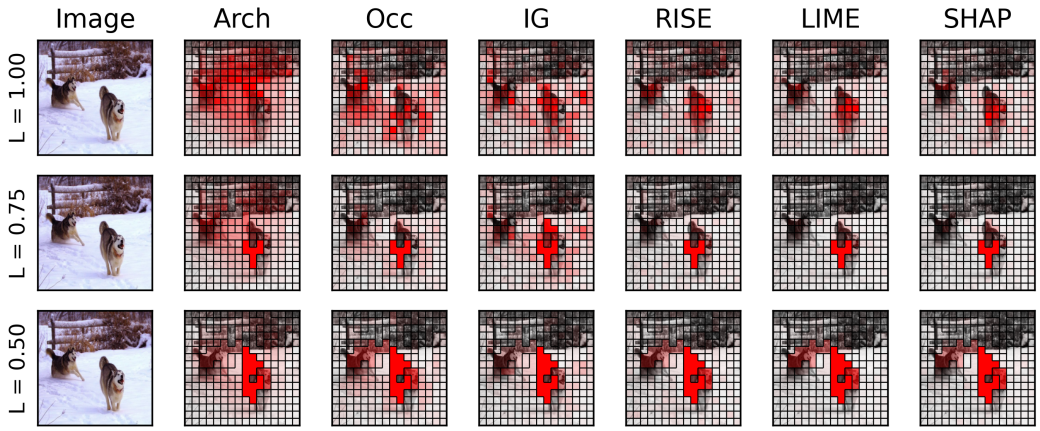


Figure 26: Explaining the "Malamute" prediction of a ResNet18. AGREED yields a partition with increased agreement between the various saliency map methods.

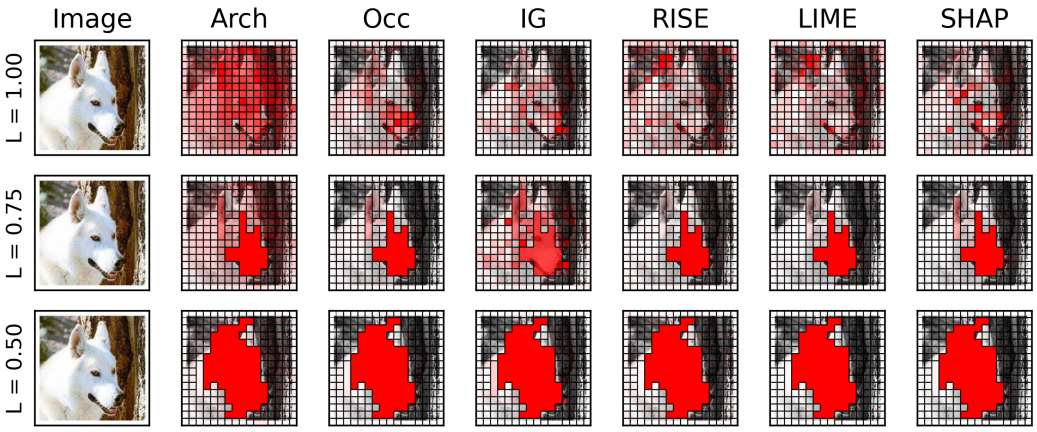


Figure 27: Explaining the "White Wolf" prediction of a ResNet18. AGREED yields a partition with increased agreement between the various saliency map methods.

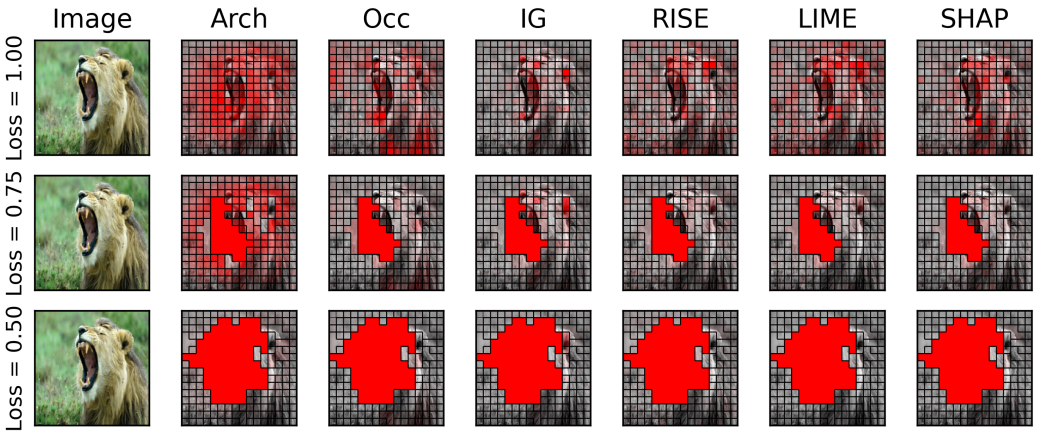


Figure 28: Explaining the "Lion" prediction of a ResNet18. AGREED yields a partition with increased agreement between the various saliency map methods.

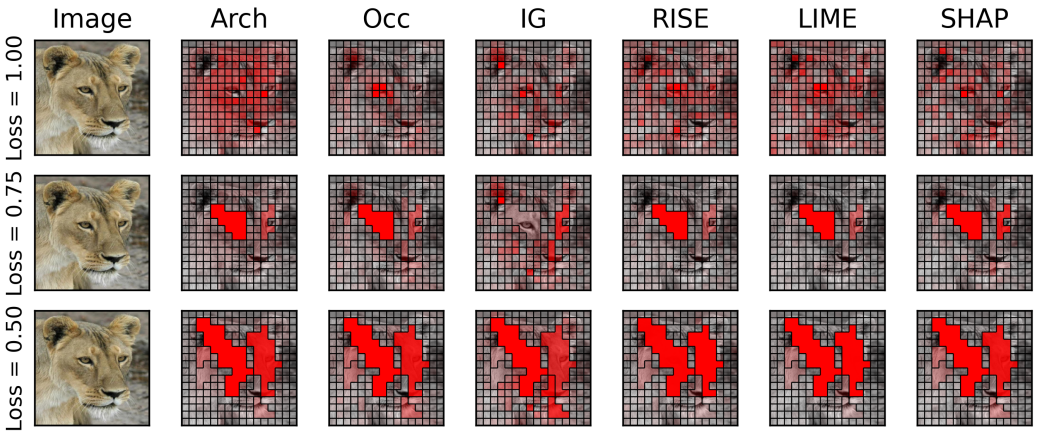


Figure 29: Explaining the "Lion" prediction of a ResNet18. AGREED yields a partition with increased agreement between the various saliency map methods.