DAG-SHAP: FEATURE ATTRIBUTION IN DAG BASED ON EDGE INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Shapley value-based feature attribution methods face challenges in scenarios with complex feature interactions and causal relationships, even when a causal structure is provided. The assumption on the attribution objects of existing methods often deviates from practical scenarios as they cannot capture the exogenous influence of features through each edge in the causal graph, leading to unreasonable interpretations. To overcome these limitations, we propose a novel feature attribution method called DAG-SHAP, which is based on edge intervention. DAG-SHAP treats the exogenous contributions in each ongoing feature edge as an individual attribution object ensuring that both externality and exogenous contributions of features are appropriately captured. Additionally, we introduce an approximation method for efficiently computing DAG-SHAP. Extensive experiments on both synthetic and real datasets validate the effectiveness of DAG-SHAP. Our code can be found in the anonymous repository at `https://anonymous.4open.science/r/dag-30F2`.

1 INTRODUCTION

In recent years, the field of model interpretation has gained significant attention (Scott et al., 2017; Ribeiro et al., 2016; Sundararajan et al., 2017; Chen et al., 2024; Machiraju et al., 2024) due to the increasing demands and complexity of machine learning (ML) models in real-world applications. Shapley value-based feature attribution methods have been extensively explored in this domain since they uniquely offer a fair allocation guarantee of cooperative contributions rooted in game theory, with desirable properties including efficiency, symmetry, redundancy, and additivity (Winter, 2002). Moreover, owing to their model-agnostic nature and ease of implementation, Shapley value-based methods are highly user-friendly and versatile. For healthcare, Shapley value-based attribution can help explain which features, such as age, blood pressure, and blood sugar levels, are important when predicting whether a patient is likely to develop diabetes (Ter-Minassian et al., 2023). Similarly, they can also be applied in credit scoring to help financial institutions understand why a customer's predicted credit score is low or high by identifying the key factors to the predicted credit score (Chen et al., 2022).

However, modeling feature interactions appropriately remains a significant challenge for Shapley value-based feature attribution methods. The off-manifold Shapley value (Scott et al., 2017), one of the pioneering Shapley value-based feature attribution methods, assumes feature independence, which is inappropriate in most practical scenarios. Meanwhile, the on-manifold Shapley value (Scott et al., 2017; Sundararajan & Najmi, 2020) considers the feature dependency. It fills excluded features with conditional expected values based on their correlations with included features when measuring the utility of feature subsets, aiming to make the interpretation more reasonable. However, filling excluded features based solely on correlation may not align with the data generation process, particularly when causal relationships exist among features, potentially leading to causal reversion (Jung et al., 2022a). To address this, researchers consider the sequential and causal relationships in data generation, like asymmetric Shapley value (Frye et al., 2020) and causal Shapley value (Heskes et al., 2020). Furthermore, existing attribution methods that focus on feature vertices fail to adequately capture interactions between features (Ter-Minassian et al., 2023). Shapley Flow (Wang et al., 2021) attributes the contribution in each cut of the feature graph by treating paths as players. Recursive Shapley value (Singal et al., 2021) attributes contributions following a top-down principle by first attributing to "source" vertices and then flowing them down the directed acyclic graph (DAG).

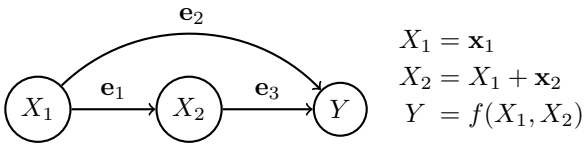


Figure 1: A toy numerical example with two features (X_1, X_2) and a target label Y, where arrows indicate direct effects and lowercase letters represent exogenous variables.

Despite the aforementioned significant efforts to enhance feature interactions in the attribution process, *externality* and *exogeneity* in existing approaches are still inadequately addressed. Attribution methods based on feature vertices adopting asymmetry samplings, such as asymmetry Shapley value and asymmetry causal Shapley value, fail to make attribution results satisfy externality. When using the asymmetrical ordering of feature attribution methods (the parent vertex must appear before the child vertex), the marginal contribution of the parent vertex cannot receive externality gains from cooperating with the child vertex. For example in Figure 1, the direct influence $X_1 \to Y$ cannot receive a marginal contribution from cooperating with X_2 which is $X_2 \to Y$. In addition, the attribution methods focusing on attributing contributions at each graph cut like Shapley Flow and recursive Shapley value, fail to recognize exogenous contributions of features. Exogenous contribution refers to the part of the contribution in each feature that is not influenced by other features in the explained input. Shapley Flow and recursive Shapley value assume that only feature vertices without incoming edges have exogenous contributions. Hence, the contribution of exogenous variable $\mathbf{x}_2$ cannot be properly captured. Apparently, the assumption that intermediate vertices have no endogenous contributions is misaligned with real-world feature attribution scenarios.

In this paper, we explore the potential to enhance the reasonableness of attribution methods through an advanced investigation of feature interactions. Specifically, we incorporate fine-grained causal relationships under the assumption that features are structured as DAG. We briefly summarize our contributions as follows. (1) We propose an edge intervention feature attribution method DAG-SHAP that allows for interventions on certain child vertices through their parent vertices without affecting other child vertices of the same parent. (2) Our attribution method not only meets the requirement of externality but also captures the exogenous contributions of each feature. (3) We present an approximate method for the practical computation of DAG-SHAP and validate its effectiveness using both synthetic and real datasets.

2 PRELIMINARIES

In this section, we present the problem setting for the feature attribution task, review the concept of Shapley value, and discuss several representative Shapley value-based feature attribution methods.

2.1 PROBLEM DEFINITION AND SHAPLEY VALUE

Problem Definition. Given a trained ML model $f(\cdot)$ and an n-dimensional input feature vector $\boldsymbol{x} \in \mathbb{R}^n$, our goal is to assign an attribution value ϕ_i to each feature i $(1 \leq i \leq n)$, reflecting its contribution to the model's prediction $f(\boldsymbol{x})$. Given a baseline f_0, and for $i \in \mathbb{V}$ where $\mathbb{V} = \{1, \cdots, n\}$ represents the n features, $f(\boldsymbol{x})$ can be expressed as follows

$$f(\boldsymbol{x}) = f_0 + \sum_{i=1}^{n} \phi_i. \tag{1}$$

Specifically, we focus on the scenario where the causal graph of features $\mathbb{V}$ forms a DAG. The attribution of each feature to the model output must capture only the exogenous contribution of the feature, where the exogenous contribution refers to the portion of the feature's impact on the output that originates from itself. It ensures that the attribution reflects the intrinsic effect of the feature and its downstream causal influence on the outcome.

Shapley Value. Consider a set of players $\mathbb{V} = \{1, \ldots, n\}$. A *coalition* $\mathcal{S}$ is a subset of $\mathbb{V}$ that cooperates to complete a task. A utility function $\mathcal{U}(\mathcal{S})$ $(\mathcal{S} \subseteq \mathbb{V})$ is the utility of coalition $\mathcal{S}$ for a task.

The *marginal contribution* of player i with respect to a coalition $\mathcal{S}$ is $\mathcal{U}(\mathcal{S} \cup \{\mathrm{i}\}) - \mathcal{U}(\mathcal{S})$. Shapley value is the unique metric that satisfies the properties of fair reward allocation, including balance, symmetry, additivity, and zero element (Winter, 2002). It measures the expectation of marginal contribution by i in all possible coalitions. That is

$$\mathcal{SV}_i = \frac{1}{n} \sum_{\mathcal{S} \subseteq \mathbb{V} \setminus \{i\}} \frac{\mathcal{U}(\mathcal{S} \cup \{i\}) - \mathcal{U}(\mathcal{S})}{\binom{n-1}{|\mathcal{S}|}}. \tag{2}$$

According to Equation 2, we can find that computing the exact Shapley value requires enumerating all utilities for all player subsets. Therefore, the computational complexity of exactly calculating the Shapley value is exponential (Deng & Papadimitriou, 1994).

3 Edge Intervention Causal Shapley value

In this section, we analyze the limitations of existing methods using a simple attribution task. We then introduce DAG-SHAP, a feature attribution method based on edge intervention that leverages the DAG of causal relationships between input features, aiming to enhance the reasonableness of feature attribution.

We provide definitions of several properties below, which will be discussed in detail afterward.
Causality: The contribution of a feature to the prediction outcome should be based on its true causal impact, rather than just its statistical correlation with other features. In essence, attribution should reflect the change in the model's output that is directly caused by the feature, independent of any dependencies on other features.
Efficiency: The sum of the attribution values of all nodes equals to the difference between the prediction outcome when all nodes are added. This can be expressed as $\sum_{i \in \mathbb{V}} \Phi(i) = f(\boldsymbol{x}) - \mathbb{E}[f(\mathbf{x})]$, where $\boldsymbol{x}$ represents the input features being explained.
Externality: The attribution value of feature vertex k should derive benefit from the cooperation of another feature vertex k' if there exists a path to the target label that passes through k but does not pass through k'.
Exogeneity: The attribution method should identify and measure the independent contribution of each feature to the prediction outcome. Each feature's effect should be attributed to its exogenous influence, without being confounded by other input features.

3.1 Motivation Example for Illustrating the Limitations of Existing Methods

Data Generation Process. We use the toy example in Figure 1 to show why existing methods can produce unreasonable interpretations for model output. Suppose there are two input features X_1 and X_2 and a target feature $Y = X_1 \cdot X_2$. The generation of features follows the process below. $X_1 = \mathbf{x}_1$, where $\mathbf{x}_1$ is a random variable uniformly distributed on $(0, 1)$, representing the exogenous influence of X_1; $X_2 = X_1 + \mathbf{x}_2$, where $\mathbf{x}_2$ is another random variable uniformly distributed on $(0, 1)$, representing the exogenous influence of X_2. In summary, X_1 directly influences Y and indirectly influences Y through X_2. X_2 influences Y with its own exogenous influence x_2 and transfers the indirect influence of X_1. We want to attribute a value to each feature of a specific explained input $\boldsymbol{x}^* = [\boldsymbol{x}_1^*, \boldsymbol{x}_2^*] = [0.2, 0.8]$ with respect to the uniform distributions of each feature.

Proposition 1. *For off-manifold Shapley value, on-manifold Shapley value, and symmetry causal Shapley value, the attribution results do not satisfy causality.*

For off-manifold Shapley value, the marginal contribution cannot account for the causal interaction between features. For example, the marginal contribution of $\boldsymbol{x}_1^* = 0.2$ when cooperating with the empty set is $\mathbb{E}[f(0.2, \mathbf{x}_1 + \mathbf{x}_2)] - \mathbb{E}[f(\mathbf{x}_1, \mathbf{x}_1 + \mathbf{x}_2)]$. However, the value of feature X_2 is influenced by $\boldsymbol{x}_1^*$ but the feature value of X_2 in the utility $\mathbb{E}[f(0.2, \mathbf{x}_1 + \mathbf{x}_2)]$ is independent from $\boldsymbol{x}_1^* = 0.2$. So the marginal contribution does not incorporate the indirect effect of $\boldsymbol{x}_1^* = 0.2$.

For on-manifold Shapley value, there is causal reversion in the explained results. For example, when calculating the marginal contribution of $\boldsymbol{x}_2^* = 0.8$ cooperating with the empty set, On-manifold Shapley value yields $\mathbb{E}[f(\mathbf{x}_1, 0.8) | \boldsymbol{x}_2^* = 0.8] - \mathbb{E}[f(\mathbf{x}_1, \mathbf{x}_1 + \mathbf{x}_2)]$, where x_1 is conditioned on $\mathbf{x}_2^*$. However, X_2 cannot influence X_1 from a causal perspective. Therefore, calculating contribution $\boldsymbol{x}_2^* = 0.8$ with condition expectation includes the part of contribution which does not belong to $\boldsymbol{x}_2^*$.

Symmetry causal Shapley value faces the same causal reversion problem as on-manifold Shapley value. Considering the marginal contribution of $\boldsymbol{x}_2^* = 0.8$ when cooperating with empty set, it yields $\mathbb{E}[f(\mathbf{x}_1, 0.8)|\,\mathrm{do}(\mathbf{x}_2 = 0.8)] - \mathbb{E}[f(\mathbf{x}_1, \mathbf{x}_1 + \mathbf{x}_2)]$. However, the utility of $\boldsymbol{x}_2^* = 0.8$ when cooperating with an empty cannot be $\mathbb{E}[f(\mathbf{x}_1, 0.8)|\,\mathrm{do}(\mathbf{x}_2 = 0.8)]$ because $\boldsymbol{x}_2^* = 0.8$ already has the influence of $\boldsymbol{x}_1^* = 0.2$.

Proposition 2. *For asymmetry Shapley value and asymmetry causal Shapley value, the attribution results do not satisfy externality.*

Asymmetry Shapley value and asymmetry causal Shapley value calculate the marginal contributions on the permutations where ancestor vertices before their descendants. However, these asymmetry methods fail to guarantee the externality property in feature attribution. For example, $\boldsymbol{x}_1^* = 0.2$ has a direct influence on Y, and the influence is independent of X_2. But the direct influence does not receive a marginal contribution from X_2 since $\boldsymbol{x}_1^*$ must appear before $\boldsymbol{x}_2^*$ in the permutations. In other words, the attributed value of $\boldsymbol{x}_1^*$ is independent of $\boldsymbol{x}_2^*$, which does not satisfy the externality in cooperation game theory (Shapley, 1953), where $\boldsymbol{x}_1^*$ should benefit from the cooperation with variable $\mathbf{x}_2$.

Proposition 3. *For Shapley Flow and recursive Shapley value, the attribution process cannot capture the exogenous contributions of intermediate vertices.*

Shapley Flow and recursive Shapley value are unsuitable for this scenario due to their cut efficiency property, which requires that the sum of contributions be equal for every cut. This implies that the contribution of $\boldsymbol{x}_1^*$ through $\boldsymbol{x}_2^*$ is considered equal to the direct contribution of $\boldsymbol{x}_2^*$, an assumption that may not hold true in real-world situations. They do not account for the independent exogenous contributions of each feature.

3.2 Our Proposed Method: DAG-SHAP

We propose DAG-SHAP, an edge-based feature attribution method suitable for the scenario where the causal relations of input features can be formulated as a DAG. DAG-SHAP enables us to gain a deeper understanding of how features collaboratively contribute to the model output using the edge intervention to capture the causal influence within the features.

Edge Intervention. Given a directed acyclic graph $\mathcal{G}$ with a set of vertices $\mathbb{V} = \{1, \cdots, n\}$ and directed edges $\boldsymbol{E} = \{\mathrm{e}_1, \cdots, \mathrm{e}_m\}$, where n represents the number of vertices and m the number of edges in $\mathcal{G}$, each vertex corresponds to a feature, and each edge denotes a causal influence between features. An edge $\mathrm{e}_i = (p_i, c_i, \mathbf{x}_{p_i})$ indicates that parent vertex $p_i \in \mathbb{V}$ has a direct causal influence on child vertex $c_i \in \mathbb{V}$, where the feature corresponding to p_i follows distribution $\mathbf{x}_{p_i}$. For a specific explained input $\boldsymbol{x}$, we instantiate the edges as $E = \{e_1, \cdots, e_m\}$, where each edge $e_i = (p_i, c_i, \boldsymbol{x}_{p_i})$ corresponds to a specific value of $\boldsymbol{x}_{p_i}$. The causal influence from p_i is then transferred to c_i. We define an edge intervention in $\mathcal{G}$ as $\mathrm{do}(\mathrm{e}_{\mathcal{S}} = e_{\mathcal{S}})$, where $\mathcal{S}$ is the set of edge indices which are intervened. Such interventions make the effects of parents transfer to childs , providing insights into how changes in one feature directly affect another. This edge intervention isolates the causal pathway's impact without altering the entire network of child vertices, enabling a targeted analysis of causality within the graph.

DAG-SHAP. Let Π denote the set of all permutations of edges E, where each permutation constitutes a topological ordering of E. This arrangement ensures that each edge appears only after all its prerequisite edges have been ordered, thereby allowing interventions to be applied sequentially without violating causal relationships. We define the attribution value of an edge e_i for the explained input $\boldsymbol{x}$ as the average marginal contribution of applying the edge intervention on e_i, given that interventions have already been applied to the edges preceding it in the permutations within Π. Consequently, the edge intervention causal Shapley value of edge e_i can be formulated as

$$\Psi(e_i) = \sum_{\pi \in \Pi} \frac{1}{|\Pi|} \{\mathbb{E}[f(\mathbf{x})|\,\mathrm{do}(\mathrm{e}_{\underline{\mathcal{S}}_\pi^i} = e_{\underline{\mathcal{S}}_\pi^i})] - \mathbb{E}[f(\mathbf{x})|\,\mathrm{do}(\mathrm{e}_{\mathcal{S}_\pi^i} = e_{\mathcal{S}_\pi^i})]\}, \tag{3}$$

where $\mathcal{S}_\pi^i = \{j|\pi(j) < \pi(i)\}$, $\underline{\mathcal{S}}_\pi^i = \{j|\pi(j) \leq \pi(i)\}$, and $\pi(i)$ is the index of i in permutation π.

Computation Example. We use example $Y = X_1 \cdot X_2$ in Section 3.1 to illustrate the calculation process of the edge intervention causal Shapley value. We denote edge $X_1 \to X_2$ as e_1, edge $X_1 \to Y$ as e_2, and edge $X_2 \to Y$ as e_3. For the input to be explained, $\boldsymbol{x}^* = [0.2, 0.8]$,

the instantiated edges are $e_1 = (X_1, X_2, 0.2)$, $e_2 = (X_1, Y, 0.2)$, and $e_3 = (X_2, Y, 0.8)$. The permutations formed by e_1, e_2, e_3 are (e_1, e_2, e_3), (e_1, e_3, e_2), (e_2, e_1, e_3), (e_2, e_3, e_1), (e_3, e_1, e_2), and (e_3, e_2, e_1). Since e_1 must precede e_3 to avoid reversing the causal effect, the valid permutations are (e_1, e_2, e_3), (e_1, e_3, e_2), and (e_2, e_1, e_3). When calculating the marginal contribution in the permutation (e_1, e_2, e_3), we first obtain the marginal contribution of adding edge e_1 from the empty set, which is $\mathbb{E}[f(\mathbf{x})|\,\mathrm{do}(\mathbf{e}_1 = e_1)] - \mathbb{E}[f(\mathbf{x})]$. When performing $\mathrm{do}(\mathbf{e}_1 = e_1)$, X_2 is conditionally dependent on $X_1 = 0.2$, but the direct effect of X_1 on Y is absent. Therefore, $\mathbb{E}[f(\mathbf{x})|\,\mathrm{do}(\mathbf{e}_1 = e_1)] = \mathbb{E}_{\overline{\mathbf{x}}_2 \sim P(\mathbf{x}_2|\boldsymbol{x}_1^* = 0.2)}[f(\mathbf{x}_1, \overline{\mathbf{x}}_2)]$ where $P(\mathbf{x}_2|\boldsymbol{x}_1^* = 0.2)$ represents the distribution of $\mathbf{x}_2$ when it is conditioned on $\boldsymbol{x}_1^* = 0.2$. Thus the marginal contribution of e_1 is $\mathbb{E}_{\overline{\mathbf{x}}_2 \sim P(\mathbf{x}_2|\boldsymbol{x}_1^* = 0.2)}[f(\mathbf{x}_1, \overline{\mathbf{x}}_2)] - \mathbb{E}[f(\mathbf{x})]$. Next, we calculate the marginal contribution of e_2, which is $\mathbb{E}[f(\mathbf{x})|\,\mathrm{do}(\mathbf{e}_{\{1,2\}} = e_{\{1,2\}})] - \mathbb{E}[f(\mathbf{x})|\,\mathrm{do}(\boldsymbol{e}_1 = e_1)]$. Executing $\mathrm{do}(\mathbf{e}_{\{1,2\}} = e_{\{1,2\}})$ applies the direct effect of $X_1 = 0.2$ to Y, and the expected value is $\mathbb{E}_{\overline{\mathbf{x}}_2 \sim P(\mathbf{x}_2|\boldsymbol{x}_1^* = 0.2)}[f(0.2, \overline{\mathbf{x}}_2)]$. Thus the marginal contribution of e_2 is $\mathbb{E}_{\overline{\mathbf{x}}_2 \sim P(\mathbf{x}_2|\boldsymbol{x}_1^* = 0.2)}[f(0.2, \overline{\mathbf{x}}_2)] - \mathbb{E}_{\overline{\mathbf{x}}_2 \sim P(\mathbf{x}_2|\boldsymbol{x}_1^* = 0.2)}[f(\mathbf{x}_1, \overline{\mathbf{x}_2})]$. When calculating the marginal contribution of e_3, $X_1 = 0.2$ and $X_2 = 0.8$ are already fixed. The marginal contribution of e_3 is $\mathbb{E}[f(0.2, 0.8)] - \mathbb{E}_{\overline{\mathbf{x}}_2 \sim P(\mathbf{x}_2|\boldsymbol{x}_1^* = 0.2)}[f(0.2, \overline{\mathbf{x}_2})]$. Similarly, we can compute the marginal contributions for the other permutations (e_1, e_3, e_2) and (e_2, e_1, e_3). The attribution values for e_1, e_2, e_3 are determined by averaging the marginal contributions across all valid permutations, while the attribution values for vertices X_1 and X_2 are calculated by summing the attribution values of their respective outgoing edges.

The edge intervention causal Shapley value of vertex k is then defined as the sum of attribution values of its outgoing edges, which can be formulated as follows

$$\Phi(k) = \sum_{e \in \mathcal{O}_k} \Psi(e), \tag{4}$$

where $\mathcal{O}_k$ is the set of edges with parent vertex k.

DAG-SHAP satisfies several well-established properties that have been extensively discussed in previous work (Sundararajan & Najmi, 2020), including Linearity, Implementation Invariance, Sensitivity, and Dummy, all of which are defined in the appendix. These properties are already satisfied by existing methods and serve as foundational criteria in attribution methods. For properties related to Causality, Efficiency, Externality, and Exogeneity, different approaches vary in their ability to satisfy these criteria. A summary of this comparison is presented in Table 1. Additionally, detailed explanations are provided in the appendix, where we outline how and why DAG-SHAP fulfills each of these properties.

Remark. do-Shapley and Causal Shapley both use node interventions, with the key difference being whether the goal is to explain the impact on the data generation process of Y or on a predictive model f. This is reflected in their utility definitions: for do-Shapley, $v(S) = \mathbb{E}\left[\mathbf{Y} \mid \mathrm{do}(\mathbf{x}_S = x_S)\right]$, and for Causal Shapley, $v(S) = \mathbb{E}\left[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{x}_S = x_S)\right]$. Node interventions can be further classified into symmetry sampling node intervention and asymmetry sampling node intervention. Shapley-ICC uses structure-preserving interventions to measure node contributions in reducing uncertainty in the generation of Y. In the appendix Section D, we provide examples to illustrate the differences between the intervention methods, specifically highlighting our proposed edge-based intervention approach.

4 Implementation of Edge Intervention Causal Shapley Value

In this section, we introduce an exact computation method for edge intervention causal Shapley value by inferring feature distributions. Additionally, we propose an approximate algorithm for cases where these distributions are difficult to determine in real datasets.

4.1 Exact Computation of Edge Intervention Causal Shapley Value

To compute the edge intervention Shapley value, the initial step involves calculating the value function $v(\mathcal{S}) = \mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_{\mathcal{S}} = e_{\mathcal{S}})]$, which necessitates determining the distribution of $\mathbf{x}$ under edge intervention. When intervening on the edges $\mathbf{e}_{\mathcal{S}}$, for those edges where the child vertex is the target label, the values of the parent vertices corresponding to these edges can be directly fixed by the feature values of the input $\boldsymbol{x}$ since these edges represent the direct influence of features to the target

Table 1: Comparison of feature attribution method in terms of having (✔) and missing (✘) desiderata.

Method	Causality	Efficiency	Externality	Exogeneity
Off-manifold SV	✘	✔	✔	✔
On-manifold SV	✘	✔	✔	✔
Asymmetry SV	✔	✔	✘	✔
Symmetry causal SV	✘	✔	✔	✔
Asymmetry causal SV	✔	✔	✘	✔
Shapley Flow	✔	✘	✘	✘
Recursive SV	✔	✘	✘	✘
DAG-SHAP	✔	✔	✔	✔

label. Let $\mathbb{D}_{\mathcal{S}}$ denote the set of features that have a direct influence on the target label, and whose direct influence on the target label is through the edges in $\mathbf{e}_{\mathcal{S}}$. $\mathbb{V}_{\mathcal{S}}$ is the complement of $\mathbb{D}_{\mathcal{S}}$. For the value function $v(\mathcal{S}) = \mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_{\mathcal{S}} = e_{\mathcal{S}})]$, the values of the features in $\mathbb{D}_{\mathcal{S}}$ are determined by the values of the features in $\boldsymbol{x}$, given that their direct influences are subjected to intervention. Consequently, the edge intervention causal Shapley value for edge e_i can be reformulated as follows

$$\begin{aligned}\Psi_v(e_i) &= \sum_{\pi\in\Pi}\frac{1}{|\Pi|}\{\mathbb{E}[f(\mathbf{x})|\,\mathrm{do}(\mathbf{e}_{\underline{\mathcal{S}}^i_\pi} = e_{\underline{\mathcal{S}}^i_\pi})] - \mathbb{E}[f(\mathbf{x})|\,\mathrm{do}(\mathbf{e}_{\mathcal{S}^i_\pi} = e_{\mathcal{S}^i_\pi})]\}\\ &= \sum_{\pi\in\Pi}\frac{1}{|\Pi|}\{\mathbb{E}[f(\mathbf{x}_{\mathbb{V}_{\underline{\mathcal{S}}^i_\pi}}, \boldsymbol{x}_{\mathbb{D}_{\underline{\mathcal{S}}^i_\pi}})|\,\mathrm{do}(\mathbf{e}_{\underline{\mathcal{S}}^i_\pi} = e_{\underline{\mathcal{S}}^i_\pi})] - \mathbb{E}[f(\mathbf{x}_{\mathbb{V}_{\mathcal{S}^i_\pi}}, \boldsymbol{x}_{\mathbb{D}_{\mathcal{S}^i_\pi}})|\,\mathrm{do}(\mathbf{e}_{\mathcal{S}^i_\pi} = e_{\mathcal{S}^i_\pi})]\}.\end{aligned}$$

To calculate $\mathbb{E}[f(\mathbf{x}_{\mathbb{V}_{\mathcal{S}}}, \boldsymbol{x}_{\mathbb{D}_{\mathcal{S}}})|\,\mathrm{do}(\mathbf{e}_{\mathcal{S}} = e_{\mathcal{S}})]$, we should determine how to get the distribution of $\mathbf{x}_{\mathbb{V}_{\mathcal{S}}}$. For a given directed acyclic graph, the distribution of the data $\mathbf{x}$ generated by this graph satisfies the Markov property. Denote the distribution of the data as $P(\mathbf{x})$, which can be expressed as follows

$$P(\mathbf{x}) = \Pi_{i\in\mathbb{V}}P(\mathbf{x}_i|\mathbf{x}_{pa(i)}),$$

where $pa(i)$ represents the set of parent vertices of i in the graph. When we intervene in the generation of data, the conditional distribution of each feature needs to be changed accordingly. The distribution formula of $\mathbf{x}_{\mathbb{V}_{\mathcal{S}}}$ after the intervention is

$$P(\mathbf{x}_{\mathbb{V}_{\mathcal{S}}}|\,\mathrm{do}(\mathbf{e}_{\mathcal{S}} = e_{\mathcal{S}})) = \Pi_{i\in\mathbb{V}_{\mathcal{S}}}P(\mathbf{x}_i|\mathbf{x}_{pa(i)\cap\mathbb{V}^i_{\mathcal{S}}}, \boldsymbol{x}_{pa(i)\cap\mathbb{D}^i_{\mathcal{S}}}),$$

where $\mathbb{D}^i_{\mathcal{S}}$ is the set of features that have a direct influence on the value of feature i and are connected to i by edges in $\mathbf{e}_{\mathcal{S}}$. $\mathbb{V}^i_{\mathcal{S}}$ is the complement of $\mathbb{D}^i_{\mathcal{S}}$. Since the permutations used to calculate the marginal contributions follow the topological order, the distributions of all parent vertices are already determined when calculating the distribution of a particular feature. Therefore, we can sequentially compute the distributions of all vertices. After calculating the marginal contribution for all topological orderings, a weighted average can be used to obtain the attribution value for each edge. Then, by summing the attribution values of all outgoing edges of a vertex, we can obtain the attribution value for the feature point.

4.2 Approximation of Edge Intervention Causal Shapley Value

In Section 4.1, we give the way to infer the distribution of each feature. However, we generally cannot accurately obtain the specific distribution expressions for each feature for real datasets. Therefore, we use approximation methods based on data sampling to estimate the value of each feature, rather than directly solving the exact form of the distribution. The detailed process is shown in Algorithm 1.

Since the computing Shapley value is an #P-hard problem (Deng & Papadimitriou, 1994; Zhang et al., 2023), the computational cost becomes very high when the number of edges increases. Therefore, we propose an approximation method based on Monte Carlo sampling for DAG-SHAP. The detailed process is shown in Algorithm 2. We first enumerate edge permutations that comply with the

Algorithm 1 Sampling $\mathbf{x}_{\mathbb{V}_S}$ under edge intervention.

Require: Edge permutation π, intervention edge set size s, explained input $\boldsymbol{x}$, baseline input $\mathcal{D}$.
Ensure: A sample of $\mathbf{x}_{\mathbb{V}_S}$, denoted as $\hat{\mathbf{x}}_{\mathbb{V}_S}$.
1: Calculate the in-degree of each vertex in the graph and store it in the array $\boldsymbol{d}$;
2: Randomly initialize $\hat{\mathbf{x}}_{\mathbb{V}_S}$ based on $\mathcal{D}$;
3: **for** $i = 1$ to length of π **do**
4: Let c denote the child vertex of edge $\pi(i)$;
5: Let p denote the parent vertex of edge $\pi(i)$;
6: **if** i smaller than s **then**
7: Assign value $\boldsymbol{x}_p$ to edge $p \to c$;
8: **end if**
9: $\boldsymbol{d}_j \leftarrow \boldsymbol{d}_j - 1$;
10: **if** $\boldsymbol{d}_j$ equals 0 **then**
11: Infer the value of $\hat{\mathbf{x}}_j$ on the values of the edges between it and its parent vertices;
12: **for** each child vertex k of vertex j **do**
13: Assign value $\hat{\mathbf{x}}_j$ to edge $j \to k$;
14: **end for**
15: **end if**
16: **end for**
17: **return** $\hat{\mathbf{x}}_{\mathbb{V}_S}$;

topological sort and then intervene on each edge to compute its marginal contribution (Lines 6-9). After obtaining the marginal contributions, we compute the weighted average to estimate the attribution value of the edge (Lines 13-14). Finally, we sum the attribution values of the outgoing edges of each feature to get the overall edge attribution estimation (Lines 16-19).

Algorithm 2 Approximation of DAG-SHAP.

Require: Graph $\mathcal{G}$, vertices $\mathbb{V}$, edges E, explained input $\boldsymbol{x}$, sampling number T.
Ensure: Approximate attribution values of each edges, approximate attribution values of each feature.
1: Initialize counter $cnt \leftarrow 0$;
2: **for** $t = 1$ to T **do**
3: let π be a random permutation of E;
4: **if** π is a valid toposort **then**
5: $u \leftarrow 0$, $cnt \leftarrow cnt + 1$;
6: **for** $i = 0$ to length of π **do**
7: Get $\hat{\mathbf{x}}_{\mathbb{V}_S}$ through intervention on edges $\{\pi(0), \cdots, \pi(i)\}$ via Algorithm 1;
8: $\overline{\Psi(\pi(i))} \leftarrow \overline{\Psi(\pi(i))} + f(\hat{\mathbf{x}}_{\mathbb{V}_S}, \hat{\mathbf{x}}_{\mathbb{D}_S}) - u$;
9: $u \leftarrow f(\hat{\mathbf{x}}_{\mathbb{V}_S}, \hat{\mathbf{x}}_{\mathbb{D}_S})$;
10: **end for**
11: **end if**
12: **end for**
13: **for** $i = 1$ to m **do**
14: $\overline{\Psi(i)} \leftarrow \overline{\Psi(i)}/cnt$;
15: **end for**
16: **for** $i = 1$ to n **do**
17: Get $\overline{\Phi(i)}$ by summing the attribution values of its ongoing edges;
18: **end for**
19: **return** $\overline{\Psi(1)}, \cdots, \overline{\Psi(m)}, \overline{\Phi(1)}, \cdots, \overline{\Phi(n)}$;

Theorem 4. *Denote by τ the sampling number of valid permutations, the asymptotic upper bound of $\mathbb{P}(|\overline{\Phi(i)} - \Phi(i)| \geq \epsilon)$ sampling based on Algorithm 2 is $\mathcal{O}(m \cdot \exp(-\frac{1}{\tau m^2}))$. The proof can be found in appendix C.*

5 EXPERIMENTS

To show the superiority of DAG-SHAP over existing methods, we conduct experiments on both synthetic and real datasets. We present the attribution values of the benchmark, DAG-SHAP, and the baselines using bar charts. Additionally, we calculate the mean absolute error between DAG-SHAP and the benchmark, as well as between each baseline and the benchmark. We utilize a Mixture

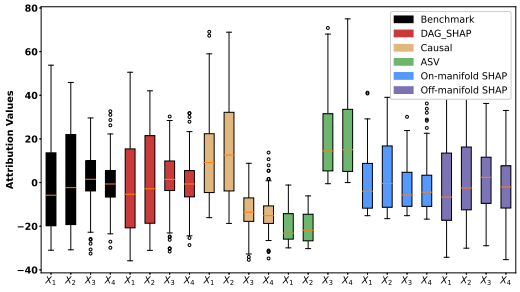


Figure 2: Attribution results of synthetic dataset.

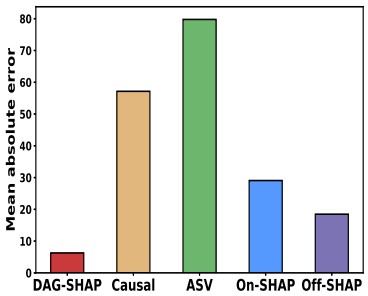
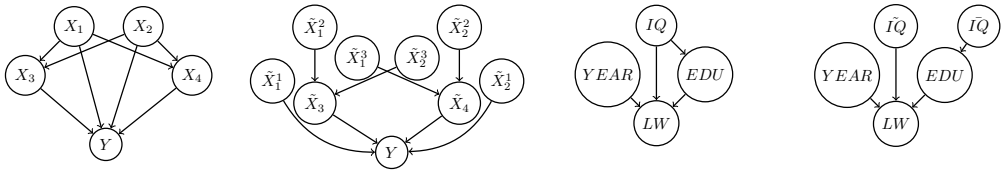


Figure 3: Mean absolute error.

Density Network to predict the distribution of child vertices after intervention on parent vertices within the input features, consistent with the approach used in causal Shapley value (Heskes et al., 2020). We experiment with two models for predicting the target feature from the input features: a deep neural network (DNN) and an XGBoost model. The experimental results for the XGBoost model are provided in the appendix E.1 due to the limited space.

5.1 Experiments on Synthetic Dataset

Data Generation Process. Consider a synthetic dataset consisting of four input features, X_1, X_2, X_3, and X_4, along with a target feature Y. The features X_1 and X_2 are independently sampled from two exogenous continuous random variables, $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. Both $\mathbf{x}_1$ and $\mathbf{x}_2$ follow a uniform distribution over the interval $[0, 10]$, i.e., $\mathbf{x}_1, \mathbf{x}_2 \sim U(0, 10)$. X_3 is influenced by X_1, X_2, and an additional exogenous variable $\mathbf{x}_3 \sim U(0, 10)$, and given by $X_3 = X_1 + X_2 + \mathbf{x}_3$. Similarly, X_4 is influenced by X_1 ,X_2, and an another exogenous variable $\boldsymbol{x}_4 \sim U(0, 10)$, and given by $X_4 = X_1 + X_2 + \mathbf{x}_4$. Finally, the target feature Y is given by $Y = X_1 \cdot X_3 + X_2 \cdot X_4$. The causal structure of this data generation process is shown in Figure 4(a).



(a) DAG of synthetic dataset. (b) DAG of splitted synthetic dataset. (c) DAG of Griliches76. (d) DAG of splitted Griliches76.

Figure 4: DAGs of used synthetic and real datasets.

Desirable Attribution. To get desirable benchmark attribution, we conduct vertex splitting for casual structure in Figure 4(b) such that asymmetric causal Shapley value can simulate attribution results satisfying all desirable properties in Section 3.2. Specifically, we split the influence of X_1 and X_2 on $X_3, X_4,$ and Y. Denote $\tilde{X}_1^1$, $\tilde{X}_1^2$, and $\tilde{X}_1^3$ as copies of X_1, while $\tilde{X}_2^1$, $\tilde{X}_2^2$, and $\tilde{X}_2^3$ are copies of X_2. The features $\tilde{X}_3$ and $\tilde{X}_4$ correspond to X_3 and X_4, respectively. Specifically, $\tilde{X}_3$ can be expressed as $\tilde{X}_3 = \tilde{X}_1^2 + \tilde{X}_2^2 + \tilde{\mathbf{x}}_3$ and $\tilde{X}_4 = \tilde{X}_1^3 + \tilde{X}_2^3 + \tilde{\mathbf{x}}_4$, where $\tilde{\mathbf{x}}_3 = \mathbf{x}_3$ and $\tilde{\mathbf{x}}_4 = \mathbf{x}_4$. The target feature is defined as $\tilde{Y} = \tilde{X}_1^1 \cdot \tilde{X}_3 + \tilde{X}_2^1 \cdot \tilde{X}_4$. The DAG is illustrated in Figure 4(b), following the data generation process. A key advantage of this new dataset is that the attributions for edges and vertices are equal, thereby eliminating externality issues for asymmetry causal Shapley value, as each vertex has only one outgoing edge. We train a ML model to predict labels based on features $X_1, X_2, X_3,$ and X_4. This model can also predict labels when provided with values of $\tilde{X}_1^1, \tilde{X}_2^1, \tilde{X}_3,$ and $\tilde{X}_4$. Attribution is conducted using the ML model on $\tilde{\boldsymbol{x}}$ with asymmetric causal Shapley value, where the values of $\tilde{X}_1^2, \tilde{X}_2^2, \tilde{X}_1^2, \tilde{X}_2^2$ are not model inputs but serve to intervene on $\tilde{X}_1^1, \tilde{X}_2^1, \tilde{X}_3, \tilde{X}_4$. The attribution results can benchmark the input tuple $\boldsymbol{x}$ in the original dataset, where $\boldsymbol{x}_1 = \tilde{\boldsymbol{x}}_1^1 = \tilde{\boldsymbol{x}}_1^2 = \tilde{\boldsymbol{x}}_1^3$ and $\boldsymbol{x}_2 = \tilde{\boldsymbol{x}}_2^1 = \tilde{\boldsymbol{x}}_2^2 = \tilde{\boldsymbol{x}}_2^3$. This is because the sum of attribution values for $\tilde{\boldsymbol{x}}_1^1, \tilde{\boldsymbol{x}}_1^2, \tilde{\boldsymbol{x}}_1^3$ should be equal to the attribution value of $\boldsymbol{x}_1$, and similarly for $\tilde{\boldsymbol{x}}_2^1, \tilde{\boldsymbol{x}}_2^2, \tilde{\boldsymbol{x}}_2^3$ with $\boldsymbol{x}_2$. Moreover, the attribution value of $\tilde{\boldsymbol{x}}_3$ must be equal to that of $\boldsymbol{x}_3$, and likewise for $\tilde{\boldsymbol{x}}_4$ and $\boldsymbol{x}_4$, in

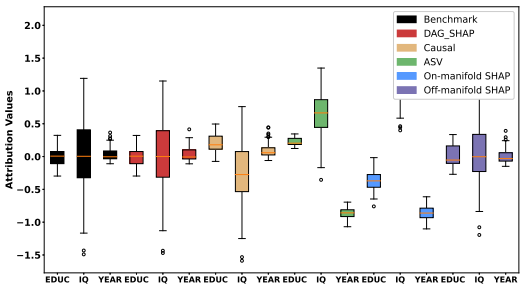


Figure 5: Attribution results of Griliches76 dataset.

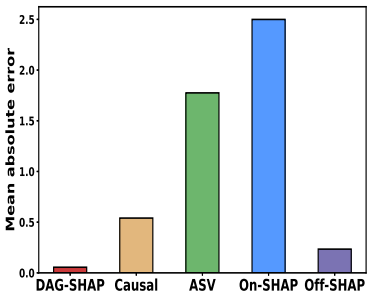


Figure 6: Mean absolute error.

accordance with implementation invariance (Sundararajan et al., 2017) as they map the original input of the model.

Experimental results. ASV represents both the asymmetric Shapley value and the asymmetric causal Shapley value of the baseline, as they are equivalent given a DAG structure. To simplify, we use 'Causal' to denote the symmetric causal Shapley value in the legend of Figure 2. The axis labels X_1, X_2, X_3, X_4 of benchmark correspond to the sum of the split nodes. Based on the data generation process, it is clear that the influence of X_1 and X_2 on Y should be greater than that of X_3 and X_4. That is, the range of attribution values for features X_1 and X_2 should be greater than X_3 and X_4. However, the results from off-manifold SHAP, on-manifold SHAP, and asymmetry SHAP with a DNN contradict this as shown in Figure 2 , as they fail to identify the impact of source vertices on their child vertices. We also conduct a statistical analysis of the attribution error for each method. The experimental results show that the DAG-SHAP values calculated on Figure 4(a) are approximately equal to the sum of the DAG-SHAP values calculated on Figure 4(b), and are also approximately equal to the attribution values in benchmark. The attribution error of other methods is significantly greater than that of DAG-SHAP as shown in Figure 3.

5.2 Experiments on Real Datasets

Experiments on Griliches76 Dataset. The first real dataset we used is the Griliches76 dataset (Griliches, 1976), consisting of 758 entries gathered from the U.S. labor market. This dataset is widely used in research to explore the impact of features on income. We selected three features: IQ, education level (EUD), and years working at the current unit (YEAR). The target variable is the logarithm of weekly income (LW). We use the same way to create the benchmark as the synthetic dataset. The original causal graph is shown in Figure 4(c) and the splitted graph is shown in Figure 4(d). We train a ML model with input features IQ, YEAR, and LW. The attribution results with a DNN of each method are shown in Figure 5. The experimental results show that the attribution value of DAG-SHAP is the closest to the benchmark, with the mean absolute error only 23.7% of that of the second smallest method, Off-manifold SHAP which is shown in Figure 6.

Experiments on Census Income Dataset. The second real dataset we used is the Adult dataset (Dua & Graff, 2017) with the causal graph shown in Figure 7. We train a binary classifier to predict whether the income of one individual exceeds $50,000 per year. We also split the direct and indirect effects of Country, Race, and Age to create a new dataset. We use the same way in the above experiments to get benchmark attribution values. We show the attribution results with DNN in Figure 9 in the appendix due to the limited space. The experimental results show that the attribution value of DAG-SHAP is the closest to the benchmark, with the mean absolute error only 75.1% of that of the second smallest method, causal Shapley value which is shown in Figure 8.

6 Conclusion

In the paper, we present an innovative approach to feature attribution by incorporating causal relationships within a DAG structure for data generation. We introduce an edge intervention method that strategically targets specific child vertices via their parent vertices to control the fine-grained causal influence. Our method uniquely addresses the limitations of traditional feature attribution

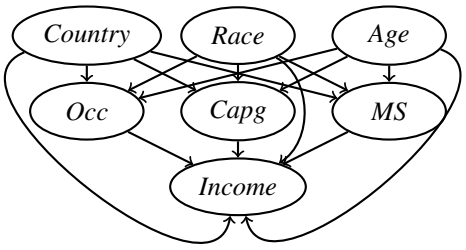


Figure 7: DAG of the Census Income dataset.

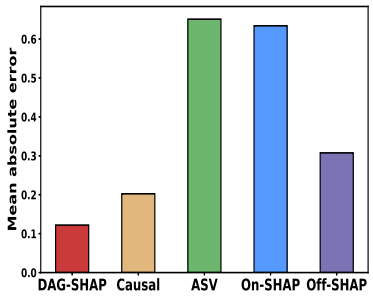


Figure 8: Mean absolute error.

techniques by capturing both the exogenous contributions of edges and the requisite externality conditions in model interpretations. Future research could expand on refining and optimizing computational strategies for DAG-SHAP to enhance its applicability in more complex and larger datasets. Additionally, it is interesting to explore integrating our method with models that handle time-dependent data, potentially opening new avenues for real-time decision-making applications.

REFERENCES

Amnon Catav, Boyang Fu, Jason Ernst, Sriram Sankararaman, and Ran Gilad-Bachrach. Marginal contribution feature importance–an axiomatic approach for the natural case. *arXiv preprint arXiv:2010.07910*, 2020.

Hugh Chen, Scott M Lundberg, and Su-In Lee. Explaining a series of models by propagating shapley values. *Nature communications*, 13(1):4512, 2022.

Xuexin Chen, Ruichu Cai, Zhengting Huang, Yuxuan Zhu, Julien Horwood, Zhifeng Hao, Zijian Li, and José Miguel Hernández-Lobato. Feature attribution with necessity and sufficiency via dual-stage perturbation test for causal explanation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6486–6502. PMLR, 21–27 Jul 2024.

Xiaotie Deng and Christos H Papadimitriou. On the complexity of cooperative solution concepts. *Mathematics of operations research*, 19(2):257–266, 1994.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in neural information processing systems*, 33:1229–1239, 2020.

Zvi Griliches. Wages of very young men. *Journal of Political Economy*, 84(4, Part 2):S69–S85, 1976.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Joseph Janssen, Vincent Guan, and Elina Robeva. Ultra-marginal feature importance: Learning from data with causal guarantees. In *International conference on artificial intelligence and statistics*, pp. 10782–10814. PMLR, 2023.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pp. 2907–2916. PMLR, 2020.

Dominik Janzing, Patrick Blöbaum, Atalanti A Mastakouri, Philipp M Faller, Lenon Minorics, and Kailash Budhathoki. Quantifying intrinsic causal contributions via structure preserving interventions. In *International Conference on Artificial Intelligence and Statistics*, pp. 2188–2196. PMLR, 2024.

Yonghan Jung, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Blöbaum, and Elias Bareinboim. On measuring causal contributions via do-interventions. In *International Conference on Machine Learning*, pp. 10476–10501. PMLR, 2022a.

Yonghan Jung, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Bloebaum, and Elias Bareinboim. On measuring causal contributions via do-interventions. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10476–10501. PMLR, 17–23 Jul 2022b.

Jinkun Lin, Anqi Zhang, Mathias Lécuyer, Jinyang Li, Aurojit Panda, and Siddhartha Sen. Measuring the effect of training data on deep learning predictions via randomized experiments. In *International Conference on Machine Learning*, pp. 13468–13504. PMLR, 2022.

Christoph Luther, Gunnar König, and Moritz Grosse-Wentrup. Efficient sage estimation via causal structure learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 11650–11670. PMLR, 2023.

Gautam Machiraju, Alexander Derry, Arjun D Desai, Neel Guha, Amir-Hossein Karimi, James Zou, Russ B Altman, Christopher Re, and Parag Mallick. Prospector heads: Generalized feature attribution for large models amp; data. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34115–34144. PMLR, 21–27 Jul 2024.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

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.

M Scott, Lee Su-In, et al. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30:4765–4774, 2017.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Lloyd S Shapley. A value for n-person games. *Contribution to the Theory of Games*, 2, 1953.

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

Raghav Singal, George Michailidis, and Hoiyi Ng. Flow-based attribution in graphical models: A recursive shapley approach. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9733–9743. PMLR, 2021.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

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 *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Yair Tauman. The aumann-shapley prices: a survey. *The shapley value*, pp. 279, 1988.

Lucile Ter-Minassian, Oscar Clivio, Karla Diazordaz, Robin J Evans, and Christopher C Holmes. Pwshap: a path-wise explanation model for targeted variables. In *International Conference on Machine Learning*, pp. 34054–34089. PMLR, 2023.

Jiaxuan Wang, Jenna Wiens, and Scott M. Lundberg. Shapley flow: A graph-based approach to interpreting model predictions. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 721–729. PMLR, 2021.

Eyal Winter. The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054, 2002.

Jiayao Zhang, Qiheng Sun, Jinfei Liu, Li Xiong, Jian Pei, and Kui Ren. Efficient sampling approaches to shapley value approximation. *Proceedings of the ACM on Management of Data*, 1(1):1–24, 2023.

APPENDIX

In the appendix of our paper, we provide comprehensive additional content. Section A reviews the related work. In Section B, we discuss why DAG-SHAP satisfies all the desirable properties. In Section C, we provide the proof of Theorem 4. Subsequently, Section E offers supplementary material related to our experimental procedures.

A RELATED WORK

In the related work section, we organize the content into three parts. First, we provide a detailed introduction to the baselines and the comparative methods used in our study. Next, we review classical feature attribution approaches. Finally, we discuss feature attribution methods related to causal inference.

A.1 SHAPLEY VALUE-BASED FEATURE ATTRIBUTION METHODS

On-manifold Shapley value is proposed by Scott et al. (2017), it is defined as follows

$$\phi_i = \sum_{j=0}^{n-1} \frac{1}{n\binom{n-1}{j}} \sum_{i \notin \mathcal{S}, |\mathcal{S}|=j} \{\mathbb{E}\left[f\left(\boldsymbol{x}_{\mathcal{S}\cup\{i\}}, \mathbf{x}_{\overline{\mathcal{S}\cup\{i\}}}\right) \mid \mathbf{x}_{\mathcal{S}\cup\{i\}} = \boldsymbol{x}_{\mathcal{S}\cup\{i\}}\right] - \mathbb{E}\left[f\left(\boldsymbol{x}_{\mathcal{S}}, \mathbf{x}_{\bar{\mathcal{S}}}\right) \mid \mathbf{x}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}\right]\}.$$

The marginal contribution of feature i in the explained input $\boldsymbol{x}$ when cooperating with a coalition $\mathcal{S}$ is expressed as $\mathbb{E}\left[f\left(\boldsymbol{x}_{\mathcal{S}\cup\{i\}}, \mathbf{x}_{\overline{\mathcal{S}\cup\{i\}}}\right) \mid \mathbf{x}_{\mathcal{S}\cup\{i\}} = \boldsymbol{x}_{\mathcal{S}\cup\{i\}}\right] - \mathbb{E}\left[f\left(\boldsymbol{x}_{\mathcal{S}}, \mathbf{x}_{\bar{\mathcal{S}}}\right) \mid \mathbf{x}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}\right]$, where $\overline{\mathcal{S}\cup\{i\}}$ represents the complement of $\mathcal{S}\cup\{i\}$. The values of features not in the coalitions are determined under conditional expectations of features in the coalitions.

Off-manifold Shapley value (Scott et al., 2017), a simplified version of on-manifold Shapley value by assuming independence of features for implementation, which is widely adopted in practical applications. The marginal contribution in off-manifold Shapley value is defined as $\mathbb{E}\left[f\left(\boldsymbol{x}_{\mathcal{S}\cup\{i\}}, \mathbf{x}_{\overline{\mathcal{S}\cup\{i\}}}\right)\right] - \mathbb{E}\left[f\left(\boldsymbol{x}_{\mathcal{S}}, \mathbf{x}_{\bar{\mathcal{S}}}\right)\right]$. The values of features $\overline{\mathcal{S}\cup\{i\}}$ and $\overline{\mathcal{S}}$ which are not in the selected coalitions are determined by the average values in the dataset.

Asymmetry Shapley value (Frye et al., 2020) sets the weight of permutations that do not satisfy the topological order to zero, thereby breaking the symmetry in Shapley value. This allows for model explainability analysis to distinguish the causal order between features when considering their contributions. It accounts for the causal relationships between features, rather than assuming that all features affect the model's output symmetrically and independently.

Causal Shapley value (Heskes et al., 2020) incorporates priori causal knowledge, using the intervention in the causal area to represent the interaction influence between features. The marginal contribution is expressed as $\mathbb{E}\left[f\left(\boldsymbol{x}_{\mathcal{S}\cup\{i\}}, \mathbf{x}_{\overline{\mathcal{S}\cup\{i\}}}\right) \mid \mathrm{do}\left(\mathbf{x}_{\mathcal{S}\cup\{i\}} = \boldsymbol{x}_{\mathcal{S}\cup\{i\}}\right)\right] - \mathbb{E}\left[f\left(\boldsymbol{x}_{\mathcal{S}}, \mathbf{x}_{\bar{\mathcal{S}}}\right) \mid \mathrm{do}\left(\mathbf{x}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}}\right)\right]$, where operator $\mathrm{do}(\cdot)$ means setting a variable to a specific value. The intervention is conducted on the feature vertex so the influence of the parent vertex will transfer to all its child vertices simultaneously if it is intervened. Note that the causal Shapley value can adopt both symmetric and asymmetric permutations, referred to as symmetry causal Shapley value and asymmetry causal Shapley value, respectively. Asymmetry causal Shapley value is equivalent to the asymmetry Shapley value when DAG is known.

Shapley Flow (Wang et al., 2021) is a method that attributes based on paths, ensuring path order consistency through depth-first search. It sums the attribution values of all paths passing through an edge as the edge's attribution, ensuring that the sum of the attribution values of all edges in any cut in the causal graph is the same. Formally, the Shapley value of path i can be represented as follows

$$\tilde{\phi}_\nu(i) = \sum_{\pi \in \Pi_{\mathrm{dfs}}} \frac{\tilde{v}([j : \pi(j) \leq \pi(i)]) - \tilde{v}([j : \pi(j) < \pi(i)])}{|\Pi_{\mathrm{dfs}}|},$$

where inequality $\pi(j) \leq \pi(i)$ denotes that path j precedes path i under ordering $\pi \in \Pi_{\mathrm{dfs}}$. To obtain the attribution of an edge e, they propose to sum the attributions of all paths $\mathbb{P}$ that contains e, i.e., $\phi_v(e) = \sum_{p \in \mathbb{P}, e \in p} \tilde{\phi}_v(p)$.

Recursive Shapley value (Singal et al., 2021) conducts a top-down feature attribution process to quantify how changes at the source vertices propagate through the graph. It assumes that only the vertices with no incoming edges have exogenous contributions, and the relationship between intermediate vertices and the source vertices is deterministic, with intermediate vertices only transmitting the effects of source vertices.

A.2 Class Feature Attribution Techniques

LIME (Local Interpretable Model-agnostic Explanations) is a popular model explanation method proposed by (Ribeiro et al., 2016) in 2016. Its core idea is to generate a local, interpretable model for each prediction made by a complex model. The advantage of LIME lies in its universality and simplicity, as it can be applied to any model and provides intuitive explanations of feature importance. Grad-CAM (Selvaraju et al., 2017) highlights the areas of the image that contribute the most to the prediction of a specific class through visualization. Specifically, it uses the feature maps of the last convolutional layer and the gradient information regarding the prediction of a specific class to generate a Class Activation Map. DeepLIFT (Shrikumar et al., 2017) explains the decision-making process of deep learning models through the differences in activation functions. The DeepLIFT method allocates contribution scores by comparing the activation of each neuron with its "reference activation," thereby revealing the contribution of input features to the model's prediction. SmoothGrad Smilkov et al. (2017) is a method designed to improve the interpretability of deep learning models, particularly for image classification tasks. It aims to enhance the quality of model prediction explanations by reducing the visual noise in gradient sensitivity maps. The core idea of this method is to add random noise to the input image to create multiple perturbed images, and then average the gradient sensitivity maps of these perturbed images to smooth the original sensitivity map. Integrated Gradients (IG) (Tauman, 1988) involves integrating the gradients of a model's output with respect to its inputs along the path from a baseline to the actual input data. It satisfies important interpretability axioms such as Sensitivity, Implementation Invariance, and Completeness, and is independent of the specific implementation details of the model. Layer-wise Relevance Propagation (LRP) (Montavon et al., 2019) is a popular method for explaining neural networks. It explains the predictions of neural networks by propagating relevance scores from the output layer back to the input layer. MCI measures feature contribution based on the maximum marginal contribution a feature can bring which cannot capture the right causal contributin. It gives some new properties like Super-efficiency and Sub-additivity (Catav et al., 2020).

A.3 Causal Feature Attribution Techniques

The Shapley Additive Global Importance (SAGE) value is a globally attributed explainability method. d-SAGE (Luther et al., 2023)observes the conditional independence between features and the model target, and utilizes Causal Structure Learning (CSL) to infer a graph that encodes (conditional) independence in the data as d-separations, which is more computationally efficient for calculating SAGE. do-Shapley value (Jung et al., 2022b), as a measure of causal contribution, provides a theoretical justification through axiomatic foundations. Like causal Shapley value, it employs interventions but generalizes previous approaches to measure the causal contributions of each feature to a target effect induced by a black-box/unknown/inaccessible model. PWSHAP (Path-Wise Shapley effects) (Ter-Minassian et al., 2023) is a method for explaining the impact of specific binary variables, such as treatment effects or ethnicity in policy models, within predictive models. It evaluates the targeted effects in complex outcome models by combining a predictive model with a user-defined Directed Acyclic Graph (DAG). Inspired by causal inference and randomized experiments, researchers have developed an algorithm to estimate AME (Average Marginal Effect), a measure of the expected (average) marginal effect of adding a data point to a subset of the training data sampled from a uniform distribution. When subsets are sampled from a uniform distribution, AME simplifies to the well-known Shapley value (Lin et al., 2022). CF-SHAP (**?**) is a method that combines counterfactual information for feature attribution. It strengthens and clarifies the link between actionable recourse and feature attributions, playing a role in advancing the development of causal feature attribution. Dominik et al. (Janzing et al., 2020) primarily focus on distinguishing between calculating Shapley values based on observational versus interventional conditional distributions. They were among the first to emphasize the role of causality in feature attribution. Besides, Dominik et al. (Janzing et al., 2024) employs structure-preserving interventions to attribute uncertainty (e.g., variance or Shannon entropy) in the target to its influencing features. Their approach can be seen as focusing on

feature interventions that assess changes in uncertainty metrics. UMFI removes the dependencies of the feature being attributed from other features before calculating its marginal contribution. As a result, it requires a causal graph to identify these dependencies. This is because, without such a graph, determining the dependencies would not be possible. Additionally, since UMFI relies on an approximately optimally preprocessed feature set for its computations, it does not satisfy properties like efficiency and additivity, which are inherent to the original Shapley value (Janssen et al., 2023).

B PROPERTIES

In this section, we introduce the definitions of several fundamental properties and then explain why DAG-SHAP possesses properties that should be satisfied.

B.1 FUNDAMENTAL PROPERTIES

Linearity, Implementation Invariance, Sensitivity, and Dummy Feature are fundamental to feature attribution. When edges are the objects of attribution, the definitions of these properties are as follows:

Linearity: For any two value functions f_u and f_w, and their sum $f = f_u + f_w$, the attributed value of each edge in f is equal to the sum of its attributed values in f_u and f_w, which is $\Psi_f(e_i) = \Psi_{f_u}(e_i) + \Psi_{f_w}(e_i)$.
Implementation Invariance: The attribution results from an attribution method will be identical for two models if their outputs are the same for all inputs, even though their implementations may differ significantly.
Sensitivity: If two inputs differ in only one feature (edge) and the model's predictions for these two inputs are different, then the attribution for that differing feature (edge) should be non-zero. Formally, let $\boldsymbol{x}$ and $\boldsymbol{x}'$ be two inputs that differ only in edge e_i, and the corresponding model predictions satisfy $f(\boldsymbol{x}) \neq f(\boldsymbol{x}')$. Then, the attribution $\Psi(e_i)$ should satisfy $\Psi(e_i) \neq 0$.
Dummy: The attributed value of a feature is zero if it does not have exgenous influence to the outcome.

B.2 WHY DAG-SHAP SATISFIES THESE PROPERTIES

Proof of Linearity. For any edge e_i, its edge intervention causal Shapley value is given by

$$\Psi_f(e_i) = \sum_{\pi \in \Pi} \frac{1}{|\Pi|} \{\mathbb{E}[f(\mathbf{x})|\,\mathrm{do}(\mathbf{e}_{\underline{S}^i_\pi} = e_{\underline{S}^i_\pi})] - \mathbb{E}[f(\mathbf{x})|\,\mathrm{do}(\mathbf{e}_{S^i_\pi} = e_{S^i_\pi})]\}. \tag{5}$$

Substituting $f = f_u + f_w$, we can get

$$\Psi_f(e_i) = \sum_{\pi \in \Pi} \frac{1}{|\Pi|} \{\mathbb{E}[f_u(\mathbf{x}) + f_w(\mathbf{x})|\,\mathrm{do}(\mathbf{e}_{\underline{S}^i_\pi} = e_{\underline{S}^i_\pi})] - \mathbb{E}[f_u(\mathbf{x}) + f_w(\mathbf{x})|\,\mathrm{do}(\mathbf{e}_{S^i_\pi} = e_{S^i_\pi})]\}. \tag{6}$$

This can be split into two parts:

$$\Psi_f(e_i) = \sum_{\pi \in \Pi} \frac{1}{|\Pi|} \{\mathbb{E}[f_u(\boldsymbol{x})|\,\mathrm{do}(\mathbf{e}_{\underline{S}^i_\pi} = e_{\underline{S}^i_\pi})] - \mathbb{E}[f_u(\boldsymbol{x})|\,\mathrm{do}(\mathbf{e}_{S^i_\pi} = e_{S^i_\pi})]\} \tag{7}$$

$$+ \sum_{\pi \in \Pi} \frac{1}{|\Pi|} \{\mathbb{E}[f_w(\boldsymbol{x})|\,\mathrm{do}(\mathbf{e}_{\underline{S}^i_\pi} = e_{\underline{S}^i_\pi})] - \mathbb{E}[f_w(\boldsymbol{x})|\,\mathrm{do}(\mathbf{e}_{S^i_\pi} = e_{S^i_\pi})]\}. \tag{8}$$

According to the definition of edge intervention causal Shapley values, the two summations correspond to $\Psi_{f_u}(e_i)$ and $\Psi_{f_w}(e_i)$, respectively. Thus, $\Psi_f(e_i) = \Psi_{f_u}(e_i) + \Psi_{f_w}(e_i)$. This proves that edge intervention causal Shapley satisfies the Linearity property. □

Proof of Implementation Invariance. If $f(\boldsymbol{x}) = g(\boldsymbol{x})$ for all inputs $\boldsymbol{x}$,the marginal contributions of edge e_i in f and g with any edge subsets $e_{S^i_\pi}$ are equal:$\mathbb{E}[f(\boldsymbol{x})|\,\mathrm{do}(\mathbf{e}_{\underline{S}^i_\pi} = e_{\underline{S}^i_\pi})] - \mathbb{E}[f(\boldsymbol{x})|\,\mathrm{do}(\mathbf{e}_{S^i_\pi} = e_{S^i_\pi})] = \mathbb{E}[g(\boldsymbol{x})|\,\mathrm{do}(\mathbf{e}_{\underline{S}^i_\pi} = e_{\underline{S}^i_\pi})] - \mathbb{E}[g(\boldsymbol{x})|\,\mathrm{do}(\mathbf{e}_{S^i_\pi} = e_{S^i_\pi})]$. This follows directly from the fact that

f and g produce identical outputs for any subset of features. The edge intervention causal Shapley value for an edge i is a weighted sum of the marginal contributions across all permutations within Π:

$$\Psi_f(e_i) = \sum_{\pi \in \Pi} \frac{1}{|\Pi|} \{\mathbb{E}[f(\mathbf{x})|\, \mathrm{do}(\mathbf{e}_{\underline{\mathcal{S}}^i_\pi} = e_{\underline{\mathcal{S}}^i_\pi})] - \mathbb{E}[f(\mathbf{x})|\, \mathrm{do}(\mathbf{e}_{\mathcal{S}^i_\pi} = e_{\mathcal{S}^i_\pi})]\}. \tag{9}$$

Substituting the equivalence of marginal contributions:

$$\Psi_g(e_i) = \sum_{\pi \in \Pi} \frac{1}{|\Pi|} \{\mathbb{E}[g(\mathbf{x})|\, \mathrm{do}(\mathbf{e}_{\underline{\mathcal{S}}^i_\pi} = e_{\underline{\mathcal{S}}^i_\pi})] - \mathbb{E}[g(\mathbf{x})|\, \mathrm{do}(\mathbf{e}_{\mathcal{S}^i_\pi} = e_{\mathcal{S}^i_\pi})]\}. \tag{10}$$

Thus, the edge intervention causal Shapley for f and g are identical for all edges e_i when $f(\boldsymbol{x}) = g(\boldsymbol{x})$ for all $\boldsymbol{x}$. This proves that edge intervention causal Shapley satisfies the Implementation Invariance property. □

Proof of Sensitivity. In each permutation π, the marginal contribution of edge e_i is:

$$\Delta^i_\pi = \mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_{\underline{\mathcal{S}}^i_\pi} = e_{\underline{\mathcal{S}}^i_\pi})] - \mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_{\mathcal{S}^i_\pi} = e_{\mathcal{S}^i_\pi})].$$

Since $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ only in e_i, and $f(\boldsymbol{x}) \neq f(\boldsymbol{x}')$, it means that changing e_i while keeping other edges fixed leads to a change in the prediction. Therefore, there exists at least one context (set of preceding edges $\mathcal{S}^i_\pi$) such that the marginal contribution $\Delta^i_\pi \neq 0$. Since at least one $\Delta^i_\pi \neq 0$, and DAG-SHAP averages these marginal contributions over all permutations, the overall attribution $\Psi(e_i)$ will be non-zero:

$$\Psi(e_i) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \Delta^i_\pi \neq 0.$$

Thus, DAG-SHAP assigns a non-zero attribution to edge e_i when it is the only differing edge between inputs that have different predictions, satisfying the Sensitivity property. □

Proof of Dummy. Since feature k has no exogenous influence on the outcome, intervening on its outgoing edges does not affect the expected model output. Therefore, for any subset $\mathcal{S} \subseteq E \setminus \{e_i\}$ where $i \in \mathcal{O}_k$:

$$\mathbb{E}\left[f(\mathbf{x}) \mid \mathrm{do}\left(\mathbf{e}_{\mathcal{S}\cup\{e_i\}} = e_{\mathcal{S}\cup\{e_i\}}\right)\right] = \mathbb{E}\left[f(\mathbf{x}) \mid \mathrm{do}\left(\mathbf{e}_{\mathcal{S}} = e_{\mathcal{S}}\right)\right].$$

This implies that the marginal contribution of any outgoing edge e_i is zero:

$$\Delta^i_\pi = \mathbb{E}\left[f(\mathbf{x}) \mid \mathrm{do}\left(\mathbf{e}_{\underline{\mathcal{S}}^i_\pi} = e_{\underline{\mathcal{S}}^i_\pi}\right)\right] - \mathbb{E}\left[f(\mathbf{x}) \mid \mathrm{do}\left(\mathbf{e}_{\mathcal{S}^i_\pi} = e_{\mathcal{S}^i_\pi}\right)\right] = 0.$$

Since the marginal contributions Δ^i_π are zero for all permutations π and all outgoing edges e_i, the DAG-SHAP attribution for each edge is:

$$\Phi(k) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \Delta^i_\pi = 0.$$

The total attribution for feature is the sum of the attributions of its outgoing edges:

$$\Phi(k) = \sum_{i \in \mathcal{O}_k} \Psi(e_i) = \sum_{i \in \mathcal{O}_k} 0 = 0.$$

Since all outgoing edges of feature k have zero attributions and the feature's attribution is the sum of these, the feature's total attribution is zero. Therefore, DAG-SHAP satisfies the Dummy property. □

Proof of Causality. In any valid topological order $\pi \in \Pi$, all edges respect the causal constraint: if $e_j \to e_k$, then e_j appears before e_k ($\pi(j) < \pi(k)$). Thus, when calculating $\Psi(e_i)$, all edges in $\mathcal{S}^i_\pi$ have already been intervened upon, ensuring that the causal effects of ancestor nodes are incorporated correctly, avoiding any reversal of causality. The intervention $\mathrm{do}(\mathbf{e}_{\underline{\mathcal{S}}^i_\pi} = e_{\underline{\mathcal{S}}^i_\pi})$ ensures that the model output depends only on the current edge e_i and the previously intervened edges in $\mathcal{S}^i_\pi$. -By computing the difference in expected model output, the calculation isolates the direct causal effect of e_i, excluding any indirect effects through other paths. Let the edge $e_i = (p_i, c_i, \boldsymbol{x}_{p_i})$, and its attribution value is defined as:

$$\Psi(e_i) = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \Delta^i_\pi,$$

where $\Delta_\pi^i = \mathbb{E}[f(\mathbf{x})|\,\mathrm{do}(\mathbf{e}_{\underline{\mathcal{S}}_\pi^i} = e_{\underline{\mathcal{S}}_\pi^i})] - \mathbb{E}[f(\mathbf{x})|\,\mathrm{do}(\mathbf{e}_{\mathcal{S}_\pi^i} = e_{\mathcal{S}_\pi^i})]$. Δ_π^i depends solely on the effect of the edge e_i under the current intervention, ensuring that it measures only the direct causal contribution of e_i. DAG-SHAP's attribution process strictly respects causal ordering and isolates the effects of individual edges through edge interventions. The resulting attribution values reflect the direct causal impact of features on the output, thus satisfying the causality property. □

Proof of Efficiency. The sum of attribution values over all edges is:

$$\sum_{e_i \in E} \Psi(e_i) = \sum_{e_i \in E} \sum_{\pi \in \Pi} \frac{1}{|\Pi|} \left[\mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_{\underline{\mathcal{S}}_\pi^i} = e_{\underline{\mathcal{S}}_\pi^i})] - \mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_{\mathcal{S}_\pi^i} = e_{\mathcal{S}_\pi^i})] \right].$$

Switching the order of summation over e_i and π:

$$\sum_{e_i \in E} \Psi(e_i) = \sum_{\pi \in \Pi} \frac{1}{|\Pi|} \sum_{e_i \in E} \left[\mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_{\underline{\mathcal{S}}_\pi^i} = e_{\underline{\mathcal{S}}_\pi^i})] - \mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_{\mathcal{S}_\pi^i} = e_{\mathcal{S}_\pi^i})] \right].$$

For a given permutation π, summing over all edges $e_i \in E$ creates a telescoping series:

$$\sum_{e_i \in E} \left[\mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_{\underline{\mathcal{S}}_\pi^i} = e_{\underline{\mathcal{S}}_\pi^i})] - \mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_{\mathcal{S}_\pi^i} = e_{\mathcal{S}_\pi^i})] \right] = \mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_E = e_E)] - \mathbb{E}[f(\mathbf{x})],$$

where $\mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_E = e_E)]$ is the model output when all edges are intervened, and $\mathbb{E}[f(\mathbf{x})]$ is the baseline output when no edges are intervened. Averaging over all permutations $\pi \in \Pi$ does not affect the telescoping result because the final and initial terms are the same for every permutation:

$$\sum_{\pi \in \Pi} \frac{1}{|\Pi|} \left[\mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_E = e_E)] - \mathbb{E}[f(\mathbf{x})] \right] = \mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_E = e_E)] - \mathbb{E}[f(\mathbf{x})].$$

Thus, the sum of the attribution values for all edges satisfies:

$$\sum_{e_i \in E} \Psi(e_i) = f(\boldsymbol{x}) - \mathbb{E}[f(\mathbf{x})].$$

This proves that DAG-SHAP satisfies the Efficiency property. □

Proof of Externality. Consider two vertices k and k': k' has a path to y, forming a path $k' \to y$, k has a path $k \to y$ independent of k'. When computing the marginal contribution of edges associated with k, the effects of $k' \to y$ are considered before intervening on $k \to y$. If $k' \to y$ is included in the intervention set $\mathcal{S}_\pi^i$ before $k \to y$, then the baseline for $k \to y$ is conditioned on the effects of $k' \to y$. In DAG-SHAP, all valid permutations $\pi \in \Pi$ are considered. For permutations where $k' \to y$ precedes $k \to y$, the cooperative effect is explicitly reflected in the marginal contribution. This summation ensures that the attribution value for k includes benefits from $k' \to y$ in all cases where k' precedes k. Therefore, DAG-SHAP satisfies the Externality property. □

Proof of Exogeneity. In DAG-SHAP, the independent contribution of feature k is derived by intervening on the edges $e_i = (k, c_i, \boldsymbol{x}_k)$ where $e_i \in \mathcal{O}_k$ while respecting the causal order defined by the directed acyclic graph (DAG). The attribution value $\Psi(e_i)$ is calculated as:

$$\Psi(e_i) = \sum_{\pi \in \Pi} \frac{1}{|\Pi|} \left[\mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_{\underline{\mathcal{S}}_\pi^i} = e_{\underline{\mathcal{S}}_\pi^i})] - \mathbb{E}[f(\mathbf{x}) \mid \mathrm{do}(\mathbf{e}_{\mathcal{S}_\pi^i} = e_{\mathcal{S}_\pi^i})] \right],$$

where $\underline{\mathcal{S}}_\pi^i = \mathcal{S}_\pi^i \cup \{e_i\}$ includes all preceding edges in the permutation π and the current edge e_k, $\mathcal{S}_\pi^i$ includes only the preceding edges, The expectation $\mathbb{E}[\cdot]$ measures the contribution of e_i under specific interventions, ensuring no confounding from subsequent edges. DAG-SHAP adheres to the topological order $\pi \in \Pi$, ensuring that for any edge $e_j \in \mathcal{O}_k$, the influences of all ancestor nodes of k (those connected via paths to p_k) are already fully propagated before e_i is intervened upon. This means $\Psi(e_j)$ reflects only the influence of k. By summing over all valid permutations $\pi \in \Pi$, the attribution value for e_j accounts for its independent contribution across all possible causal configurations. Nodes not causally connected to k do not influence $\Psi(e_j)$ because they are excluded from the intervention sets $\mathcal{S}_\pi^k$ and $\underline{\mathcal{S}}_\pi^k$. Ancestor nodes' effects are already fully propagated by the time e_j is intervened upon. Thus, DAG-SHAP satisfies the Exogeneity property. □

C PROOF

In this section, we give the proof of Theorem 4. Denote $1, \cdots, m$ the ongoing edges of feature vertex i. We have

$$\begin{aligned}\mathbb{P}(|\overline{\Phi(i)} - \Phi(i)| \geq \epsilon) \leq & \mathbb{P}(\sum_{j=1}^{m} |\overline{\Psi(j)} - \Psi(j)| \geq \epsilon) \\ \leq & \sum_{j=1}^{m} \mathbb{P}(|\overline{\Psi(j)} - \Psi(j)| \geq \frac{\epsilon}{m}) \\ \leq & 2m \exp(-\frac{2(\frac{\epsilon}{m})^2}{\sum_{k=1}^{\tau}(b_j - a_j)^2}) \leq 2m \exp(-\frac{2(\frac{\epsilon}{m})^2}{\tau r^2}),\end{aligned}$$

where (a_j, b_j) denotes the range of marginal contribution of edge j, and r is $\max(b_1 - a_1, \cdots, b_j - a_j)$. Equation (6) is derived from Hoeffding's inequality. Then, we have

$$\mathcal{O}(2m \exp(-\frac{2(\frac{\epsilon}{m})^2}{\tau r^2})) = \mathcal{O}(2m \exp(-\frac{2\epsilon^2}{m^2 \tau r^2})) = \mathcal{O}(m \cdot \exp(-\frac{1}{\tau m^2})).$$

Thus, we complete the proof.

D A COMPARATIVE STUDY OF DIFFERENT INTERVENTION METHODS

D.1 A COMPARISON OF EDGE INTERVENTION AND ASYMMETRIC SAMPLING NODE INTERVENTION

For asymmetric sampling node intervention, we use the following example to explain why it may fail. Let $X_1 = \mathbf{x}_1$, where $\mathbf{x}_1$ is a random variable uniformly distributed on $[0, 1]$, representing the exogenous influence of X_1; $X_2 = X_1 \cdot \mathbf{x}_2$, where $\mathbf{x}_2$ is another random variable uniformly distributed on $[0, 1]$, representing the exogenous influence of X_2. The generation of Y follows $Y = X_1 \cdot X_2$. In summary, X_1 directly influences Y and indirectly influences Y through X_2. X_2 influences Y with its own exogenous influence $\mathbf{x}_2$ and transfers the indirect influence of X_1. We aim to attribute values to each feature of a specific explained input $x^* = [x_1^*, x_2^*] = [1, 1]$ with respect to the baseline [0, 0].

For asymmetric sampling node interventions, the only valid sample permutation is (x_1^*, x_2^*). The marginal contributions of x_1^* and x_2^* in the permutation (x_1^*, x_2^*) are shown in the table below:

x_1^* in (x_1^*, x_2^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{x}_{\{1\}} = x_{\{1\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\emptyset\right)\right]$=1/2-0=1/2
x_2^* in (x_1^*, x_2^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{x}_{\{1,2\}} = x_{\{1,2\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{x}_{\{1\}} = x_{\{1\}}^*\right)\right]$ =1-1/2=1/2

Thus, the attribution value assigned to x_1^* by the asymmetric sampling node intervention is $1/2$, and the attribution value assigned to x_2^* is also $1/2$.

For edge intervention, we denote the edge $X_1 \to X_2$ as $\mathbf{e}_1$, $X_1 \to Y$ as $\mathbf{e}_2$, and $X_2 \to Y$ as $\mathbf{e}_3$. We denote the edges of the instance x^* as e_1^*, e_2^*, e_3^*. According to the definition of DAG-SHAP , there are three valid edge permutations: (e_1^*, e_2^*, e_3^*), (e_1^*, e_3^*, e_2^*), and (e_2^*, e_1^*, e_3^*). The marginal contributions of each edge in these permutations are as follows:

e_1^* in (e_1^*, e_2^*, e_3^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1\}} = e_1^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\emptyset\right)\right]$=0-0=0
e_1^* in (e_1^*, e_3^*, e_2^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1\}} = e_1^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\emptyset\right)\right]$=0-0=0
e_1^* in (e_2^*, e_1^*, e_3^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2\}} = e_{\{1,2\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{2\}} = e_2^*\right)\right]$ =1/2-0=1/2
e_2^* in (e_1^*, e_2^*, e_3^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2\}} = e_{\{1,2\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1\}} = e_1^*\right)\right]$=1/2-0=1/2
e_2^* in (e_1^*, e_3^*, e_2^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2,3\}} = e_{\{1,2,3\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,3\}} = e_{\{1,3\}}^*\right)\right]$ =1-0=1
e_2^* in (e_2^*, e_1^*, e_3^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{2\}} = e_2^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\emptyset\right)\right]$ =0-0=0
e_3^* in (e_1^*, e_2^*, e_3^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2,3\}} = e_{\{1,2,3\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2\}} = e_{\{1,2\}}^*\right)\right]$=1-1/2=1/2
e_3^* in (e_1^*, e_3^*, e_2^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,3\}} = e_{\{1,3\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1\}} = e_1^*\right)\right]$ =0-0=0
e_3^* in (e_2^*, e_1^*, e_3^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2,3\}} = e_{\{1,2,3\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2\}} = e_{\{1,2\}}^*\right)\right]$ =1-1/2=1/2

Thus, the attribution value assigned to x_1^* by DAG-SHAP is $(0+0+1/2)/3+(1/2+1+0)/3 = 2/3$,and the attribution value assigned to x_2^* is $(1/2+0+1/2)/3 = 1/3$. As $x_1^* = 1$ directly influences Y through X_2 and the interaction $Y = X_1 \cdot X_2$, it is evident that $x_1^* = 1$ is more important than x_2^*. Therefore, the asymmetric sampling node intervention provides incorrect attribution, as it fails to account for the external contribution of $x_1^* = 1$.

D.2 A Comparison of Edge Intervention and Symmetric Sampling Node Intervention

For symmetric intervention on the sampling nodes, we use this example to explain why it fails. $X_1 = \mathbf{x}_1$, where $\mathbf{x}_1$ is a random variable uniformly distributed on $[0, 1]$, representing the exogenous influence of X_1; $X_2 = X_1 + \mathbf{x}_2$, where $\mathbf{x}_2$ is another random variable uniformly distributed on $[0, 1]$, representing the exogenous influence of X_2. The generation of Y follows $Y = max(X_1, X_2)$. We want to attribute a value to each feature of a specific explained input $x^* = [x_1^*, x_2^*] = [1, 2]$ with respect to the baseline set as [0,0].

For symmetric node intervention, both the sampling permutations (x_1^*, x_2^*) and (x_2^*, x_1^*) are valid, and the marginal contributions of x_1^*, x_2^* in the permutations (x_1^*, x_2^*) and (x_2^*, x_1^*) are shown in the following table:

x_1^* in (x_1^*, x_2^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{x}_{\{1\}} = x_{\{1\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\emptyset\right)\right]$=3/2-0=3/2
x_1^* in (x_2^*, x_1^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{x}_{\{1,2\}} = x_{\{1,2\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{x}_{\{2\}} = x_2^*\right)\right]$=2-2=0
x_2^* in (x_2^*, x_1^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{x}_{\{2\}} = x_{\{2\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\emptyset\right)\right]$=2-0=2
x_2^* in (x_1^*, x_2^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{x}_{\{1,2\}} = x_{\{1,2\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{x}_{\{1\}} = x_{\{1\}}^*\right)\right]$ =2-3/2=1/2

Thus, the attribution value for x_1^* is $(3/2+0)/2 = 3/4$, while the attribution value for x_2^* is $(2+1/2)/2 = 5/4$.

For DAG-SHAP edge interventions, the edges for the instance x^* are denoted as e_1^*, e_2^*, e_3^*, and the marginal contributions of each edge in each arrangement are as follows:

e_1^* in (e_1^*, e_2^*, e_3^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1\}} = e_1^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\emptyset\right)\right]$=3/2-0=3/2
e_1^* in (e_1^*, e_3^*, e_2^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1\}} = e_1^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\emptyset\right)\right]$=3/2-0=3/2
e_1^* in (e_2^*, e_1^*, e_3^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2\}} = e_{\{1,2\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{2\}} = e_2^*\right)\right]$ =3/2-5/4=1/4
e_2^* in (e_1^*, e_2^*, e_3^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2\}} = e_{\{1,2\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1\}} = e_1^*\right)\right]$=3/2-3/2=0
e_2^* in (e_1^*, e_3^*, e_2^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2,3\}} = e_{\{1,2,3\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,3\}} = e_{\{1,3\}}^*\right)\right]$ =2-2=0
e_2^* in (e_2^*, e_1^*, e_3^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{2\}} = e_2^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\emptyset\right)\right]$ =5/4-0=5/4
e_3^* in (e_1^*, e_2^*, e_3^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2,3\}} = e_{\{1,2,3\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2\}} = e_{\{1,2\}}^*\right)\right]$=2-3/2=1/2
e_3^* in (e_1^*, e_3^*, e_2^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,3\}} = e_{\{1,3\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1\}} = e_1^*\right)\right]$ =2-3/2=1/2
e_3^* in (e_2^*, e_1^*, e_3^*)	$\mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2,3\}} = e_{\{1,2,3\}}^*\right)\right] - \mathbb{E}\left[\mathbf{Y} \mid \text{do}\left(\mathbf{e}_{\{1,2\}} = e_{\{1,2\}}^*\right)\right]$=2-3/2=1/2

Therefore, the attribution value for x_1 in DAG-SHAP is $(3/2+3/2+1/4)/3+(0+0+5/4)/3 = 3/2$, and the attribution value for x_2 is $(1/2+1/2+1/2)/3 = 1/2$. Similarly, because $x_1^* = 1$ directly influences X_2 and in turn directly influences $Y = \max(X_1, X_2)$ through X_1, it is clear that $x_1^* = 1$ is more important than x_2^*. Therefore, the attribution value provided by symmetric sampling node interventions is misleading, because it includes the contribution of x_1^* in the marginal contribution of X_2 based on an empty set, i.e., $\mathbb{E}[Y|\text{do}(X_2 = x_2^*)] - \mathbb{E}[Y|\text{do}(\emptyset)] = 2$.

D.3 A Comparison of Edge Intervention and Structure-preserving interventions

Structure-preserving interventions are based on the assumption that the distribution of exogenous variables for each feature is known, and then intervene on the exogenous variable of each feature accordingly. However, in reality, the distribution of exogenous variables is unknown. The authors use an additive structural causal model to fit the data generation process. By subtracting the distribution of the parent node from the distribution of the child node, they obtain the exogenous variable distribution

for the child node. However, most real-world data cannot be represented by an additive structural causal model. For example, in this case $X_1 = \mathbf{x}_1$, where $\mathbf{x}_1$ is a random variable uniformly distributed on $[0, 1]$, representing the exogenous influence of X_1 and $X_2 = X_1 \cdot \mathbf{x}_2$, where $\mathbf{x}_2$ is another random variable uniformly distributed on $[0, 1]$. In the generation of the X_2 feature, the exogenous variable $\mathbf{x}_2$ acts multiplicatively. Thus, fit the data using an additive causal model, we cannot determine the right causal structure, so the attribution values calculated using this method would not be accurate. Consider a special case where when $x_1^* = 1$ and $x_2^* = 1$, the resulting exogenous variable for x_2 is 0. This implies that x_2 has no effect on the data generation process, which clearly contradicts the true data generation process. DAG-SHAP does not require calculating the distribution of exogenous variables, nor does it assume the influence of exogenous variables in the data generation process is linear. It only need the causal direction between features. We think this is a significant distinction.

E Experiment Supplement

E.1 Attribution Results and Errors

In this section, we supplement the attribution results on the census dataset using deep neural networks (DNN), as well as the attribution results and error analysis using the XGBoost model across the synthetic and real datasets. The attribution result distribution using the DNN model on the Census dataset is shown in Figure 9 where $X_1, \cdots, X_6$ represent the following features: Race, Country, Age, Occupation, Marital Status, and Capital Gain.. The attribution results using the XGBoost model on the synthetic dataset and the mean absolute error between the results and the benchmark are shown in Figures 10 and 11, respectively.

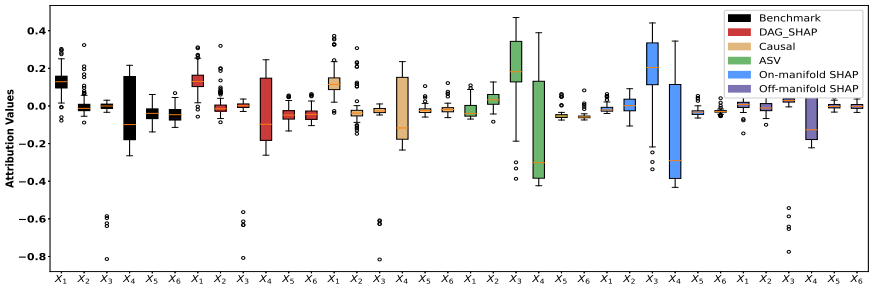


Figure 9: Attribution results of Census dataset(DNN).

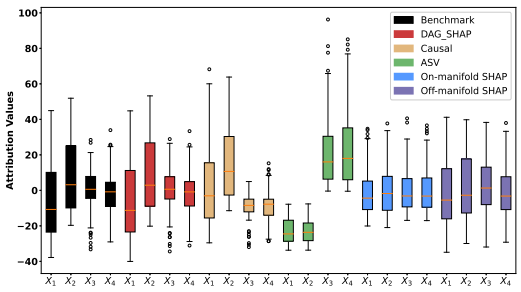


Figure 10: Results of synthetic dataset(XGBoost).

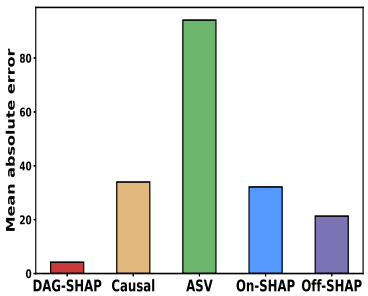


Figure 11: MAE(XGBoost).

The attribution results using the XGBoost model on the Griliches dataset and the mean absolute error between the results and the benchmark are shown in Figures 12 and 13, respectively. The attribution results using the XGBoost model on the census dataset and the mean absolute error between the results and the benchmark are shown in Figures 14 and 15, respectively.

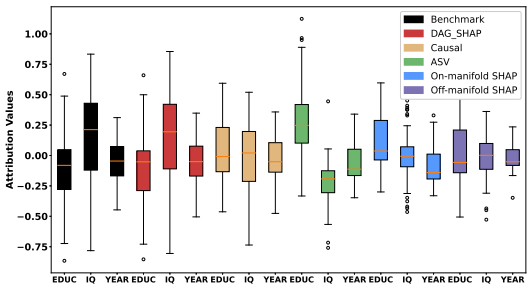


Figure 12: Results of Griliches76 dataset(XGBoost).

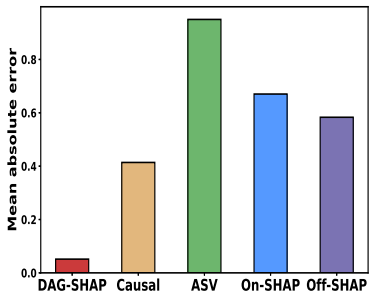


Figure 13: MAE(XGBoost).

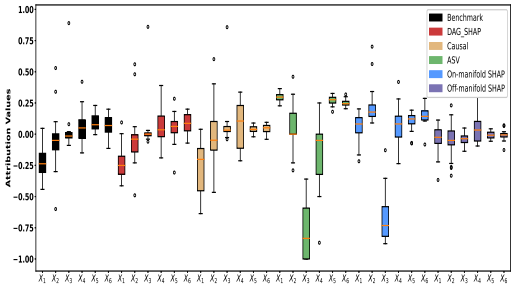


Figure 14: Attribution results of Census dataset(XGBoost).

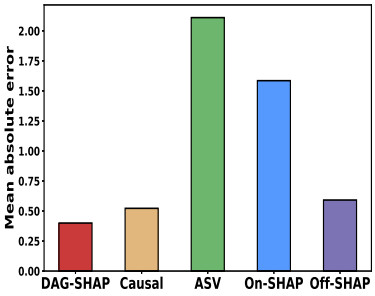


Figure 15: MAE(XGBoost).

E.2 Approximation Error Evaluation

Table 2: MAE of the SHAP values estimation.

Datasets	DAG-SHAP	Causal	ASV	On-SHAP	Off-SHAP
Synthetic(DNN)	3.717	1.921	2.935	3.322	2.988
Griliches76(DNN)	0.032	0.025	0.017	0.021	0.016
Census(DNN)	0.082	0.064	0.075	0.056	0.088
Synthetic(XGBoost)	3.913	2.897	3.564	3.341	2.926
Griliches76(XGBoost)	0.035	0.020	0.035	0.029	0.018
Census(XGBoost)	0.133	0.089	0.122	0.071	0.098

The attribution values in our experiments are obtained using approximate sampling methods. Since sampling estimates can introduce errors, this aspect is incorporated into the comparison between each attribution method and the benchmark. Therefore, we conducte a separate experiment to estimate the errors arising from the approximation. In the above experiments, 200 permutations are used for sampling during the attribution of each data point. We repeat the attribution process twice independently for each expalined data point, with each attribution method still using 200 permutations. The average error between the results of these two attributions is shown in Table 2. From the results, it can be observed that the errors introduced by sampling are much smaller than the discrepancies between each method and the benchmark. This indicates that the attribution errors compared to the benchmark are primarily due to inherent differences in the methods rather than sampling errors.

E.3 Scalability Evaluation

When dealing with large datasets, parallel computing can effectively accelerating the feature attribution calculations. Feature attribution for different data points can be executed in parallel. Additionally, within the feature attribution process for a single data point, sampling different permutations can

also be executed in parallel. We extende synthetic dataset (a) such that each data point had 100 features, and the DAG contained 200 edges. Specifically, we replicate X_1, X_2, X_3, and X_4 from the synthetic dataset (a) 25 times, resulting in 100 features in total. Each replicated feature maintains its causal relationship with Y, ensuring that the collective contribution of all features to Y remains consistent with the contribution in the original structure. We use the same neural network structure as in the experiments in Section 5.1 and use a server equipped with two AMD EPYC 9754 128-Core Processors, providing a total of 512 logical processors. For each data point, we conduct two independent rounds of sampling and compare the mean absolute error (MAE) between the sampling results. The experimental results show that with 128 sampling permutations per data point, the MAE is 5.17%, which is significantly smaller than the errors caused by the different feature attribution algorithms. For instance, the absolute error between off-SHAP which has the smallest MAE in baseline and the benchmark was 17.24%. Using 128 threads for parallel sampling of 128 permutations, the computation time was only 57.5 seconds. Additionally, we conducted experiments with 256 and 384 permutations, resulting in mean absolute errors of 4.23% and 3.71%, respectively.