IS THE DEEP MODEL REPRESENTATION SPARSE AND SYMBOLIC WITH CAUSAL PATTERNS?

Anonymous authors

Paper under double-blind review

Abstract

This paper aims to show that the inference logic of a deep model can be faithfully approximated as a sparse, symbolic causal graph. Such a causal graph potentially bridges the gap between connectionism and symbolism. To this end, the faithfulness of the causal graph is theoretically guaranteed, because we show that the causal graph can well mimic the model's output on an exponential number of different masked samples. Besides, such a causal graph can be further simplified and rewritten as an And-Or graph (AOG), which explains the logical relationship between interactive concepts encoded by the deep model, without losing much explanation accuracy. *The code will be released when the paper is accepted*.

1 INTRODUCTION

Can we faithfully explain the logic encoded by a deep model using a symbolic causal graph? Using a sparse and symbolic model to explain a black-box deep model may potentially bridge the gap between connectionism and symbolism. Recently, a line of studies (Frye et al., 2020; Heskes et al., 2020; Wang et al., 2021) have proposed to use manually defined causal relationship between input variables to compute their attributions w.r.t. the model output. However, previous studies mainly used heuristically pre-defined causality to explain deep models, instead of discovering and theoretically verifying the actual causal patterns used by the deep model for inference.

In this paper, we discover and prove that the inference logic of a deep model on a specific input sample can usually be represented as a sparse causal graph. Let a deep model have *n* input variables (*e.g.* a sentence with *n* words). Then, a three-layer causal graph in Fig. 1(b) can mimic the inference logic of the deep model. Each source node X_i (i = 1, ..., n) in the bottom layer represents the binary state of whether the *i*-th input variable is masked ($X_i = 0$) or not ($X_i = 1$). Each intermediate node C_S ($S = S_1, ..., S_K$) in the causal graph represents the AND relationship between a subset of input variables. C_S indicates the binary state of whether a specific causal pattern is triggered ($C_S = 1$) or not ($C_S = 0$). For example, in Fig. 1(b), the causal pattern C_S is triggered and makes an effect of "calm down," if and only if the three words in $S = \{X_4 = take, X_5 = it, X_6 = easy\}$ co-appear. The sink node Y in the top layer represents the output of the causal graph.

Then, let us define the faithfulness and conciseness of using the causal graph to explain a deep model.

• Faithfulness. Given an input sample with n variables, there are 2^n different ways to randomly mask input variables. Given any one of all the 2^n masked input sample, we prove that the output Y of the causal graph can always mimic the deep model's output. This guarantees that the causal graph encodes the same logic (*i.e.* the same set of interactive concepts) as the deep model. Thus, we can consider such a causal graph as a faithful explanation for the inference logic of the deep model.

• **Conciseness.** Theoretically, the totally precise fitting of the model's outputs on all the 2^n masked samples requires the causal graph to contain $K = 2^n$ nodes in the second layer, which is quite complex. However, we discover that we can use a very sparse graph with a small number of salient causal patterns to approximate the deep model's output in real applications. This is because most causal patterns have almost zero causal effects on the output *Y*, and thus can be ignored. In this way, we propose to learn a small causal graph with a few salient causal patterns to explain the deep model.

Note that since the deep model encodes complex inference logic, different input samples may activate different sets of salient causal patterns and generate different causal graphs.



Figure 1: We prove that the inference logic of a deep model (a) can be represented as a causal graph (b). The causal graph faithfully extracts concepts encoded by the deep model. Besides, the causal graph can be further simplified as an And-Or graph (AOG) (c,d), which extracts common coalitions.

Furthermore, we propose to summarize common coalitions shared by salient causal patterns to simplify the causal graph to a deep And-Or graph (AOG), as Fig. 1(c,d) shows.

• Concept discovery. Because we find that we can use a sparse causal graph to faithfully explain the inference logic of a deep model, we can consider each causal pattern in the graph represents a specific interactive concept encoded by the deep model. For example, in Fig. 1, the causal pattern C_S represents the AND relationship between words in $S = \{X_4 = take, X_5 = it, X_6 = easy\}$. Their co-appearance will make a causal effect $w_S = 2.0$ of the "calm down" emotion on the model output. Otherwise, the absence of any words in S will remove the "calm down" effect from the model output. In particular, the deep model's output can be written as a structural causal model (SCM) (Pearl, 2009), which sums up all triggered causal effects, *i.e.* $Y = \sum_S w_S \cdot C_S$.

• **Generality of causal patterns.** In this study, the causal patterns in the AOG are learned to satisfy the both requirements for faithfulness and conciseness. More crucially, we prove that the causal pattern can also explain the elementary mechanism of typical interaction metrics and attribution metrics for deep models, including the Shapley value (Shapley, 1953), the Shapley interaction index (Grabisch & Roubens, 1999), and the Shapley-Taylor interaction index (Sundararajan et al., 2020).

Contributions of this paper can be summarized as follows: (1) We discover and prove that the inference logic of a complex deep model on a certain sample can be represented as a relatively simple causal graph. (2) Furthermore, such a causal graph can be further simplified as an AOG. (3) The trustworthiness of using the AOG to explain a DNN is verified in experiments.

2 RELATED WORKS

Explanations for deep models. Many explanation methods have been proposed to explain the knowledge learned by deep models. Typical explanation methods include the visualization of features learned by the DNN (Simonyan et al., 2013; Zeiler & Fergus, 2014; Yosinski et al., 2015; Dosovitskiy & Brox, 2016), and the estimation of the pixel-wise attribution/saliency of input samples (Ribeiro et al., 2016; Lundberg & Lee, 2017; Fong & Vedaldi, 2017; Zhou et al., 2015; 2016; Selvaraju et al., 2017). Some studies explained a deep model's internel logic by distilling the model into another interpretable symbolic model, *e.g.* an additive model (Vaughan et al., 2018; Tan et al., 2018) or a decision tree (Frosst & Hinton, 2017; Che et al., 2016; Wu et al., 2018). Meanwhile, another direction is to compile the deep model into a set of logical formulas (Ignatiev et al., 2019a;b; Marques-Silva et al., 2021) or a logical decision graph (Shih et al., 2019). Zhang et al. (2018) used an explanatory graph to explain a deep model. *However, most of these explainer models were mainly learned to fit the model output, but whether their explanation can faithfully reflect the logic in the deep model is still an open problem without being fully investigated. In this study, we represent the inference logic of a deep model as a causal graph, and theoretically prove the faithfulness of such a representation.*

Using causality to explain deep models. The framework of causality was originally proposed to study the causal structure between a set of observed variables (Pearl, 2009; Hoyer et al., 2008). Recently, many studies have explained DNNs based on causality. For example, some studies (Frye et al., 2020; Heskes et al., 2020; Wang et al., 2021) proposed attribution methods that considered manually defined causal relationship hidden in the input. Similarly, Alvarez-Melis & Jaakkola (2017) and Harradon et al. (2018) extracted the inferring associations between inputs and intermediate features / outputs. Besides, Chattopadhyay et al. (2019) considered the output of an intermediate layer of the network as the causal effects of the input of this layer. Janzing et al. (2020) manually set a number of causal relationships between the input and the output of the DNN to explore algorithmic

flaws of attribution methods. Instead of manually setting or assuming the causal relationship, we quantify the exact interactive concepts encoded by the deep model as the causal patterns for inference, of which the faithfulness is both theoretically guaranteed and experimentally verified. Note that although the SCM in Eq. (2) seems like a linear model, our method does not explain the DNN in a similar manner like a bag-of-words model (Sivic & Zisserman, 2003; Csurka et al., 2004). It is because given different input samples, the DNN may activate different sets of causal patterns.

Interactions. Causal patterns in the proposed causal graph can actually be considered as a specific type of interactions in game theory. Just like causal effects, interactions are widely used to quantify numerical effects of interactive concepts between input variables on the deep model's output (Sorokina et al., 2008; Murdoch et al., 2018; Singh et al., 2018; Jin et al., 2019; Janizek et al., 2020). In game theory, the Shapley interaction index (Grabisch & Roubens, 1999) was used by (Lundberg et al., 2018) to analyze tree ensembles. Sundararajan et al. (2020) defined the Shapley-Taylor interaction index, and Tsai et al. (2022) proposed the Faith-Shap interaction index. Deng et al. (2022) proved that DNNs were less likely to encode interactive concepts of intermediate complexity, which was counter-intuitive. Unlike previous studies, we find that we can use a few causal patterns (*i.e.*, a few interactive concepts) to faithfully represent the inference logic of a deep model, which is experimentally verified.

3 Method

3.1 REPRESENTING A DEEP MODEL'S INFERENCE LOGIC AS A CAUSAL GRAPH

Given a pre-trained deep model $v(\cdot)$ and an input sample x with n variables $\mathcal{N} = \{1, 2, \ldots, n\}$ (e.g., a sentence with n words), we find that the inference logic of the deep model on sample x can be represented as a causal graph. As Fig. 1(b) shows, each source node X_i (i = 1, ..., n) in the bottom layer represents the binary state of whether the *i*-th input variable is masked ($X_i = 0$) or not ($X_i = 1$). The second layer consists of all causal patterns. Each causal pattern \mathcal{S} ($\mathcal{S} = \mathcal{S}_1, ..., \mathcal{S}_K$) represents the AND relationship between a subset of input variables $\mathcal{S} \subseteq \mathcal{N}$. For example, in Fig. 1(b), the co-appearance of the three words in $\mathcal{S} = \{take, it, easy\}$ forms a phrase meaning "calm down". In other words, only when all the three words are present, the causal pattern \mathcal{S} will be triggered, denoted by $C_{\mathcal{S}} = 1$; otherwise, $C_{\mathcal{S}} = 0$. As the output of the causal graph, the single sink node Y depends on triggering states $C_{\mathcal{S}}$ of all causal patterns in $\Omega = \{\mathcal{S}_1, ..., \mathcal{S}_K\}$. Thus, the transition probability in this causal graph is given as follows, and Theorem 1 proves that the inference logic of a trained deep model can be faithfully mimicked by this causal graph.

$$P(C_{\mathcal{S}} = 1 | X_1, X_2, ..., X_n) = \prod_{i \in \mathcal{S}} X_i, \quad P(Y | \{ C_{\mathcal{S}} | \mathcal{S} \in \Omega \}) = \mathbb{1} \left(Y = \sum_{\mathcal{S} \in \Omega} w_{\mathcal{S}} \cdot C_{\mathcal{S}} \right), \quad (1)$$

where $P(C_S = 0|X_1, X_2, ..., X_n) = 1 - P(C_S = 1|X_1, X_2, ..., X_n)$, and $\mathbb{1}(\cdot)$ refers to the indicator function. w_S can be understood as the causal effect of the pattern S to the output Y. Specifically, each triggered causal pattern C_S will contribute a certain causal effect w_S to the deep model's output. For example, the triggered causal pattern "take it easy" would contribute a significant additional effect $w_S > 0$ that pushes the deep model's output towards the positive meaning "calm down". The quantification of the causal effect w_S will be introduced later. According to Eq. (1), the causal relationship between C_S ($S = S_1, ..., S_K$) and the output Y of the causal graph can be rewritten as the following structural causal model (SCM) (Pearl, 2009), *i.e.*, the output Y of the causal graph sums up causal effects of all triggered causal patterns.

$$Y\big|_{X} = \sum_{\mathcal{S}\in\Omega} w_{\mathcal{S}} \cdot C_{\mathcal{S}}\big|_{X} \tag{2}$$

• Faithfulness of the causal graph. Given an input sample x with n variables, we prove that the inference logic of the deep model on this sample can be faithfully explained as the above causal graph. Specifically, we have 2^n ways to mask input variables in x, and generate 2^n different masked samples. If the output Y of a causal graph can always mimic the deep model's output on all the 2^n input samples, we can consider that the causal graph is faithful. Note that no matter whether input variables are dependent or not, the faithfulness will not be affected, *i.e.*, the causal graph can always accurately mimic the deep model's output on all 2^n possible masked input samples. To this end, given a subset of input variables $S \subseteq N$, let x_S denote the masked sample, where variables in $N \setminus S$ are masked, and other variables in S keep unchanged. Let $v(x_S)$ and $Y(x_S)$ denote the deep model's output on this sample x_S , respectively.

Theorem 1 (Proof in Appendix C). Given a certain input x, let the causal graph in Fig. 1 encode 2^n causal patterns, i.e., $K = 2^n$ and $\Omega = 2^N = \{S : S \subseteq N\}$. If the causal effect w_S of each causal

pattern $S \in \Omega$ is measured by the Harsanyi dividend (Harsanyi, 1963), i.e. $w_S \triangleq \sum_{S' \subseteq S} (-1)^{|S| - |S'|} \cdot v(\mathbf{x}_{S'})$, then the causal graph faithfully encodes the inference logic of the deep model, as follows.

$$\forall \mathcal{S} \subseteq \mathcal{N}, \ Y(\boldsymbol{x}_{\mathcal{S}}) = v(\boldsymbol{x}_{\mathcal{S}}) \tag{3}$$

More crucially, the Harsanyi dividend is the unique metric that satisfies the faithfulness requirement.

Theorem 1 proves the faithfulness of using such a causal graph to represent the inference logic of the deep model on a certain sample x. However, different samples mainly trigger different sets of causal patterns and generate different causal graphs. For example, given a cat image, pixels on the head (in S) may form a head pattern, and the DNN may assign a significant effect w_S on the pattern. Whereas, we cannot find the head pattern in a bus image, so the same set of pixels S in the bus image probably do not form any meaningful pattern and have ignorable effect $w_S \approx 0$.

Given the sample x, model outputs on the 2^n masked samples can all be accurately mimicked by the causal graph's output. Specifically, each masked sample x_S is obtained by masking all variables in $\mathcal{N} \setminus S$ using baseline values (Dabkowski & Gal, 2017; Ancona et al., 2019), as follows.

$$(\boldsymbol{x}_{\mathcal{S}})_{i} = \begin{cases} x_{i}, & i \in \mathcal{S} \\ r_{i}, & i \in \mathcal{N} \setminus \mathcal{S} \end{cases},$$
(4)

where $\mathbf{r} = [r_1, r_2, \dots, r_n]$ denotes the baseline values of the *n* input variables. The deep model's output $v(\mathbf{x}_S)$ is computed by taking the masked sample \mathbf{x}_S as the input. According to the SCM in Eq. (2), the corresponding output $Y(\mathbf{x}_S)$ of the causal graph is computed as $Y(\mathbf{x}_S) = \sum_{S \in \Omega} w_S \cdot C_S(\mathbf{x}_S) = \sum_{S' \subset S} w_S$. In particular, $Y(\mathbf{x} = \mathbf{x}_N) = \sum_{S \in \Omega} w_S$.

• Generality of causal patterns. Besides, we also prove that the above causal effects w_S based on Harsanyi dividends satisfy *the efficiency, linearity, dummy, symmetry, anonymity, recursive, and interaction distribution axioms* in game theory (see Appendix B), which further demonstrates the trustworthiness of the causal effects. More crucially, we also prove that causal effects w_S can explain the elementary mechanism of existing game-theoretic metrics. For example, both interaction metrics in Theorems 3 and 4 can be understood as the assignment of causal effects w_S to each involved coalition. Please see Appendix D for the proof and further discussions.

Theorem 2 (Connection to the Shapley value, proved by (Harsanyi, 1963)). Let $\phi(i)$ denote the Shapley value (Shapley, 1953) of an input variable *i*. Then, the Shapley value $\phi(i)$ can be explained as the result of uniformly assigning causal effects to each involving variable *i*, i.e., $\phi(i) = \sum_{S \subseteq N \setminus \{i\}} \frac{1}{|S|+1} w_{S \cup \{i\}}.$

Theorem 3 (Connection to the Shapley interaction index). Given a subset of input variables $\mathcal{T} \subseteq \mathcal{N}$, the Shapley interaction index (Grabisch & Roubens, 1999) $I^{\text{Shapley}}(\mathcal{T})$ can be represented as $I^{\text{Shapley}}(\mathcal{T}) = \sum_{S \subseteq \mathcal{N} \setminus \mathcal{T}} \frac{1}{|S|+1} w_{S \cup \mathcal{T}}$. In other words, the index $I^{\text{Shapley}}(\mathcal{T})$ can be explained as uniformly allocating causal effects $w_{S'}$ s.t. $S' = S \cup \mathcal{T}$ to the compositional variables of S', if we treat the coalition of variables in \mathcal{T} as a single variable.

Theorem 4 (Connection to the Shapley Taylor interaction index). Given a subset of input variables $\mathcal{T} \subseteq \mathcal{N}$, the k-th order Shapley Taylor interaction index (Sundararajan et al., 2020) $I^{\text{Shapley-Taylor}}(\mathcal{T})$ can be represented as weighted sum of causal effects, i.e., $I^{\text{Shapley-Taylor}}(\mathcal{T}) = w_{\mathcal{T}}$ if $|\mathcal{T}| < k$; $I^{\text{Shapley-Taylor}}(\mathcal{T}) = \sum_{S \subseteq \mathcal{N} \setminus \mathcal{T}} {|S|+k \choose k}^{-1} w_{S \cup \mathcal{T}}$ if $|\mathcal{T}| = k$; and $I^{\text{Shapley-Taylor}}(\mathcal{T}) = 0$ if $|\mathcal{T}| > k$.

3.2 DISCOVERING AND BOOSTING THE CONCISENESS OF THE CAUSAL GRAPH

Remark 1. Given a deep model $v(\cdot)$ and an input sample x with n variables, we can find a small set of causal patterns Ω subject to $|\Omega| \ll 2^n$, such that the deep model's output can be approximated by the causal graph's output, i.e. $\forall S \subseteq \mathcal{N}, Y(\mathbf{x}_S) \approx v(\mathbf{x}_S)$.

• Discovering the conciseness. Although Theorem 1 indicates that the causal graph needs to encode 2^n causal patterns to precisely fit the deep model's output on all the 2^n masked samples, Remark 1 shows a common phenomenon that the causal effects w_S extracted from the deep model are usually very sparse. The sparsity of causal effects is demonstrated in Fig. 2 and 8. Most causal patterns have little influence on the output with negligible values $|w_S| \approx 0$, and they are termed *noisy causal patterns*. Only a few causal patterns have considerable effects $|w_S|$, and they are termed *salient causal patterns*. To this end, we conducted experiments in Section 4.2, and Fig. 4 shows that we could use a small number of causal patterns (empirically 10 to 100 causal patterns for most deep models) in Ω to approximate the deep model's output, as stated in Remark 1.

• Boosting the conciseness. Inspired by Remark 1, we aim to learn a more concise causal graph. To this end, we propose the following objective of learning faithful and sparse causal effects w_S .

$$\min_{\boldsymbol{w},\Omega} \operatorname{unfaith}(\boldsymbol{w}_{\Omega}) \, s.t. \, |\Omega| \leq M \Leftrightarrow \min_{\boldsymbol{w},\Omega} \operatorname{unfaith}(\boldsymbol{w}_{\Omega}) \, s.t. \, \|\boldsymbol{w}_{\Omega}\|_{0} \leq M, \, \operatorname{unfaith}(\boldsymbol{w}_{\Omega}) = \sum_{\mathcal{S} \subseteq \mathcal{N}} \left[v(\boldsymbol{x}_{\mathcal{S}}) - Y(\boldsymbol{x}_{\mathcal{S}}) \right]^{2}$$
(5)

where $w_{\Omega} \stackrel{\text{def}}{=} [w'_{S_1}, w'_{S_1}, ..., w'_{S_{2^n}}]$. If $S \in \Omega$, then $w'_S = w_S$; otherwise, $w'_S = 0$. The L_0 -norm $||w_{\Omega}||_0$ refers to the number of non-zero elements in w_{Ω} , thereby $||w_{\Omega}||_0 = |\Omega|$. In this way, the above objective function enables people to use a small number of causal patterns to explain the deep model.

In this section, we propose several techniques to learn sparse causal effects based on Eq. (5) to faithfully mimic the deep model's outputs on numerous masked samples. The following paragraphs will introduce how to relax the Harsanyi dividend in Theorem 1 by removing noisy causal patterns and learning the optimal baseline value, so as to boost the sparsity of causal effects. Besides, we also discovered that adversarial training (Madry et al., 2018) can make the deep model encode much more sparse causal effects.

First, boosting conciseness by learning the optimal baseline value. In fact, the sparsity of causal patterns does not only depend on the deep model itself, but it is also determined by the choice of baseline values in Eq. (4). Specifically, input variables are masked by their baseline values $r = [r_1, r_2, ..., r_n]$ to represent their absence states in the computation of causal effects. Thus, w_{Ω} can be represented as a function of r, *i.e.*, $w_{\Omega}(r)$. To this end, some recent studies (Ancona et al., 2019; Dabkowski & Gal, 2017; Ren et al., 2021) defined baseline values from a heuristic perspective, *e.g.* simply using mean/zero baseline values (Dabkowski & Gal, 2017; Sundararajan et al., 2017). However, it still remains an open problem to define optimal baseline values.

Thus, we further boost the sparsity of causal patterns by learning the optimal baseline values that enhance the conciseness of the causal graph. However, it is difficult to learn optimal baseline values by directly optimizing Eq. (5). To this end, we relax the optimization problem in Eq. (5) (L_0 regression) as a Lasso regression (L_1 regression) as follows.

$$\min_{\Omega, \boldsymbol{r}} \operatorname{unfaith}(\boldsymbol{w}_{\Omega}) \ s.t. \ \|\boldsymbol{w}_{\Omega}\|_{0} \leq M \iff \min_{\Omega, \boldsymbol{r}} \ \operatorname{unfaith}(\boldsymbol{w}_{\Omega}) + \lambda \|\boldsymbol{w}_{\Omega}\|_{0} \xrightarrow{\operatorname{relax}} \min_{\Omega, \boldsymbol{r}} \ \operatorname{unfaith}(\boldsymbol{w}_{\Omega}) + \lambda \|\boldsymbol{w}_{\Omega}\|_{1}$$
(6)

We learn optimal baseline values by minimizing the loss $\mathcal{L}(\mathbf{r}, \Omega) = \text{unfaith}(\mathbf{w}_{\Omega}) + \lambda \cdot ||\mathbf{w}_{\Omega}||_1$. More curcially, the learning of baseline values is the most *safe* way of optimizing $\mathcal{L}(\mathbf{r}, \Omega)$, because the change of baseline values always ensures $\text{unfaith}(\mathbf{w}) = 0$ and just affects $||\mathbf{w}_{\Omega}||_1$. In this way, learning baseline values significantly boosts the conciseness of causal effects. In practice, we usually initialize the baseline value r_i as the mean value of the variable *i* over all samples, and then we constrain r_i within a relatively small range, *i.e.*, $||r_i - r_i^{\text{initial}}||^2 \leq \tau$, to represent the absence state¹.

Second, boosting conciseness by neglecting noisy causal patterns. Besides, considering the optimization problem, we use a greedy strategy to remove the noisy causal patterns from $2^{\mathcal{N}} = \{S : S \subseteq \mathcal{N}\}$ and keep the salient causal patterns to construct the set $\Omega \subseteq 2^{\mathcal{N}}$ that minimizes the loss $\mathcal{L}(\mathbf{r}, \Omega)$ in Eq. (6). It is worth noting that we do not directly learn causal effects by blindly optimizing Eq. (6), because automatically optimized causal effects usually lacks sufficient support for their physical meanings, while the setting of Harsanyi dividends is a meaningful interaction metric in game theory (Harsanyi, 1963). The Harsanyi dividend satisfies *the efficiency, linearity, dummy, symmetry axioms* axioms, which ensures the trustworthiness of this metric. In other words, although automatically optimized causal effects can minimize unfaith(\boldsymbol{w}), they still cannot be considered as reliable explanations from the perspective of game theory. Thus, we only recursively remove noisy causal patterns from Ω to update Ω , *i.e.*, $\Omega \leftarrow \Omega \setminus \{S\}$, without creating any new causal effect outside the paradigm of the Harsanyi dividends in Theorem 1. Specifically, we remove noisy causal patterns by following a greedy strategy, *i.e.*, iteratively removing the noisy causal patterns such that unfaith(\boldsymbol{w}_{Ω}) is minimized in each step. In this way, we just use the set of retained causal patterns, denoted by Ω , to approximate the output, *i.e.*, $\boldsymbol{v}(\boldsymbol{x}) \approx Y(\boldsymbol{x}) = \sum_{\boldsymbol{s} \in \Omega} \boldsymbol{w}_{\boldsymbol{s}} \cdot C_{\boldsymbol{s}}(\boldsymbol{x}) = \sum_{\boldsymbol{s} \in \Omega} \boldsymbol{w}_{\boldsymbol{s}}$.

Ratio of the explained causal effects R_{Ω} . We propose a metric R_{Ω} to quantify the ratio of the explained causal effects (*i.e.*, salient causal patterns in Ω) to the overall model output.

$$R_{\Omega} = \frac{\sum_{\mathcal{S} \in \Omega} |w_{\mathcal{S}}|}{\sum_{\mathcal{S} \in \Omega} |w_{\mathcal{S}}| + |\Delta|}$$
(7)

¹The setting of τ is introduced in Section 4.2. Please see Appendix E and F for more discussions.



Figure 2: AOGs that explained correct predictions made by the neural network. The networks were trained on (left) the CoLA dataset and (right) the SST-2 dataset, respectively. The red color of nodes in the second layer indicates causal patterns with positive effects, while the blue color represents patterns with negative effects. Red edges indicate the parse graph of a causal pattern.

where $\Delta = v(x) - \sum_{S \in \Omega} w_S$ denotes effects of the unexplained causal patterns. Besides R_{Ω} , Appendix H.9 also shows another metric for the explained effects.

Third, discovering that adversarial training boosts the conciseness. As discussed in Section 4.3, we also discover that adversarial training (Madry et al., 2018) makes the deep model encode more sparse causal patterns than standard training, thus boosting the conciseness of the causal graph.

3.3 REWRITING THE CAUSAL GRAPH AS AN AOG

In this section, we show that the above causal graph can be rewritten into an And-Or graph (AOG), which summarizes common coalitions shared by different causal patterns to further simplify the explanation. According to the SCM in Eq. (2), the causal graph in Section 3.1 actually represents the And-Sum representation encoded by the deep model, *i.e.*, $v(x) \approx \sum_{S \in \Omega} w_S \cdot C_S(x) = \sum_{S \in \Omega} w_S$. In fact, such And-Sum representation can be equivalently transformed into an AOG.

The AOG is a hierarchical graphical model that encodes how semantic patterns are formed for inference, which has been widely used for interpretable knowledge representation (Li et al., 2019; Zhang et al., 2020), object detection (Song et al., 2013), *etc.* The structure of a simple three-layer AOG is shown in Fig. 1(c). Just like the causal graph in Fig. 1(b), at the bottom layer of the AOG in Fig. 1(c), there are *n* leaf nodes representing *n* variables of the input sample. The second layer of the AOG has multiple AND nodes, each representing the AND relationship between its child nodes. For example, the AND node $x_4x_5x_6$ indicates the causal pattern $S = \{x_4, x_5, x_6\}$ with the causal effect $w_S = 2.0$. The root node is a *noisy OR* node (as discussed in (Li et al., 2019)), which sums up effects of all its child AND nodes to mimic the model output, *i.e.*, *output* = $\sum_{S \in \Omega} w_S \cdot C_S$.

Furthermore, in order to simplify the AOG, we extract common coalitions shared by different causal patterns as new nodes to construct a deeper AOG. For example, in Fig. 1(c), input variables x_5 and x_6 frequently co-appear in different causal patterns. Thus, we consider x_5, x_6 as a coalition and add an AND node $\beta = \{x_5, x_6\}$ to represent their co-appearance. Accordingly, the pattern $\{x_4, x_5, x_6\}$ is simplified as $\{x_4, \beta\}$ (see Fig. 1(d)). Therefore, for each coalition / causal pattern S in an intermediate layer, its triggering state $C_S = \prod_{S' \in \text{Child}(S)} C_{S'}$, where Child(S) denotes all input variables or coalitions composing S. *I.e.*, each coalition / causal pattern S is triggered if and only if all its child nodes in Child(S) are triggered.

In order to extract common coalitions, we use the minimum description length (MDL) principle (Hansen & Yu, 2001) to learn the AOG g as the simplest description of causal patterns. The MDL principle is widely used in information theory to balance the model (graph) complexity and the complexity of using the model to describe data (causal patterns). Given an AOG g and input variables \mathcal{N} , let $\mathcal{M} = \mathcal{N} \cup \Omega^{\text{coalition}}$ denote the set of all leaf nodes and AND nodes in the bottom two layers, *e.g.* $\mathcal{M} = \mathcal{N} \cup \Omega^{\text{coalition}} = \{x_1, x_2, ..., x_6\} \cup \{\alpha, \beta\}$ in Fig. 1(d). The objective of minimizing the description length $L(g, \mathcal{M})$ is given as follows.

$$\min_{\mathcal{M}} L(g, \mathcal{M}) \quad s.t. \quad L(g, \mathcal{M}) = \underbrace{L(\mathcal{M})}_{\text{complexity (length) of describing the set of nodes } \mathcal{M}} + \underbrace{L_{\mathcal{M}}(g)}_{\text{complexity (length) of using nodes in } \mathcal{M} \text{ to describe patterns in } g}$$
(8)

The MDL principle usually formulates the complexity (description length) of the set of nodes \mathcal{M} as the entropy $L(\mathcal{M}) = -\kappa \sum_{m \in \mathcal{M}} p(m) \log p(m)$. We set the occurring probability p(m) of the node $m \in \mathcal{M}$ proportional to the overall strength of causal effects of the node *m*'s all parent nodes \mathcal{S} , Child $(\mathcal{S}) \ni m$. $\forall m \in \mathcal{M}, p(m) = count(m) / \sum_{m' \in \mathcal{M}} count(m')$ s.t. $count(m) = \sum_{\mathcal{S} \in \Omega: \text{Child}(\mathcal{S}) \ni m} |w_{\mathcal{S}}|$. $\kappa = 10/Z$ is a scalar weight, where $Z = \sum_{\mathcal{S} \in \Omega} |w_{\mathcal{S}}|$. The second term $L_{\mathcal{M}}(g) = -\mathbb{E}_{\mathcal{S} \sim p(\mathcal{S}|g)} \sum_{m \in \mathcal{S}} \log p(m)$ represents



towards Table 1: IoU (\uparrow) on the synthesized datasets. The AOG explainer correctly extracted causal patterns.

		Average IoU					
Dataset	Model	SI	STI (k=2)	STI (k=3)	ours		
Add-Mul dataset (Zhang et al., 2021)	functions in	0.61	0.27	0.55	1.00		
(Ren et al., 2021)	the dataset	0.99	0.50	0.59	1.00		
Manually labeled	MLP-5 PerMLP 5	0.87	0.35	0.69	0.97		
Anu-Or ualaset	KUSWILF-J	0.90	0.55	0.09	0.90		

Figure 3: AOGs for a network trained on the CoLA dataset. We randomly highlight a parse graph (blue) in the AOG.

the complexity (description length) of using nodes in \mathcal{M} to describe all causal patterns in g. The appearing probability of the causal pattern S in the AOG g is sampled as $p(S|g) \propto |w_S|$. The loss $L(g, \mathcal{M})$ can be minimized by recursively adding common coalitions into \mathcal{M} via the greedy strategy by following (Hansen & Yu, 2001). Please see Appendix G for more discussions.

Limitations of the AOG explainer. Since the AOG explainer is built on a specific input sample, an AOG can only explain the inference logic of the deep model on a specific sample, instead of encoding the common logic of different samples. In other words, we cannot replace the deep model with the AOG explainer for inference. Besides, although we prove that the AOG explainer is the unique faithful explanation, it is still far from a computationally efficient explanation. Thus, extending the theoretical solution to the practical one is our future work, *e.g.* developing approximated methods or accelerating techniques for computation.

4 **EXPERIMENTS**

Datasets and models. We focused on both tasks of natural language processing and the classification/regression task based on tabular datasets. For the natural language processing, we explained LSTMs (Hochreiter & Schmidhuber, 1997) and CNNs used in (Rakhlin, 2016). Each model was trained for sentiment classification on the SST-2 dataset (Socher et al., 2013) or for linguistic acceptability classification on the CoLA dataset (Warstadt et al., 2019), respectively. The tabular datasets included the UCI census income dataset (Dua & Graff, 2017), the UCI bike sharing dataset (Dua & Graff, 2017), and the UCI TV news channel commercial detection dataset (Dua & Graff, 2017). These datasets were termed *census, bike,* and *TV news* for simplicity. Each tabular dataset was used to train MLPs, LightGBM (Ke et al., 2017), and XGBoost (Chen & Guestrin, 2016). For MLPs, we used two-layer MLPs (namely *MLP-2*) and five-layer MLPs (namely *MLP-5*), where each layer contained 100 neurons. Besides, we added a skip-connection to each layer of the MLP-5 to build *ResMLP-5*. Please see Appendix H.1 for more details.

Explaining representation flaws of deep models. Fig. 2 and Fig. 3 show AOG explanations for correct predictions and incorrect predictions, respectively. The highlighted parse graph in each figure correspond to a single causal pattern. We only visualized a single parse graph in each AOG for clarity. Results show that the AOG explainer could reveal the representation flaws that were responsible for incorrect predictions. For example, local correct grammar "she was" in Fig. 3(left) was mistakenly learned to make negative impacts on the linguistic acceptability of the whole sentence. The phrase "John placed" in Fig. 3(right) directly hurt the linguistic acceptability without considering the complex structure of the sentence. Please see Appendix H.4 for more results.

4.1 EXAMINING WHETHER THE AOG EXPLAINER REFLECTS FAITHFUL CAUSALITY

In this section, we proposed two metrics to examine whether the AOG explainer faithfully reflected the inference logic encoded by deep models.

Metric 1: intersection over union (IoU) between causal patterns in the AOG explainer and ground-truth causal patterns. This metric evaluated whether causal patterns (nodes) in the AOG explainer correctly reflected the interactive concepts encoded by the model. Given a model and an input sample, let *m* denote the number of ground-truth causal patterns $m = |\Omega^{\text{truth}}|$ in the input. Then, for fair comparisons, we also used *m* causal patterns $\Omega^{\text{top-}m}$ in the AOG explainer with the top-*m* causal effects $|w_S|$. We measured the IoU between Ω^{truth} and $\Omega^{\text{top-}m}$ as $IoU = |\Omega^{\text{top-}m} \cap \Omega^{\text{truth}}|/|\Omega^{\text{top-}m} \cup \Omega^{\text{truth}}|$ to evaluate the correctness of the extracted causal patterns in the AOG explainer. A higher IoU value

Explanation methods		TV news		census		bike		ers were	more si	innar i	nan tv	vo nor-
		MLP-5	ResMLP-5	MLP-5	ResMLP-5	MLP-5	ResMLP-5	mally trained ones				
Attribution	Buylian Shapley 125.5 130.8 55.6 51.4 1.1E+4 7953.9 many trained off				incu one.	5.						
hasad	I×G	738.7	2586.1	408.1	1325.1	1.4E+5	1.1E+5			TV news	census	bike
evolutions	LRP	317.6	9.4E+4	155.1	1.4E+04	1.4E+5	5.8E+8	MIDO	normal	0.5965	0.4899	-
explanations	OCC	1386.2	1117.5	638.7	287.4	6.2E+4	3.7E+4	NILF-2	adversarial	0.6109	0.6292	-
Interaction	SI	6231.2	5598.6	2726.1	2719.0	1.2E+5	1.2E+5	MID 5	normal	0.3664	0.2482	0.3816
-based	STI (k=2)	182.0	236.0	34.7	38.8	7685.0	5219.8	MLF-3	adversarial	0.6304	0.4971	0.4741
explanations	STI (k=3)	177.7	252.4	41.0	60.5	1.0E+4	5045.8	DecMID 5	normal	0.3480	0.2764	0.3992
ours		9.4E-12	1.1E-11	8.5E-12	8.5E-12	2.6E-9	1.9E-9	ResMLP-5	adversarial	0.5731	0.4489	0.4491

Table 2: Unfaithfulness ρ^{unfaith} (\downarrow) of different explanation methods. Our AOG exhibited the lowest unfaithfulness. Table 3: Jaccard similarity between two models. Two adversarially trained models.

means a larger overlap between the ground-truth causal patterns and the extracted causal patterns, which indicates higher correctness of the extracted causal patterns.

However, for most datasets and models, people could not annotate the ground-truth patterns, as discussed in (Zhang et al., 2021). Therefore, we used the off-the-shelf functions with ground-truth causal patterns in the Addition-Multiplication (Add-Mul) dataset (Zhang et al., 2021) and the dataset proposed in (Ren et al., 2021), to test whether the learned AOGs could faithfully explain these functions. The ground-truth causal patterns of functions in both datasets can be easily determined. For example, for the function $y = x_1x_3 + x_3x_4x_5 + x_4x_6, x_i \in \{0, 1\}$ in the Add-Mul dataset, the ground-truth causal patterns are $\Omega^{\text{truth}} = \{\{x_1, x_3\}, \{x_3, x_4, x_5\}, \{x_4, x_6\}\}$ given the input sample x = [1, 1, ..., 1]. It was because the co-appearance of variables in each causal pattern would contribute 1 to the output score y. Similarly, we also constructed a dataset containing pre-defined And-Or functions with ground-truth causal patterns, namely the *manually labeled And-Or dataset* (see Section H.5). Then, we learned the aforementioned MLP-5 and ResMLP-5 networks to regress each And-Or function. We considered causal patterns in such And-Or functions as ground-truth causal patterns in the neural networks.

Actually, previous studies usually did not directly extract causal patterns from a trained DNN at as a low level as input units. To this end, interaction metrics (such as the Shapley interaction (SI) index (Grabisch & Roubens, 1999) and the Shapley-Taylor interaction (STI) index (Sundararajan et al., 2020)) were widely used to quantify numerical effects of different interactive patterns between input variables on the network output. Thus, we computed interactive patterns with top-ranked SI values, or patterns with top-ranked STI values of orders k=2 and k=3, as competing causal patterns for comparison. Based on the IoU score defined above, Table 1 shows that our AOG explainer successfully explained much more causal patterns than other interaction metrics.

Metric 2: evaluating faithfulness of the the AOG explainer. We also proposed a metric ρ^{unfaith} to evaluate whether an explanation method faithfully extracted causal effects encoded by deep models. As discussed in Section 3.2, if the quantified causal effects w are faithful, then they are supposed to minimize unfaith(w). Therefore, according to the SCM in Eq. (2), we defined $\rho^{\text{unfaith}} = \mathbb{E}_{S \subseteq \mathcal{N}}[v(\boldsymbol{x}_S) - \sum_{S' \subseteq S} w_{S'})]^2$ to measure the unfaithfulness. As mentioned above, we considered the SI values and STI values as numerical effects w_S of different interactive patterns S on a DNN's inference. Besides, we could also consider that attribution-based explanations quantified the causal effect $w_{\{i\}}$ of each single variable i. Therefore, Table 2 compares the extracted causal effects in the AOG with SI values, STI values, and attribution-based explanations (including the Shapley value (Shapley, 1953), Input×Gradient (Shrikumar et al., 2016), LRP (Bach et al., 2015), and Occlusion (Zeiler & Fergus, 2014)). Our AOG explainer exhibited much lower ρ^{unfaith} values than baseline methods.

4.2 CONCISENESS OF THE AOG EXPLAINER

The conciseness of an AOG depends on a trade-off between the ratio of the explained causal effects R_{Ω} and simplicity of the explanation. In this section, we evaluated effects of baseline values on the simplicity of the AOG explainer, and examined the relationship between the ratio of causal effects being explained and the simplicity of the AOG explainer.

Effects of baseline values on the conciseness of explanations. In this experiment, we explored whether the learning of baseline values in Section 3.2 could boost the sparsity of causal patterns. To this end, we followed (Dabkowski & Gal, 2017) to initialize baseline values of input variables as their mean values over different samples. Then, we learned baseline values via Eq. (6). The baseline



Figure 4: (a) The change of R_{Ω} (the ratio of the explained causal effects) along with the number of causal patterns $|\Omega|$ in AOGs. (b, c) The change of the node number and the edge number in AOGs along with R_{Ω} . AOGs corresponding to adversarially trained models were less complex than AOGs *w.r.t.* normally trained models. (d) The histogram of the re-scaled causal effects. The learned baseline values boosted the sparsity of causal patterns in the AOG explainer.

value r_i of each input variable *i* was constrained within a certain range around the data average, *i.e.*, $||r_i - \mathbb{E}_x[x_i]||^2 \leq \tau$. In experiments, we set $\tau = 0.01 \cdot \operatorname{Var}_x[x_i]$, where $\operatorname{Var}_x[x_i]$ denotes the variance of the *i*-th input variable over different samples. Fig. 4(d) shows the histogram of the relative strength of causal effects $\frac{|w_S|}{\max_{S' \subseteq \mathcal{N}} |w_{S'}|}$, which was re-scaled to the range of [0, 1]. Compared with mean baseline values, the learned baseline values usually generated fewer causal patterns with significant strengths, which boosted the sparsity of causal effects and enhanced the conciseness of explanations. In this experiment, we used MLP-5 and computed the re-scaled strengths of causal effects with training samples in the TV news dataset. Please see Appendix H.10 for more experimental results.

Ratio of the explained causal effects R_{Ω} . There was a trade-off between faithfulness (the ratio of explained causal effects) and conciseness of the AOG. A good explanation was supposed to improve the simplicity while keeping a large ratio of causal effects being explained. As discussed in Section 3.2, we just used causal patterns in Ω to approximate the model output. Fig. 4(a) shows the relationship between $|\Omega|$ and the ratio of the explained causal effects R_{Ω} in different models based on the TV news dataset. When we used a few causal patterns, we could explain most effects of causal patterns to the model output. Fig. 4(b,c) shows that the node and edge number of the AOG increased along with the increase of R_{Ω} . Please see Appendix H.8 for results on other datasets.

4.3 EFFECTS OF ADVERSARIAL TRAINING

In this experiment, we learned MLP-2, MLP-5, and ResMLP-5 on the TV news dataset via adversarial training (Madry et al., 2018). Fig. 4(a) shows that compared with normally trained models, we could use less causal patterns (smaller $|\Omega|$) to explain the same ratio of causal effects R_{Ω} in adversarially trained models. Moreover, Fig. 4(b,c) also shows that AOGs for adversarially trained models contained even less nodes and edges than AOGs for normally trained models. This indicated that adversarial training made models encode more sparse causal patterns than normal training.

Besides, adversarial training also made different models encode common patterns. To this end, we trained different pairs of models with the same architecture but with different initial parameters. Given the same input, we measured the Jaccard similarity coefficient between causal effects of each pair of models, in order to examine whether the two models encoded similar causal patterns. Let w_s and w'_s denote causal effects in the two models. The Jaccard similarity coefficient was computed as $J = \frac{\sum_{s \subseteq N} \min(|w_s|, |w'_s|)}{\sum_{s \subseteq N} \max(|w_s|, |w'_s|)}$. A high Jaccard similarity indicated that the two models encoded similar causal patterns for inference. Table 3 shows that the similarity between two adversarially trained models was significantly higher than that between two normally trained models. This indicated adversarial training made different models encode common causal patterns for inference.

5 CONCLUSION

In this paper, we show that the inference logic of a deep model can usually be represented as a sparse causal graph. To this end, we theoretically prove and experimentally verify the faithfulness of using a sparse causal graph to explain the interactive concepts encoded in a DNN. We also propose several techniques to boost the conciseness of such causal representation. Furthermore, we show that such a causal graph can be rewritten as an AOG, which further simplifies the explanation. The AOG explainer provides new insights for understanding the inference logic of deep models.

ETHICS STATEMENT

This study focused on understanding the inference logic of a trained deep model. Our work shed new insights for research fields including explainable AI (XAI), and provided new mathematical tools to analyze deep models. Specifically, we proposed the AOG explainer for understanding the deep model, whose faithfulness was theoretically guaranteed (Section 3.1). We also developed several techniques to further simplify the explanation (Section 3.2, 3.3). As a generic mathematical tool to analyze deep models, we have applied the proposed method to various deep models, including LightGBM, XGBoost, and deep neural networks. Experimental results provided new insights for understanding the inference logic of deep models.

Reproducibility Statement

Most work in this study is to theoretically derive the faithfulness of the proposed causal graph and AOG explainer. Therefore, there is no problem with the reproducibility. Appendix C and Appendix D provide proofs for all theoretical results in the paper. Besides, we have discussed all experimental details about datasets, models, and details for building the AOG in Section 3.3 and Appendix H, which ensures the reproducibility. Furthermore, we will release the code when the paper is accepted.

REFERENCES

- David Alvarez-Melis and Tommi S Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *EMNLP*, 2017.
- Marco Ancona, Cengiz Oztireli, and Markus Gross. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In *International Conference on Machine Learning*, pp. 272–281. PMLR, 2019.
- Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.
- Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective. In *ICML*, 2019.
- Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable deep models for icu outcome prediction. In *AMIA annual symposium proceedings*, volume 2016, pp. 371. American Medical Informatics Association, 2016.
- Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.
- Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 3457–3465. PMLR, 2021.
- Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33, 2020.
- Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision*, *ECCV*. Prague, 2004.
- Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *arXiv preprint arXiv:1705.07857*, 2017.
- Huiqi Deng, Qihan Ren, Hao Zhang, and Quanshi Zhang. Discovering and explaining the representation bottleneck of dnns. In *International Conference on Learning Representations*, 2022.

- Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4829– 4837, 2016.
- Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive. ics.uci.edu/ml.
- Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2950–2958, 2019.
- Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.
- Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. *arXiv* preprint arXiv:1711.09784, 2017.
- 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.
- Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4):547–565, 1999.
- Mark H Hansen and Bin Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.
- Michael Harradon, Jeff Druce, and Brian Ruttenberg. Causal learning and explanation of deep neural networks via autoencoded activations. *arXiv preprint arXiv:1802.00541*, 2018.
- John C Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963.
- Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv* preprint arXiv:1503.02531, 2015.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.
- Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, Bernhard Schölkopf, et al. Nonlinear causal discovery with additive noise models. In *NIPS*, 2008.
- Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *AAAI*, 2019a.
- Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. *Advances in Neural Information Processing Systems*, 32:15883–15893, 2019b.
- Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *arXiv preprint arXiv:2002.04138*, 2020.
- 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.
- Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *International Conference on Learning Representations*, 2019.

- Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30:3146–3154, 2017.
- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- Xilai Li, Xi Song, and Tianfu Wu. Aognets: Compositional grammatical architectures for deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6220–6230, 2019.
- Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777, 2017.
- Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.
- Joao Marques-Silva, Thomas Gerspacher, Martin C Cooper, Alexey Ignatiev, and Nina Narodytska. Explanations for monotonic classifiers. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7469–7479. PMLR, 18–24 Jul 2021. URL https://proceedings. mlr.press/v139/marques-silva21a.html.
- W James Murdoch, Peter J Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from lstms. In *International Conference on Learning Representations*, 2018.
- Judea Pearl. Causality. Cambridge university press, 2009.
- A Rakhlin. Convolutional neural networks for sentence classification. *GitHub*, 2016.
- Jie Ren, Zhanpeng Zhou, Qirui Chen, and Quanshi Zhang. Learning baseline values for shapley values. *arXiv preprint arXiv:2105.10719*, 2021.
- 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.
- 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. *Contributions to the Theory of Games*, 2(28):307–317, 1953.
- Andy Shih, Arthur Choi, and Adnan Darwiche. Compiling bayesian network classifiers into decision graphs. In *AAAI*, 2019.
- Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- Chandan Singh, W James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*, 2018.

- Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, volume 3, pp. 1470–1470. IEEE Computer Society, 2003.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.
- Xi Song, Tianfu Wu, Yunde Jia, and Song-Chun Zhu. Discriminatively trained and-or tree models for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3278–3285, 2013.
- Daria Sorokina, Rich Caruana, Mirek Riedewald, and Daniel Fink. Detecting statistical interactions with additive groves of trees. In *Proceedings of the 25th international conference on Machine learning*, pp. 1000–1007, 2008.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3319–3328, 2017.
- Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International Conference on Machine Learning*, pp. 9259–9268. PMLR, 2020.
- Sarah Tan, Rich Caruana, Giles Hooker, Paul Koch, and Albert Gordo. Learning global additive explanations for neural nets using model distillation. *arXiv preprint arXiv:1801.08640*, 2018.
- Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-shap: The faithful shapley interaction index. *arXiv preprint arXiv:2203.00870*, 2022.
- Joel Vaughan, Agus Sudjianto, Erind Brahimi, Jie Chen, and Vijayan N Nair. Explainable neural networks based on additive index models. *arXiv preprint arXiv:1806.01933*, 2018.
- Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 721–729. PMLR, 2021.
- Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.
- Mike Wu, Michael C Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Beyond sparsity: Tree regularization of deep models for interpretability. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.
- Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.
- Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. Interpreting multivariate shapley interactions in dnns. In AAAI, 2021.
- Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. Interpreting cnn knowledge via an explanatory graph. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- Quanshi Zhang, Jie Ren, Ge Huang, Ruiming Cao, Ying Nian Wu, and Song-Chun Zhu. Mining interpretable aog representations from convolutional networks via active question answering. *IEEE transactions on pattern analysis and machine intelligence*, 2020.
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *In ICLR*, 2015.
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

A INTERACTIVE DEMOS OF AOGS

We provide several interactive demos of AOGs (named by *.html under the supplementary zip file) for better visualization and understanding of AOGs. Each html file corresponds to the AOG explainer for the model inference on an input sample. When the user hovers on each causal pattern in the AOG, its corresponding parse graph will be highlighted, which shows the causal effect of a certain causal pattern. Note that these interactive demos are all local html scripts rather than online web pages, and they contain no affiliation and personal information. Please see Section H.2 and Section H.4 for more discussions and static visualizations of the AOGs.

B HARSANYI DIVIDEND

This section revisits the definition of the Harsanyi dividend (Harsanyi, 1963), which is a typical metric in game theory. In this paper, the causal effect w_S of each pattern S is quantified based on the Harsanyi dividend. In game theory, a complex system (*e.g.* a deep model) can usually be considered as a game. Each input variable is a player of the game, and the output of this system is the reward obtained by some subset of players. Specifically, let us consider a deep model and an input sample x with n variables (*e.g.* a sentence with n words) $\mathcal{N} = \{1, 2, ..., n\}$. The deep model can be understood as a game $v(\cdot)$. In this game, input variables in \mathcal{N} do not contribute to the model output individually. Instead, they interact with each other to form concepts (causal patterns) for inference. Each concept $S \subseteq \mathcal{N}$ will add a certain causal effect to the model output. In this paper, we prove in Theorem 1 that the Harsanyi dividend w_S is the unique faithful metric to quantify such causal effects.

$$w_{\mathcal{S}} = \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}'| - |\mathcal{S}|} \cdot v(\boldsymbol{x}_{\mathcal{S}'}), \tag{9}$$

where $v(x_S)$ denote the model output when only variables in the subset $S \subseteq N$ are given, and all other variables are masked using their baseline values.

In this paper, we also prove that the Harsanyi dividend w_S satisfies seven desirable axioms, including the *efficiency, linearity, dummy, symmetry, anonymity, recursive* and *interaction distribution* axioms, which demonstrates its trustworthiness.

(1) *Efficiency axiom* (proved by (Harsanyi, 1963)). The output score of a model can be decomposed into effects of different causal patterns, *i.e.* $v(x) = \sum_{S \subseteq N} w_S$.

(2) *Linearity axiom.* If we merge output scores of two models $t(\cdot)$ and $u(\cdot)$ as the output of model $v(\cdot)$, *i.e.* $\forall S \subseteq \mathcal{N}, v(\boldsymbol{x}_S) = t(\boldsymbol{x}_S) + u(\boldsymbol{x}_S)$, then the corresponding causal effects w_S^t and w_S^u can also be merged as $\forall S \subseteq \mathcal{N}, w_S^v = w_S^t + w_S^u$.

(3) Dummy axiom. If a variable $i \in \mathcal{N}$ is a dummy variable, *i.e.* $\forall S \subseteq \mathcal{N} \setminus \{i\}, v(\boldsymbol{x}_{S \cup \{i\}}) = v(\boldsymbol{x}_S) + v(\boldsymbol{x}_{\{i\}})$, then it has no causal effect with other variables, $\forall S \subseteq \mathcal{N} \setminus \{i\}, w_{S \cup \{i\}} = 0$.

(4) Symmetry axiom. If input variables $i, j \in \mathcal{N}$ cooperate with other variables in the same way, $\forall S \subseteq \mathcal{N} \setminus \{i, j\}, v(\boldsymbol{x}_{S \cup \{i\}}) = v(\boldsymbol{x}_{S \cup \{j\}})$, then they have same causal effects with other variables, $\forall S \subseteq \mathcal{N} \setminus \{i, j\}, w_{S \cup \{i\}} = w_{S \cup \{j\}}.$

(5) Anonymity axiom. For any permutations π on \mathcal{N} , we have $\forall S \subseteq \mathcal{N}, w_S^v = w_{\pi S}^{\pi v}$, where $\pi S \triangleq \{\pi(i) | i \in S\}$, and the new model πv is defined by $(\pi v)(\boldsymbol{x}_{\pi S}) = v(\boldsymbol{x}_S)$. This indicates that causal effects are not changed by permutation.

(6) *Recursive axiom.* The causal effects can be computed recursively. For $i \in \mathcal{N}$ and $\mathcal{S} \subseteq \mathcal{N} \setminus \{i\}$, the causal effect of the pattern $\mathcal{S} \cup \{i\}$ is equal to the causal effect of \mathcal{S} with the presence of i minus the causal effect of \mathcal{S} with the absence of i, *i.e.* $\forall \mathcal{S} \subseteq \mathcal{N} \setminus \{i\}, w_{\mathcal{S} \cup \{i\}} = w_{\mathcal{S}|i \text{ present}} - w_{\mathcal{S}}$. $w_{\mathcal{S}|i \text{ present}}$ denotes the causal effect when the variable i is always present as a constant context, *i.e.* $w_{\mathcal{S}|i \text{ present}} = \sum_{\mathcal{S}' \subset \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} \cdot v(\boldsymbol{x}_{\mathcal{S}' \cup \{i\}}).$

(7) Interaction distribution axiom. This axiom characterizes how causal effects are distributed for a class of "interaction functions" (Sundararajan et al., 2020). An interaction function $v_{\mathcal{T}}$ parameterized by a subset of variables \mathcal{T} is defined as follows. $\forall S \subseteq \mathcal{N}$, if $\mathcal{T} \subseteq S$, $v_{\mathcal{T}}(\boldsymbol{x}_S) = c$; otherwise, $v_{\mathcal{T}}(\boldsymbol{x}_S) = 0$. The function $v_{\mathcal{T}}$ purely models the causal effect of the pattern \mathcal{T} , because only if all variables in \mathcal{T} are present, the output value will be increased by c. The causal effects encoded in the function $v_{\mathcal{T}}$ satisfy $w_{\mathcal{T}} = c$, and $\forall S \neq \mathcal{T}$, $w_S = 0$.

More crucially, we have also proved that causal effects w_S based on the Harsanyi dividend can explain the elementary mechanism of existing game-theoretic attributions/interactions, as follows.

Theorem 5 (Connection to the marginal benefit (Grabisch & Roubens, 1999)). $\Delta v_{\mathcal{T}}(x_{\mathcal{S}}) = \sum_{\mathcal{T}' \subseteq \mathcal{T}} (-1)^{|\mathcal{T}| - |\mathcal{T}'|} v(x_{\mathcal{T}' \cup \mathcal{S}})$ denotes the marginal benefit of variables in $\mathcal{T} \subseteq \mathcal{N} \setminus \mathcal{S}$ given the environment \mathcal{S} . We have proven that $\Delta v_{\mathcal{T}}(x_{\mathcal{S}})$ can be decomposed into the sum of causal effects inside \mathcal{T} and sub-environments of \mathcal{S} , i.e. $\Delta v_{\mathcal{T}}(x_{\mathcal{S}}) = \sum_{\mathcal{S}' \subseteq \mathcal{S}} w_{\mathcal{T} \cup \mathcal{S}'}$.

Theorem 2 (Connection to the Shapley value (Shapley, 1953), proved by (Harsanyi, 1963)). Let $\phi(i)$ denote the Shapley value of an input variable *i*. Then, the Shapley value $\phi(i)$ can be explained as the result of uniformly assigning causal effects to each involving variable *i*, i.e., $\phi(i) = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{|S|+1} w_{S \cup \{i\}}$. This theorem also proves that the Shapley value is a fair assignment of attributions from the perspective of causal effects, as shown in Figure 5.

Theorem 3 (Connection to the Shapley interaction index (Grabisch & Roubens, 1999)). Given a subset of input variables $\mathcal{T} \subseteq \mathcal{N}$, the Shapley interaction index $I^{Shapley}(\mathcal{T})$ can be represented as $I^{Shapley}(\mathcal{T}) = \sum_{S \subseteq \mathcal{N} \setminus \mathcal{T}} \frac{1}{|S|+1} w_{S \cup \mathcal{T}}$. In other words, the index $I^{Shapley}(\mathcal{T})$ can be explained as uniformly allocating causal effects $w_{S'}$ s.t. $S' = S \cup \mathcal{T}$ to the compositional variables of S', if we treat the coalition of variables in \mathcal{T} as a single variable.

Theorem 4 (Connection to the Shapley Taylor interaction index (Sundararajan et al., 2020)). Given a subset of input variables $\mathcal{T} \subseteq \mathcal{N}$, the k-th order Shapley Taylor interaction index $I^{Shapley-Taylor}(\mathcal{T})$ can be represented as weighted sum of causal effects, i.e., $I^{Shapley-Taylor}(\mathcal{T}) = w_{\mathcal{T}}$ if $|\mathcal{T}| < k$; $I^{Shapley-Taylor}(\mathcal{T}) = \sum_{S \subseteq \mathcal{N} \setminus \mathcal{T}} {\binom{|S|+k}{k}}^{-1} w_{S \cup \mathcal{T}}$ if $|\mathcal{T}| = k$; and $I^{Shapley-Taylor}(\mathcal{T}) = 0$ if $|\mathcal{T}| > k$.

C THE PROOF OF THEOREM 1 IN THE MAIN PAPER

Theorem 1. Given a certain input \mathbf{x} , let the causal graph in Fig. 1 encodes 2^n causal patterns, i.e., $K = 2^n$ and $\Omega = 2^N = \{S : S \subseteq N\}$. If the causal effect w_S of each causal pattern $S \in \Omega$ is measured by the Harsanyi dividend (Harsanyi, 1963), i.e. $w_S \triangleq \sum_{S' \subseteq S} (-1)^{|S| - |S'|} \cdot v(\mathbf{x}_{S'})$, then the causal graph faithfully encodes the inference logic of the deep model, as follows.

$$\forall \mathcal{S} \subseteq \mathcal{N}, \ Y(\boldsymbol{x}_{\mathcal{S}}) = v(\boldsymbol{x}_{\mathcal{S}}) \tag{10}$$

More crucially, the Harsanyi dividend is the unique metric that satisfies the faithfulness requirement.

• *Proof:* We only need to prove the following two statements. (1) Necessity: the causal graph based on Harsanyi dividends w_S satisfies the faithfulness requirement $\forall S \subseteq \mathcal{N}, Y(x_S) = v(x_S)$. (2) Sufficiency: if there exists another metric \tilde{w}_S that also satisfies the faithfulness requirement, then, it is equivalent to the Harsanyi dividend, *i.e.* $\forall S \subseteq \mathcal{N}, \tilde{w}_S = w_S$.

According to the SCM in Eq. (2), we have $Y(\boldsymbol{x}_{\mathcal{S}}) = \sum_{\mathcal{S}' \in \Omega} w_{\mathcal{S}'} \cdot C_{\mathcal{S}'}(\boldsymbol{x}_{\mathcal{S}}) = \sum_{\mathcal{S}' \subseteq \mathcal{S}} w_{\mathcal{S}'}$. Therefore, the faithfulness requirement can be equivalently re-written as $\forall \mathcal{S} \subseteq \mathcal{N}, v(\boldsymbol{x}_{\mathcal{S}}) = \sum_{\mathcal{S}' \subseteq \mathcal{S}} w_{\mathcal{S}'}$.

Proof for necessity. According to the definition of the Harsanyi dividend, we have $\forall S \subseteq N$,

$$\sum_{\mathcal{S}' \subseteq \mathcal{S}} w_{\mathcal{S}'} = \sum_{\mathcal{S}' \subseteq \mathcal{S}} \sum_{\mathcal{L} \subseteq \mathcal{S}'} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} v(\boldsymbol{x}_{\mathcal{L}})$$
$$= \sum_{\mathcal{L} \subseteq \mathcal{S}} \sum_{\mathcal{S}' \subseteq \mathcal{S}: \mathcal{S}' \supseteq \mathcal{L}} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} v(\boldsymbol{x}_{\mathcal{L}})$$
$$= \sum_{\mathcal{L} \subseteq \mathcal{S}} \sum_{s' = |\mathcal{L}|} \sum_{\substack{\mathcal{S}' \subseteq \mathcal{S}: \mathcal{S}' \supseteq \mathcal{L} \\ |\mathcal{S}'| = s'}} (-1)^{s' - |\mathcal{L}|} v(\boldsymbol{x}_{\mathcal{L}})$$
$$= \sum_{\mathcal{L} \subseteq \mathcal{S}} v(\boldsymbol{x}_{\mathcal{L}}) \sum_{m=0}^{|\mathcal{S}| - |\mathcal{L}|} {|\mathcal{S}| - |\mathcal{L}| \choose m} (-1)^m = v(\boldsymbol{x}_{\mathcal{S}})$$

Proof for sufficiency. Suppose there exists another metric $\tilde{w}_{\mathcal{S}}$ that satisfies $\forall \mathcal{S} \subseteq \mathcal{N}, v(\boldsymbol{x}_{\mathcal{S}}) = \sum_{\mathcal{S}' \subseteq \mathcal{S}} \tilde{w}_{\mathcal{S}'}$. Then, we prove $\tilde{w}_{\mathcal{S}} = w_{\mathcal{S}}$ by induction on the number of variables $|\mathcal{S}|$ in the causal pattern.

(Basis step) When $|\mathcal{S}| = 0$, *i.e.* $\mathcal{S} = \emptyset$, we have $\tilde{w}_{\emptyset} = v(\boldsymbol{x}_{\emptyset}) = w_{\emptyset}$. Similarly, it can be directly derived that when $|\mathcal{S}| = 1$, *i.e.* $\mathcal{S} = \{i\}$, $\tilde{w}_{\{i\}} = v(\boldsymbol{x}_{\{i\}}) - v(\boldsymbol{x}_{\emptyset}) = w_{\{i\}}$; when $|\mathcal{S}| = 2$, *i.e.* $\mathcal{S} = \{i, j\}$, $\tilde{w}_{\{i,j\}} = v(\boldsymbol{x}_{\{i,j\}}) - v(\boldsymbol{x}_{\{i\}}) - v(\boldsymbol{x}_{\{j\}}) + v(\boldsymbol{x}_{\emptyset}) = w_{\{i,j\}}$.

(Induction step) Suppose $\tilde{w}_{S} = w_{S}$ holds for any S with $|S| = s \ge 2$. Then, for |S| = s + 1, we have

$$\begin{split} v(\boldsymbol{x}_{\mathcal{S}}) &= \sum_{\mathcal{S}' \subseteq \mathcal{S}} \tilde{w}_{\mathcal{S}'} = \tilde{w}_{\mathcal{S}} + \sum_{\mathcal{S}' \subseteq \mathcal{S}} \tilde{w}_{\mathcal{S}'} \\ &= \tilde{w}_{\mathcal{S}} + \sum_{\mathcal{S}' \subseteq \mathcal{S}} \sum_{\mathcal{L} \subseteq \mathcal{S}'} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} v(\boldsymbol{x}_{\mathcal{L}}) \qquad // \text{ by the induction assumption} \\ &= \tilde{w}_{\mathcal{S}} + \sum_{\mathcal{L} \subseteq \mathcal{S}} \sum_{\mathcal{S}' \subseteq \mathcal{S}: \mathcal{L} \subseteq \mathcal{S}'} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} \cdot v(\boldsymbol{x}_{\mathcal{L}}) \\ &= \tilde{w}_{\mathcal{S}} + \sum_{\mathcal{L} \subseteq \mathcal{S}} \sum_{\mathcal{S}' = |\mathcal{L}|} \sum_{\mathcal{S}' \subseteq \mathcal{S}: \mathcal{L} \subseteq \mathcal{S}'} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} \cdot v(\boldsymbol{x}_{\mathcal{L}}) \\ &= \tilde{w}_{\mathcal{S}} + \sum_{\mathcal{L} \subseteq \mathcal{S}} v(\boldsymbol{x}_{\mathcal{L}}) \sum_{\mathcal{S}' = |\mathcal{L}|}^{|\mathcal{S}| - 1} \binom{|\mathcal{S}| - |\mathcal{L}|}{|\mathcal{S}' - |\mathcal{L}|} (-1)^{|\mathcal{S}' - |\mathcal{L}|} \\ &= \tilde{w}_{\mathcal{S}} + \sum_{\mathcal{L} \subseteq \mathcal{S}} v(\boldsymbol{x}_{\mathcal{L}}) \underbrace{\sum_{\mathcal{M} = 0}^{|\mathcal{S}| - |\mathcal{L}| - 1} \binom{|\mathcal{S}| - |\mathcal{L}|}{m} (-1)^m}_{0 - (-1)^{|\mathcal{S}| - |\mathcal{L}|}} \\ &= \tilde{w}_{\mathcal{S}} - \sum_{\mathcal{L} \subseteq \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{L}|} v(\boldsymbol{x}_{\mathcal{L}}). \end{split}$$

In this way, we have

$$\tilde{w}_{\mathcal{S}} = v(\boldsymbol{x}_{\mathcal{S}}) + \sum_{\mathcal{L} \subsetneq \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{L}|} v(\boldsymbol{x}_{\mathcal{L}}) = \sum_{\mathcal{L} \subseteq \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{L}|} v(\boldsymbol{x}_{\mathcal{L}}) = w_{\mathcal{S}}$$

Therefore, the Harsanyi dividend is the unique metric that satisfies the faithfulness requirement.

D PROOFS OF AXIOMS AND THEOREMS FOR THE HARSANYI DIVIDEND

D.1 PROOFS OF AXIOMS

In this section, we prove that the Harsanyi dividend w_S satisfies the *efficiency*, *linearity*, *dummy*, *symmetry*, *anonymity*, *recursive*, and *interaction* distribution axioms.

(1) Efficiency axiom. The output score of a model can be decomposed into effects of different causal patterns, *i.e.* $v(x) = \sum_{S \subseteq N} w_S$.

• Proof: According to the definition of the Harsanyi dividend, we have

$$\sum_{\mathcal{S}\subseteq\mathcal{N}} w_{\mathcal{S}} = \sum_{\mathcal{S}\subseteq\mathcal{N}} \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} \cdot v(\boldsymbol{x}_{\mathcal{S}'})$$
$$= \sum_{\mathcal{S}'\subseteq\mathcal{N}} \sum_{\mathcal{S}:\mathcal{S}'\subseteq\mathcal{S}\subseteq\mathcal{N}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} \cdot v(\boldsymbol{x}_{\mathcal{S}'})$$
$$= \sum_{\mathcal{S}'\subseteq\mathcal{N}} \sum_{\substack{s=|\mathcal{S}'|}} \sum_{\substack{\mathcal{S}:\mathcal{S}'\subseteq\mathcal{S}\subseteq\mathcal{N} \\ |\mathcal{S}|=s}} (-1)^{s-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'})$$
$$= \sum_{\mathcal{S}'\subseteq\mathcal{N}} v(\boldsymbol{x}_{\mathcal{S}'}) \sum_{\substack{m=0}}^{n-|\mathcal{S}'|} \binom{n-|\mathcal{S}'|}{m} (-1)^{m}$$
$$= v(\boldsymbol{x}) \quad // \text{ the only case that cannot be cancelled out is } \mathcal{S}' = \mathcal{N}$$

(2) Linearity axiom. If we merge output scores of two models $t(\cdot)$ and $u(\cdot)$ as the output of model $v(\cdot)$, *i.e.* $\forall S \subseteq \mathcal{N}, v(\boldsymbol{x}_S) = t(\boldsymbol{x}_S) + u(\boldsymbol{x}_S)$, then the corresponding causal effects w_S^t and w_S^u can also be merged as $\forall S \subseteq \mathcal{N}, w_S^v = w_S^t + w_S^u$.

• Proof: According to the definition of the Harsanyi dividend, we have

$$w_{\mathcal{S}}^{v} = \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}})$$
$$= \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} [t(\boldsymbol{x}_{\mathcal{S}}) + u(\boldsymbol{x}_{\mathcal{S}})]$$
$$= \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} t(\boldsymbol{x}_{\mathcal{S}}) + \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} u(\boldsymbol{x}_{\mathcal{S}})$$
$$= w_{\mathcal{S}}^{t} + w_{\mathcal{S}}^{u}.$$

(3) **Dummy axiom.** If a variable $i \in \mathcal{N}$ is a dummy variable, *i.e.* $\forall S \subseteq \mathcal{N} \setminus \{i\}, v(\boldsymbol{x}_{S \cup \{i\}}) = v(\boldsymbol{x}_S) + v(\boldsymbol{x}_{\{i\}})$, then it has no causal effect with other variables, $\forall S \subseteq \mathcal{N} \setminus \{i\}, w_{S \cup \{i\}} = 0$.

• Proof: According to the definition of the Harsanyi dividend, we have

$$w_{\mathcal{S}\cup\{i\}} = \sum_{\mathcal{S}'\subseteq\mathcal{S}\cup\{i\}} (-1)^{|\mathcal{S}|+1-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'})$$

$$= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|+1-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'}) + \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'\cup\{i\}})$$

$$= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|+1-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'}) + \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} [v(\boldsymbol{x}_{\mathcal{S}}) + v(\boldsymbol{x}_{\{i\}})]$$

$$= \left[\sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|}\right] \cdot v(\boldsymbol{x}_{\{i\}})$$

$$= 0.$$

(4) Symmetry axiom. If input variables $i, j \in \mathcal{N}$ cooperate with other variables in the same way, $\forall S \subseteq \mathcal{N} \setminus \{i, j\}, v(\boldsymbol{x}_{S \cup \{i\}}) = v(\boldsymbol{x}_{S \cup \{j\}})$, then they have same causal effects with other variables, $\forall S \subseteq \mathcal{N} \setminus \{i, j\}, w_{S \cup \{i\}} = w_{S \cup \{j\}}$.

• Proof: According to the definition of the Harsanyi dividend, we have

$$\begin{split} w_{\mathcal{S}\cup\{i\}} &= \sum_{\mathcal{S}'\subseteq\mathcal{S}\cup\{i\}} (-1)^{|\mathcal{S}|+1-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'}) \\ &= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|+1-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'}) + \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'\cup\{i\}}) \\ &= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|+1-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'}) + \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'\cup\{j\}}) \\ &= \sum_{\mathcal{S}'\subseteq\mathcal{S}\cup\{j\}} (-1)^{|\mathcal{S}|+1-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'}) \\ &= w_{\mathcal{S}\cup\{j\}}. \end{split}$$

(5) Anonymity axiom. For any permutations π on \mathcal{N} , we have $\forall S \subseteq \mathcal{N}, w_S^v = w_{\pi S}^{\pi v}$, where $\pi S \triangleq \{\pi(i) | i \in S\}$, and the new model πv is defined by $(\pi v)(\boldsymbol{x}_{\pi S}) = v(\boldsymbol{x}_S)$. This indicates that causal effects are not changed by permutation.

• Proof: According to the definition of the Harsanyi dividend, we have

$$w_{\pi\mathcal{S}}^{\pi v} = \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} (\pi v) (\boldsymbol{x}_{\pi\mathcal{S}'})$$
$$= \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'})$$
$$= w_{\mathcal{S}}^{v}.$$

(6) **Recursive axiom.** The causal effects can be computed recursively. For $i \in \mathcal{N}$ and $\mathcal{S} \subseteq \mathcal{N} \setminus \{i\}$, the causal effect of the pattern $\mathcal{S} \cup \{i\}$ is equal to the causal effect of \mathcal{S} with the presence of i

minus the causal effect of S with the absence of i, *i.e.* $\forall S \subseteq \mathcal{N} \setminus \{i\}, w_{S \cup \{i\}} = w_{S|i \text{ present}} - w_S$. $w_{S|i \text{ present}}$ denotes the causal effect when the variable i is always present as a constant context, *i.e.* $w_{S|i \text{ present}} = \sum_{S' \subset S} (-1)^{|S| - |S'|} \cdot v(\boldsymbol{x}_{S' \cup \{i\}}).$

• Proof: According to the definition of the Harsanyi dividend, we have

$$w_{\mathcal{S}\cup\{i\}} = \sum_{\mathcal{S}'\subseteq\mathcal{S}\cup\{i\}} (-1)^{|\mathcal{S}|+1-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'})$$
$$= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|+1-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'}) + \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'\cup\{i\}})$$
$$= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'\cup\{i\}}) - \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'})$$
$$= w_{\mathcal{S}|i \text{ present}} - w_{\mathcal{S}}.$$

(7) Interaction distribution axiom. This axiom characterizes how causal effects are distributed for a class of "interaction functions" (Sundararajan et al., 2020). An interaction function $v_{\mathcal{T}}$ parameterized by a subset of variables \mathcal{T} is defined as follows. $\forall S \subseteq \mathcal{N}$, if $\mathcal{T} \subseteq S$, $v_{\mathcal{T}}(x_S) = c$; otherwise, $v_{\mathcal{T}}(x_S) = 0$. The function $v_{\mathcal{T}}$ purely models the causal effect of the pattern \mathcal{T} , because only if all variables in \mathcal{T} are present, the output value will be increased by *c*. The causal effects encoded in the function $v_{\mathcal{T}}$ satisfy $w_{\mathcal{T}} = c$, and $\forall S \neq \mathcal{T}$, $w_S = 0$.

• *Proof:* If $S \subsetneq T$, we have

$$w_{\mathcal{S}} = \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} \cdot \underbrace{v(\boldsymbol{x}_{\mathcal{S}'})}_{\forall \mathcal{S}' \subseteq \mathcal{S} \subsetneq \mathcal{T}, v(\boldsymbol{x}_{\mathcal{S}'}) = 0} = 0.$$

If S = T, we have

$$w_{\mathcal{S}} = w_{\mathcal{T}} = \sum_{\mathcal{S}' \subseteq \mathcal{T}} (-1)^{|\mathcal{T}| - |\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'})$$
$$= v(\mathcal{T}) + \sum_{\mathcal{S}' \subsetneq \mathcal{T}} (-1)^{|\mathcal{T}| - |\mathcal{S}'|} \underbrace{v(\boldsymbol{x}_{\mathcal{S}'})}_{=0} = c$$

If $S \subsetneq T$, we have

$$w_{\mathcal{S}} = \sum_{\substack{\mathcal{S}' \subseteq \mathcal{S} \\ \mathcal{S}' \subseteq \mathcal{S}}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'})$$
$$= c \cdot \sum_{\substack{\mathcal{S}' \subseteq \mathcal{S} \\ \mathcal{S}' \supseteq \mathcal{T}}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|}$$
$$= c \cdot \sum_{m=0}^{|\mathcal{S}| - |\mathcal{T}|} {|\mathcal{S}| - |\mathcal{T}| \choose m} (-1)^m = 0$$

D.2 PROOFS OF THEOREMS

In this section, we prove connections between the Harsanyi dividend w_S and several game-theoretic attributions/interactions. We first prove Theorem 5, which can be seen as the foundation for proofs of Theorem 2, 3, and 4.

Theorem 5 (Connection to the marginal benefit). Let $\Delta v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}}) = \sum_{\mathcal{T}' \subseteq \mathcal{T}} (-1)^{|\mathcal{T}| - |\mathcal{T}'|} v(\boldsymbol{x}_{\mathcal{T}' \cup \mathcal{S}})$ denote the marginal benefit of variables in $\mathcal{T} \subseteq \mathcal{N} \setminus \mathcal{S}$ given the environment \mathcal{S} . We have proven that $\Delta v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}})$ can be decomposed into the sum of causal effects inside \mathcal{T} and sub-environments of \mathcal{S} , *i.e.* $\Delta v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}}) = \sum_{\mathcal{S}' \subset \mathcal{S}} w_{\mathcal{T} \cup \mathcal{S}'}$.

• *Proof:* By the definition of the marginal benefit, we have

$$\Delta v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}}) = \sum_{\mathcal{L} \subseteq \mathcal{T}} (-1)^{|\mathcal{T}| - |\mathcal{L}|} v(\boldsymbol{x}_{\mathcal{L} \cup \mathcal{S}})$$
$$= \sum_{\mathcal{L} \subseteq \mathcal{T}} (-1)^{|\mathcal{T}| - |\mathcal{L}|} \sum_{\mathcal{K} \subseteq \mathcal{L} \cup \mathcal{S}} w_{\mathcal{K}} \quad // \text{ by Theorem 1}$$

$$\begin{split} &= \sum_{\mathcal{L}\subseteq\mathcal{T}} (-1)^{|\mathcal{T}| - |\mathcal{L}|} \sum_{\mathcal{L}'\subseteq\mathcal{L}} \sum_{\mathcal{S}'\subseteq\mathcal{S}} w_{\mathcal{L}'\cup\mathcal{S}'} \quad /\!\!/ \operatorname{since} \mathcal{L} \cap \mathcal{S} = \emptyset \\ &= \sum_{\mathcal{S}'\subseteq\mathcal{S}} \left[\sum_{\mathcal{L}\subseteq\mathcal{T}} (-1)^{|\mathcal{T}| - |\mathcal{L}|} \sum_{\mathcal{L}'\subseteq\mathcal{L}} w_{\mathcal{L}'\cup\mathcal{S}'} \right] \\ &= \sum_{\mathcal{S}'\subseteq\mathcal{S}} \left[\sum_{\mathcal{L}'\subseteq\mathcal{T}} \sum_{\mathcal{L}\subseteq\mathcal{T}} (-1)^{|\mathcal{T}| - |\mathcal{L}|} w_{\mathcal{L}'\cup\mathcal{S}'} \right] \\ &= \sum_{\mathcal{S}'\subseteq\mathcal{S}} \left[\underbrace{w_{\mathcal{S}'\cup\mathcal{T}}}_{\mathcal{L}'=\mathcal{T}} + \underbrace{\sum_{\mathcal{L}'\subseteq\mathcal{T}} \left(\sum_{l=|\mathcal{L}'|}^{|\mathcal{T}|} \left(|\mathcal{T}| - |\mathcal{L}'| \right) (-1)^{|\mathcal{T}| - |\mathcal{L}|} w_{\mathcal{L}'\cup\mathcal{S}'} \right) \right] \\ &= \sum_{\mathcal{S}'\subseteq\mathcal{S}} \left[w_{\mathcal{S}'\cup\mathcal{T}} + \sum_{\mathcal{L}'\subseteq\mathcal{T}} \left(w_{\mathcal{L}'\cup\mathcal{S}'} \cdot \underbrace{\sum_{l=|\mathcal{L}'|}^{|\mathcal{T}|} \left(|\mathcal{T}| - |\mathcal{L}'| \right) (-1)^{|\mathcal{T}| - |\mathcal{L}|} \right) \right] \\ &= \sum_{\mathcal{S}'\subseteq\mathcal{S}} w_{\mathcal{S}'\cup\mathcal{T}} \quad \Box \end{split}$$

In particular, if \mathcal{T} is a singleton set, *i.e.* $\mathcal{T} = \{i\}$, we can obtain a similar conclusion to (Ren et al., 2021) that $\Delta v_{\{i\}}(\boldsymbol{x}_{\mathcal{S}}) = \sum_{\mathcal{L} \subseteq \mathcal{S}} w_{\mathcal{L} \cup \{i\}}$.

Theorem 2 (Connection to the Shapley value). Let $\phi(i)$ denote the Shapley value (Shapley, 1953) of an input variable *i*. Then, the Shapley value $\phi(i)$ can be represented as a weighted sum of causal effects involving the variable *i*, *i.e.*, $\phi(i) = \sum_{S \subseteq N \setminus \{i\}} \frac{1}{|S|+1} w_{S \cup \{i\}}$. In other words, the effect of a causal pattern with *m* variables should be equally assigned to the *m* variables in the computation of Shapley values.

• Proof: By the definition of the Shapley value, we have

$$\begin{split} \phi(i) &= & \mathbb{E}_{\substack{M \leq \sum_{\substack{C \leq N \setminus \{i\} \\ |S| = m}}} [v(\boldsymbol{x}_{S \cup \{i\}}) - v(\boldsymbol{x}_{S})] \\ &= & \frac{1}{|\mathcal{N}|} \sum_{m=0}^{|\mathcal{N}| - 1} \frac{1}{(|\mathcal{M}| - 1)} \sum_{\substack{S \leq N \setminus \{i\} \\ |S| = m}} \left[v(\boldsymbol{x}_{S \cup \{i\}}) - v(\boldsymbol{x}_{S}) \right] \\ &= & \frac{1}{|\mathcal{N}|} \sum_{m=0}^{|\mathcal{N}| - 1} \frac{1}{(|\mathcal{M}| - 1)} \sum_{\substack{S \leq N \setminus \{i\} \\ |S| = m}} \Delta v_{\{i\}}(\boldsymbol{x}_{S}) \\ &= & \frac{1}{|\mathcal{N}|} \sum_{m=0}^{|\mathcal{N}| - 1} \frac{1}{(|\mathcal{M}| - 1)} \sum_{\substack{S \leq N \setminus \{i\} \\ |S| = m}} \left[\sum_{\mathcal{L} \subseteq S} w_{\mathcal{L} \cup \{i\}} \right] \quad \text{// by Theorem 5} \\ &= & \frac{1}{|\mathcal{N}|} \sum_{\substack{\mathcal{L} \subseteq \mathcal{N} \setminus \{i\}}} \sum_{m=0}^{|\mathcal{N}| - 1} \frac{1}{(|\mathcal{M}| - 1)} \sum_{\substack{S \leq N \setminus \{i\} \\ |S| = m}} w_{\mathcal{L} \cup \{i\}} \\ &= & \frac{1}{|\mathcal{N}|} \sum_{\substack{\mathcal{L} \subseteq \mathcal{N} \setminus \{i\}}} \sum_{m=|\mathcal{L}|} \frac{|\mathcal{N}|^{-1}}{(|\mathcal{M}| - 1)} \sum_{\substack{S \leq N \setminus \{i\} \\ |S| = m}} w_{\mathcal{L} \cup \{i\}} \quad \text{// since } S \supseteq \mathcal{L}, |S| = m \ge |\mathcal{L}| \\ &= & \frac{1}{|\mathcal{N}|} \sum_{\substack{\mathcal{L} \subseteq \mathcal{N} \setminus \{i\}}} \sum_{m=|\mathcal{L}|} \frac{|\mathcal{N}|^{-1}}{(|\mathcal{M}| - 1)} \frac{1}{(|\mathcal{M}| - 1)} \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - 1}{m - |\mathcal{L}|} \cdot w_{\mathcal{L} \cup \{i\}} \end{split}$$

$$= \frac{1}{|\mathcal{N}|} \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \{i\}} w_{\mathcal{L} \cup \{i\}} \underbrace{\sum_{k=0}^{|\mathcal{N}| - |\mathcal{L}| - 1} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{L}| + k}} \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - 1}{k}}{\alpha_{\mathcal{L}}}_{\alpha_{\mathcal{L}}}$$

Then, we leverage the following properties of combinatorial numbers and the Beta function to simplify the term $w_{\mathcal{L}} = \sum_{k=0}^{|\mathcal{N}| - |\mathcal{L}| - 1} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{L}| + k}} \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - 1}{k}$.

(i) A property of combinitorial numbers. $m \cdot \binom{n}{m} = n \cdot \binom{n-1}{m-1}$.

(ii) The definition of the Beta function. For p, q > 0, the Beta function is defined as $B(p,q) = \int_0^1 x^{p-1} (1-x)^{1-q} dx$.

(iii) Connections between combinitorial numbers and the Beta function.

• When $p, q \in \mathbb{Z}^+$, we have $B(p,q) = \frac{1}{q \cdot \binom{p+q-1}{p-1}}$.

• For $m, n \in \mathbb{Z}^+$ and n > m, we have $\binom{n}{m} = \frac{1}{m \cdot B(n-m+1,m)}$.

$$\begin{aligned} \alpha_{\mathcal{L}} &= \sum_{k=0}^{|\mathcal{N}| - |\mathcal{L}| - 1} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{L}| + k}} \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - 1}{k} \\ &= \sum_{k=0}^{|\mathcal{N}| - |\mathcal{L}| - 1} \binom{|\mathcal{N}| - |\mathcal{L}| - 1}{k} \cdot (|\mathcal{L}| + k) \cdot B(|\mathcal{N}| - |\mathcal{L}| - k, |\mathcal{L}| + k) \\ &= \sum_{k=0}^{|\mathcal{N}| - |\mathcal{L}| - 1} |\mathcal{L}| \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - 1}{k} \cdot (\frac{|\mathcal{N}| - |\mathcal{L}| - 1}{k}) \cdot B(|\mathcal{N}| - |\mathcal{L}| - k, |\mathcal{L}| + k) \quad \cdots \oplus \\ &+ \sum_{k=0}^{|\mathcal{N}| - |\mathcal{L}| - 1} k \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - 1}{k} \cdot S(|\mathcal{N}| - |\mathcal{L}| - k, |\mathcal{L}| + k) \quad \cdots \oplus \end{aligned}$$

Then, we solve and respectively. For , we have

$$\begin{split} \textcircled{D} &= \int_{0}^{1} |\mathcal{L}| \sum_{k=0}^{|\mathcal{N}| - |\mathcal{L}| - 1} \left(\frac{|\mathcal{N}| - |\mathcal{L}| - 1}{k} \right) \cdot x^{|\mathcal{N}| - |\mathcal{L}| - k - 1} \cdot (1 - x)^{|\mathcal{L}| + k - 1} \, dx \\ &= \int_{0}^{1} |\mathcal{L}| \cdot \underbrace{\left[\sum_{k=0}^{|\mathcal{N}| - |\mathcal{L}| - 1} \left(\frac{|\mathcal{N}| - |\mathcal{L}| - 1}{k} \right) \cdot x^{|\mathcal{N}| - |\mathcal{L}| - k - 1} \cdot (1 - x)^{k} \right]}_{=1} \cdot (1 - x)^{|\mathcal{L}| - 1} \, dx \\ &= \int_{0}^{1} |\mathcal{L}| (1 - x)^{|\mathcal{L}| - 1} \, dx = 1 \end{split}$$

For (2), we have

$$\begin{split} \textcircled{O} &= \sum_{k=1}^{|\mathcal{N}| - |\mathcal{L}| - 1} (|\mathcal{N}| - |\mathcal{L}| - 1) \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - 2}{k - 1} \cdot B(|\mathcal{N}| - |\mathcal{L}| - k, |\mathcal{L}| + k) \\ &= (|\mathcal{N}| - |\mathcal{L}| - 1) \sum_{k'=0}^{|\mathcal{N}| - |\mathcal{L}| - 2} \binom{|\mathcal{N}| - |\mathcal{L}| - 2}{k'} \cdot B(|\mathcal{N}| - |\mathcal{L}| - k' - 1, |\mathcal{L}| + k' + 1) \\ &= (|\mathcal{N}| - |\mathcal{L}| - 1) \int_{0}^{1} \sum_{k'=0}^{|\mathcal{N}| - |\mathcal{L}| - 2} \binom{|\mathcal{N}| - |\mathcal{L}| - 2}{k'} \cdot x^{|\mathcal{N}| - |\mathcal{L}| - k' - 2} \cdot (1 - x)^{|\mathcal{L}| + k'} dx \\ &= (|\mathcal{N}| - |\mathcal{L}| - 1) \int_{0}^{1} \underbrace{\left[\sum_{k'=0}^{|\mathcal{N}| - |\mathcal{L}| - 2} \binom{|\mathcal{N}| - |\mathcal{L}| - 2}{k'} \cdot x^{|\mathcal{N}| - |\mathcal{L}| - k' - 2} \cdot (1 - x)^{k'} \right]}_{=1} \cdot (1 - x)^{|\mathcal{L}|} dx \\ &= (|\mathcal{N}| - |\mathcal{L}| - 1) \int_{0}^{1} (1 - x)^{|\mathcal{L}|} dx = \frac{|\mathcal{N}| - |\mathcal{L}| - 1}{|\mathcal{L}| + 1} \end{split}$$





Figure 5: Connection between the Harsanyi dividend and Shapley values. If we uniformly assign the causal effect w_S with m variables in S, then the overall effect received by each variable equals to its Shapley value. See Theorem 2.

Figure 6: The cosine similarity between the accurate Shapley value ϕ and the estimated Shapley value $\tilde{\phi}$, when we used different ratios of salient patterns for estimation.

Hence, we have

$$\alpha_{\mathcal{L}} = (\mathbb{D} + (\mathbb{Z})) = 1 + \frac{|\mathcal{N}| - |\mathcal{L}| - 1}{|\mathcal{L}| + 1} = \frac{|\mathcal{N}|}{|\mathcal{L}| + 1}$$

Therefore, we proved $\phi(i) = \frac{1}{|\mathcal{N}|} \sum_{\mathcal{S} \subseteq \mathcal{N} \setminus \{i\}} \alpha_{\mathcal{L}} \cdot w_{\mathcal{L} \cup \{i\}} = \sum_{\mathcal{S} \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{|\mathcal{S}| + 1} \cdot w_{\mathcal{S} \cup \{i\}}.$

• Experimental verification: We conducted an experiment to verify Theorem 2, *i.e.* the Shapley value can be accurately approximated by the Harsanyi dividend. Let $\phi(i)$ denote the accurate Shapley value, and $\tilde{\phi}(i) = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{|S|+1} w_{S \cup \{i\}}$ denote the estimated Shapley value using the Harsanyi dividend w_S . In order to estimate the Shapley value via causal effects w_S , we first selected the most salient k causal patterns $\Omega_{\text{top-}k}^{\text{salient}}$. Then, according to Theorem 2, we computed the estimated Shapley value $\tilde{\phi}(i)$ as $\tilde{\phi}(i) = \sum_{S \in \Omega_{\text{top-}k}^{\text{salient}} : S \ni i} \frac{1}{|S|} w_S$. Figure 6 shows the cosine similarity between the accurate Shapley values $\phi = [\phi(1), \phi(2), \dots, \phi(n)] \in \mathbb{R}^n$ and the estimated Shapley values $\tilde{\phi} = [\tilde{\phi}(1), \tilde{\phi}(2), \dots, \tilde{\phi}(n)] \in \mathbb{R}^n$, when we used different ratios of salient patterns $\frac{k}{2^n}$ to approximate the Shapley value.

Theorem 2 (Connection to the Shapley interaction index). Given a subset of input variables $\mathcal{T} \subseteq \mathcal{N}$, $I^{\text{Shapley}}(\mathcal{T}) = \sum_{\mathcal{S} \subseteq \mathcal{N} \setminus \mathcal{T}} \frac{|\mathcal{S}|! |\mathcal{N}| - |\mathcal{S}| - |\mathcal{T}|)!}{(|\mathcal{N}| - |\mathcal{T}| + 1)!} \Delta v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}})$ denotes the Shapley interaction index (Grabisch & Roubens, 1999) of \mathcal{T} . We have proved that the Shapley interaction index can be represented as the weighted sum of causal effects involving \mathcal{T} , *i.e.*, $I^{\text{Shapley}}(\mathcal{T}) = \sum_{\mathcal{S} \subseteq \mathcal{N} \setminus \mathcal{T}} \frac{1}{|\mathcal{S}| + 1} w_{\mathcal{S} \cup \mathcal{T}}$. In other words, the index $I^{\text{Shapley}}(\mathcal{T})$ can be explained as uniformly allocating causal effects $w_{\mathcal{S}'}$ s.t. $\mathcal{S}' = \mathcal{S} \cup \mathcal{T}$ to the compositional variables of \mathcal{S}' , if we treat the coalition of variables in \mathcal{T} as a single variable.

• Proof:

Ι

$$\begin{aligned} ^{\text{Shapley}}(\mathcal{T}) &= \sum_{\mathcal{S} \subseteq \mathcal{N} \setminus \mathcal{T}} \frac{|\mathcal{S}|! (|\mathcal{N}| - |\mathcal{S}| - |\mathcal{T}|)!}{(|\mathcal{N}| - |\mathcal{T}| + 1)!} \Delta v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}}) \\ &= \frac{1}{|\mathcal{N}| - |\mathcal{T}| + 1} \sum_{m=0}^{|\mathcal{N}| - |\mathcal{T}|} \frac{1}{\binom{|\mathcal{N}| - |\mathcal{T}|}{m}} \sum_{\substack{S \subseteq \mathcal{N} \setminus \mathcal{T} \\ |S| = m}} \Delta v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}}) \\ &= \frac{1}{|\mathcal{N}| - |\mathcal{T}| + 1} \sum_{m=0}^{|\mathcal{N}| - |\mathcal{T}|} \frac{1}{\binom{|\mathcal{N}| - |\mathcal{T}|}{m}} \sum_{\substack{S \subseteq \mathcal{N} \setminus \mathcal{T} \\ |S| = m}} \left[\sum_{\mathcal{L} \subseteq \mathcal{S}} w_{\mathcal{L} \cup \mathcal{T}} \right] \\ &= \frac{1}{|\mathcal{N}| - |\mathcal{T}| + 1} \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \mathcal{T}} \sum_{m=|\mathcal{L}|}^{|\mathcal{N}| - |\mathcal{T}|} \frac{1}{\binom{|\mathcal{N}| - |\mathcal{T}|}{m}} \sum_{\substack{S \subseteq \mathcal{N} \setminus \mathcal{T} \\ |S| = m}} w_{\mathcal{L} \cup \mathcal{T}} \\ &= \frac{1}{|\mathcal{N}| - |\mathcal{T}| + 1} \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \mathcal{T}} \sum_{m=|\mathcal{L}|}^{|\mathcal{N}| - |\mathcal{T}|} \frac{1}{\binom{|\mathcal{N}| - |\mathcal{T}|}{m}} \binom{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|}{m - |\mathcal{L}|} w_{\mathcal{L} \cup \mathcal{T}} \end{aligned}$$

$$=\frac{1}{|\mathcal{N}|-|\mathcal{T}|+1}\sum_{\mathcal{L}\subseteq\mathcal{N}\setminus\mathcal{T}}w_{\mathcal{L}\cup\mathcal{T}}\underbrace{\sum_{k=0}^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|}\frac{1}{\binom{|\mathcal{N}|-|\mathcal{T}|}{|\mathcal{L}|+k}}\binom{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|}{k}}{k}$$

Just like the proof of Theorem 1, we leverage the properties of combinitorial numbers and the Beta function to simplify α_L .

$$\begin{aligned} \alpha_{\mathcal{L}} &= \sum_{k=0}^{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|} \frac{1}{\binom{|\mathcal{N}| - |\mathcal{T}|}{|\mathcal{L}| + k}} \binom{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|}{k} \\ &= \sum_{k=0}^{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|} \binom{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|}{k} \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{L}|}{k} \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{L}|}{k} \cdot \binom{|\mathcal{N}| - |\mathcal{L}|}$$

Then, we solve (1) and (2) respectively. For (1), we have

$$\begin{split} \textcircled{D} &= \int_{0}^{1} |\mathcal{L}| \sum_{k=0}^{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|} \left(|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| \right) \cdot x^{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - k} \cdot (1 - x)^{|\mathcal{L}| + k - 1} dx \\ &= \int_{0}^{1} |\mathcal{L}| \cdot \underbrace{\left[\sum_{k=0}^{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|} \left(|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| \right) \cdot x^{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - k} \cdot (1 - x)^{k} \right]}_{=1} \cdot (1 - x)^{|\mathcal{L}| - 1} dx \\ &= \int_{0}^{1} |\mathcal{L}| \cdot (1 - x)^{|\mathcal{L}| - 1} dx = 1 \end{split}$$

For (2), we have

$$\begin{split} & (\mathcal{D}) = \sum_{k=1}^{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|} (|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|) \binom{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - 1}{k - 1} \cdot B (|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - k + 1, |\mathcal{L}| + k) \\ &= (|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|) \sum_{k'=0}^{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - 1} \binom{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - 1}{k'} \cdot B (|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - k', |\mathcal{L}| + k' + 1) \\ &= (|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|) \int_{0}^{1} \sum_{k'=0}^{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - 1} \binom{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - 1}{k'} \cdot \frac{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - 1}{k'} \cdot (1 - x)^{|\mathcal{L}| + k'} dx \\ &= (|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|) \int_{0}^{1} \underbrace{\left[\sum_{k'=0}^{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - 1} \binom{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - 1}{k'} \cdot x^{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}| - 1} \cdot (1 - x)^{k'} \right]}_{=1} \cdot (1 - x)^{|\mathcal{L}|} dx \\ &= (|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|) \int_{0}^{1} (1 - x)^{|\mathcal{L}|} dx = \frac{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|}{|\mathcal{L}| + 1} \end{split}$$

Hence, we have

$$\alpha_{\mathcal{L}} = (\mathbb{D} + (\mathbb{Z})) = 1 + \frac{|\mathcal{N}| - |\mathcal{L}| - |\mathcal{T}|}{|\mathcal{L}| + 1} = \frac{|\mathcal{N}| - |\mathcal{T}| + 1}{|\mathcal{L}| + 1}$$

Therefore, we proved that $I^{\text{Shapley}}(\mathcal{T}) = \frac{1}{|\mathcal{N}| - |\mathcal{T}| + 1} \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \mathcal{T}} \alpha_{\mathcal{L}} \cdot w_{\mathcal{L} \cup \mathcal{T}} = \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \mathcal{T}} \frac{1}{|\mathcal{L}| + 1} w_{\mathcal{L} \cup \mathcal{T}}.$

Theorem 3 (Connection to the Shapley Taylor interaction index). Given a subset of input variables $\mathcal{T} \subseteq \mathcal{N}$, the *k*-th order Shapley Taylor interaction index $I^{\text{Shapley-Taylor}}(\mathcal{T})$ can be represented as weighted sum of causal effects, *i.e.*, $I^{\text{Shapley-Taylor}}(\mathcal{T}) = w_{\mathcal{T}}$ if $|\mathcal{T}| < k$; $I^{\text{Shapley-Taylor}}(\mathcal{T}) = \sum_{\mathcal{S} \subseteq \mathcal{N} \setminus \mathcal{T}} {\binom{|\mathcal{S}|+k}{k}}^{-1} w_{\mathcal{S} \cup \mathcal{T}}$ if $|\mathcal{T}| = k$; and $I^{\text{Shapley-Taylor}}(\mathcal{T}) = 0$ if $|\mathcal{T}| > k$.

• Proof: By the definition of the Shapley Taylor interaction index,

$$I^{\text{Shapley-Taylor}(k)}(\mathcal{T}) = \begin{cases} \Delta v_{\mathcal{T}}(\boldsymbol{x}_{\emptyset}) & \text{if } |\mathcal{T}| < k \\ \frac{k}{|\mathcal{N}|} \sum_{\mathcal{S} \subseteq \mathcal{N} \setminus \mathcal{T}} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{S}|}} \Delta v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}}) & \text{if } |\mathcal{T}| = k \\ 0 & \text{if } |\mathcal{T}| > k \end{cases}$$

When $|\mathcal{T}| < k$, by the definition of the Harsanyi dividend, we have

$$I^{\text{Shapley-Taylor}(k)}(\mathcal{T}) = \Delta v_{\mathcal{T}}(\boldsymbol{x}_{\emptyset}) = \sum_{\mathcal{L} \subseteq \mathcal{T}} (-1)^{|\mathcal{T}| - |\mathcal{L}|} \cdot v(\boldsymbol{x}_{\mathcal{L}}) = w_{\mathcal{T}}.$$

When $|\mathcal{T}| = k$, we have

$$\begin{split} I^{\text{Shapley-Taylor}(k)}(\mathcal{T}) &= \frac{k}{|\mathcal{N}|} \sum_{S \subseteq \mathcal{N} \setminus \mathcal{T}} \frac{1}{\binom{|\mathcal{N}| - 1}{|S|}} \cdot \Delta v_{\mathcal{T}}(\boldsymbol{x}_{S}) \\ &= \frac{k}{|\mathcal{N}|} \sum_{m=0}^{|\mathcal{N}| - k} \sum_{\substack{S \subseteq \mathcal{N} \setminus \mathcal{T} \\ |S| = m}} \frac{1}{\binom{|\mathcal{N}| - 1}{|S|}} \cdot \Delta v_{\mathcal{T}}(\boldsymbol{x}_{S}) \\ &= \frac{k}{|\mathcal{N}|} \sum_{m=0}^{|\mathcal{N}| - k} \sum_{\substack{S \subseteq \mathcal{N} \setminus \mathcal{T} \\ |S| = m}} \frac{1}{\binom{|\mathcal{N}| - 1}{|S|}} \left[\sum_{\mathcal{L} \subseteq S} w_{\mathcal{L} \cup \mathcal{T}} \right] \\ &= \frac{k}{|\mathcal{N}|} \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \mathcal{T}} \sum_{m=|\mathcal{L}|}^{|\mathcal{N}| - k} \frac{1}{\binom{|\mathcal{N}| - 1}{|S|}} \sum_{\substack{S \subseteq \mathcal{N} \setminus \mathcal{T} \\ |S| = m}} w_{\mathcal{L} \cup \mathcal{T}} \\ &= \frac{k}{|\mathcal{N}|} \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \mathcal{T}} \sum_{m=|\mathcal{L}|}^{|\mathcal{N}| - k} \frac{1}{\binom{|\mathcal{N}| - 1}{|S|}} \binom{|\mathcal{N}| - |\mathcal{L}| - k}{m - |\mathcal{L}|} w_{\mathcal{L} \cup \mathcal{T}} \\ &= \frac{k}{|\mathcal{N}|} \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \mathcal{T}} w_{\mathcal{L} \cup \mathcal{T}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - |\mathcal{L}| - k}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - |\mathcal{L}| - k}{m}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - |\mathcal{L}| - k}}_{m=|\mathcal{L}|} (|\mathcal{N}| - |\mathcal{L}| - k) \\ &= \frac{k}{|\mathcal{N}|} \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \mathcal{T}} w_{\mathcal{L} \cup \mathcal{T}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - |\mathcal{L}| - k}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - |\mathcal{L}| - k}{m}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - |\mathcal{L}| - k}}_{m} (|\mathcal{N}| - |\mathcal{L}| - k) \\ &= \frac{k}{|\mathcal{N}|} \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \mathcal{T}} w_{\mathcal{L} \cup \mathcal{T}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - |\mathcal{L}| - k}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{L}| + m}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{L}| - k}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{L}| - k}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{L}| - k}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{L}| - 1}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{N}| - 1}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{L}| - 1}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{L}| - 1}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - 1}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - 1}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - 1}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \frac{1}{\binom{|\mathcal{N}| - 1}} \underbrace{\sum_{m=0}^{|\mathcal{N}| - 1}}_{m=|\mathcal{L}|} \underbrace{\sum_{m=0}^{|\mathcal{N}$$

Just like the proof of Theorem 1, we leverage the properties of combinitorial numbers and the Beta function to simplify α_L .

$$\begin{aligned} \alpha_{\mathcal{L}} &= \sum_{m=0}^{|\mathcal{N}| - |\mathcal{L}| - k} \frac{1}{\binom{|\mathcal{N}| - 1}{|\mathcal{L}| + m}} \binom{|\mathcal{N}| - |\mathcal{L}| - k}{m} \\ &= \sum_{m=0}^{|\mathcal{N}| - |\mathcal{L}| - k} \binom{|\mathcal{N}| - |\mathcal{L}| - k}{m} \cdot \binom{|\mathcal{L}| - k}{m} \cdot \binom{|\mathcal{L}| + m}{m} \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - k}{m} \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - k}{m} \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - m}{m} \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - k}{m} \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - k}{m} \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - m}{m} \cdot \binom{|\mathcal{N}| - |\mathcal{N}| - m}{m} \cdot \binom{|\mathcal{N}| - |\mathcal{N}| - m}{m} \cdot \binom{|\mathcal{N}| - |\mathcal{N}| - m}{m} \cdot \binom{|\mathcal{N}| - m}{$$

Then, we solve and respectively. For , we have

$$\begin{split} \textcircled{\textbf{T}} &= \int_{0}^{1} |\mathcal{L}| \cdot \sum_{m=0}^{|\mathcal{N}| - |\mathcal{L}| - k} \binom{|\mathcal{N}| - |\mathcal{L}| - k}{m} \cdot x^{|\mathcal{N}| - |\mathcal{L}| - m - 1} \cdot (1 - x)^{|\mathcal{L}| + m - 1} dx \\ &= \int_{0}^{1} |\mathcal{L}| \cdot \underbrace{\left[\sum_{m=0}^{|\mathcal{N}| - |\mathcal{L}| - k} \binom{|\mathcal{N}| - |\mathcal{L}| - k}{m} \cdot x^{|\mathcal{N}| - |\mathcal{L}| - m - k} \cdot (1 - x)^{m} \right]}_{=1} \cdot x^{k-1} \cdot (1 - x)^{|\mathcal{L}| - 1} dx \end{split}$$

$$= \int_0^1 |\mathcal{L}| \cdot x^{k-1} \cdot (1-x)^{|\mathcal{L}|-1} \, dx = |\mathcal{L}| \cdot B(k, |\mathcal{L}|) = \frac{1}{\binom{|\mathcal{L}|+k-1}{k-1}}$$

For ②, we have

$$\begin{split} & (\mathcal{D}) = \sum_{m=1}^{|\mathcal{N}| - |\mathcal{L}| - k} (|\mathcal{N}| - |\mathcal{L}| - k) \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - k - 1}{m - 1} \cdot B(|\mathcal{N}| - |\mathcal{L}| - m, |\mathcal{L}| + m) \\ &= \sum_{m'=0}^{|\mathcal{N}| - |\mathcal{L}| - k - 1} (|\mathcal{N}| - |\mathcal{L}| - k) \cdot \binom{|\mathcal{N}| - |\mathcal{L}| - k - 1}{m'} \cdot B(|\mathcal{N}| - |\mathcal{L}| - m' - 1, |\mathcal{L}| + m' + 1) \\ &= \int_{0}^{1} (|\mathcal{N}| - |\mathcal{L}| - k) \sum_{m'=0}^{|\mathcal{N}| - |\mathcal{L}| - k - 1} \binom{|\mathcal{N}| - |\mathcal{L}| - k - 1}{m'} \cdot x^{|\mathcal{N}| - |\mathcal{L}| - m' - 2} \cdot (1 - x)^{|\mathcal{L}| + m'} dx \\ &= \int_{0}^{1} (|\mathcal{N}| - |\mathcal{L}| - k) \underbrace{\left[\sum_{m'=0}^{|\mathcal{N}| - |\mathcal{L}| - k - 1} \binom{|\mathcal{N}| - |\mathcal{L}| - k - 1}{m'} \cdot x^{|\mathcal{N}| - |\mathcal{L}| - m' - 4} \cdot (1 - x)^{m'} \right]}_{=1} \cdot x^{k-1} \cdot (1 - x)^{|\mathcal{L}|} dx \\ &= \int_{0}^{1} (|\mathcal{N}| - |\mathcal{L}| - k) \cdot x^{k-1} \cdot (1 - x)^{|\mathcal{L}|} dx = (|\mathcal{N}| - |\mathcal{L}| - k) \cdot B(k, |\mathcal{L}| + 1) \\ &= \frac{|\mathcal{N}| - |\mathcal{L}| - k}{(|\mathcal{L}| + 1)\binom{|\mathcal{L}| + k}{k-1}} \end{split}$$

Hence, we have

$$\begin{split} \alpha_{\mathcal{L}} = & (\mathbb{D} + \mathbb{Q}) = \frac{1}{\binom{|\mathcal{L}| + k - 1}{k - 1}} + \frac{|\mathcal{N}| - |\mathcal{L}| - k}{(|\mathcal{L}| + 1)\binom{|\mathcal{L}| + k}{k - 1}} \\ &= \frac{|\mathcal{L}|! \cdot (k - 1)!}{(|\mathcal{L}| + k - 1)!} + \frac{|\mathcal{N}| - |\mathcal{L}| - k}{|\mathcal{L}| + 1} \cdot \frac{(|\mathcal{L}| + 1)! \cdot (k - 1)!}{(|\mathcal{L}| + k)!} \\ &= \frac{|\mathcal{L}|! \cdot (k - 1)!}{(|\mathcal{L}| + k - 1)!} + \frac{|\mathcal{N}| - |\mathcal{L}| - k}{|\mathcal{L}| + k} \cdot \frac{|\mathcal{L}|! \cdot (k - 1)!}{(|\mathcal{L}| + k - 1)!} \\ &= \left[1 + \frac{|\mathcal{N}| - |\mathcal{L}| - k}{|\mathcal{L}| + k}\right] \cdot \frac{|\mathcal{L}|! \cdot (k - 1)!}{(|\mathcal{L}| + k - 1)!} \\ &= \frac{|\mathcal{N}|}{|\mathcal{L}| + k} \cdot \frac{|\mathcal{L}|! \cdot (k - 1)!}{(|\mathcal{L}| + k - 1)!} \\ &= \frac{|\mathcal{N}|}{k} \cdot \frac{|\mathcal{L}|! \cdot k!}{(|\mathcal{L}| + k - 1)!} \\ &= \frac{|\mathcal{N}|}{k} \cdot \frac{1}{\binom{|\mathcal{L}| + k}{k}} \end{split}$$

Therefore, we proved that when $|\mathcal{T}| = k$, $I^{\text{Shapley-Taylor}}(\mathcal{T}) = \frac{k}{|\mathcal{N}|} \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \mathcal{T}} \alpha_{\mathcal{L}} \cdot w_{\mathcal{L} \cup \mathcal{T}} = \frac{k}{|\mathcal{N}|} \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \mathcal{T}} \frac{|\mathcal{N}|}{k} \cdot \frac{1}{\binom{|\mathcal{L}|+k}{k}} \cdot w_{\mathcal{L} \cup \mathcal{T}} = \sum_{\mathcal{L} \subseteq \mathcal{N} \setminus \mathcal{T}} \binom{|\mathcal{L}|+k}{k}^{-1} w_{\mathcal{L} \cup \mathcal{T}}.$

E SIMPLIFYING THE EXPLANATION USING THE OPTIMAL BASELINE VALUES

In this section, we proved that *the sparsity of causal patterns does not only depend on the deep model itself, but it is also determined by the choice of baseline values,* as is mentioned in Section 3.2 of the main paper. Lemma 1 and Theorem 6 below proved that baseline values will affect the overall sparsity of salient causal patterns.

Lemma 1. The inference of the deep model on an input sample x can be written in the form of $v(x) = \sum_{S \subseteq \mathcal{N}} w'_S \cdot \prod_{i \in S} (x_i - r_i^*)$, where r_i^* is the ground-truth baseline value of the *i*-th input variable.

• **Proof:** According to the SCM in Eq. (2), we have $v(x) = \sum_{S \subseteq \mathcal{N}} w_S \cdot C_S(x)$ for a specific input sample x. For the effect of each causal pattern $w_S \cdot C_S(x)$, we can rewrite it as follows.

$$w_{\mathcal{S}} \cdot C_{\mathcal{S}}(\boldsymbol{x}) = \frac{w_{\mathcal{S}}}{\lambda} \cdot \prod_{i \in \mathcal{S}} (x_i - r_i^*)$$

where r_i^* denotes the ground-truth baseline value of the *i*-th input variable. $\lambda = \prod_{j \in S} (x_j - r_j^*)$ denotes the product of all variables that are not masked. Then, for any masked input sample $\boldsymbol{x}^{\text{masked}}$, if the *i*-th variable ($i \in S$) is absent (is replaced by its baseline value), then the pattern is deactivated and $w_S \cdot C_S(\boldsymbol{x}^{\text{masked}}) = \frac{w_S}{\lambda} \cdot \prod_{i \in S} (x_i^{\text{masked}} - r_i^*) = 0$.

Therefore, the causal effect of the pattern S can be represented as $w_S \cdot C_S(\mathbf{x}) = w'_S \prod_{i \in S} (x_i - r_i^*)$, and $w'_S = \frac{w_S}{\lambda}$. The model output a specific input sample (may be a masked sample) can be written as $v(\mathbf{x}) = \sum_{S \subseteq \mathcal{N}} w'_S \prod_{i \in S} (x_i - r_i^*)$.

Theorem 6. Given an input sample x, according to Lemma 1, the effect of a single causal pattern S can be represented as the function $v_S(x) = w'_S \prod_{j \in S} (x_j - r_j^*)$. Accordingly, ground-truth baseline values of variables are obviously $\{r_i^*\}$. This is because

(1) Setting any variable $x_j = r_j^*$ will deactivate this pattern. If we explain the pattern using groundtruth baseline values $\{r_i^*\}$, there is only one causal pattern S with a non-zero causal effect w_S .

(2) However, if we use m' incorrect baselines values $\{r'_j\}$ to replace correct ones, $\sum_{j \in S} \mathbb{1}_{r'_j \neq r^*_j} = m'$, then the function will be explained to contain at most $2^{m'}$ causal patterns. Specifically, a total of $\binom{m'}{k-|S|+m'}$ causal patterns of the k-th order emerge in this case, $k \geq |S| - m'$. The order k of a causal pattern means that this causal pattern contains k variables.

• *Proof:* Without loss of generality, let us consider an input sample x, with $\forall j \in S, x_j \neq r_j^*$. Based on the ground-truth baseline vaule $\{r_j^*\}$, we have

(1)
$$v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}) = w'_{\mathcal{S}} \prod_{j \in \mathcal{S}} (x_j - r_j^*) \neq 0$$

(2) $\forall \mathcal{S}' \subsetneq \mathcal{S}, v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}'}) = w'_{\mathcal{S}} \prod_{j \in \mathcal{S}'} (x_j - r_j^*) \prod_{j \in \mathcal{S} \setminus \mathcal{S}'} (r_j^* - r_j^*) = 0.$

Therefore, the causal effect $w_{\mathcal{S}}$ of the pattern \mathcal{S} is $w_{\mathcal{S}} = \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}| - |\mathcal{S}'|} v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}'}) = v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}) \neq 0$. For all the other patterns $\mathcal{S}' \subsetneq \mathcal{S}$, we have $w_{\mathcal{S}'} = \sum_{\mathcal{L} \subseteq \mathcal{S}'} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{L}}) = \sum_{\mathcal{L} \subseteq \mathcal{S}'} 0 = 0$. Hence, there is only one causal pattern \mathcal{S} with a non-zero effect $w_{\mathcal{S}}$.

In comparison, if we use m' incorrect baseline values $\{r'_j\}$, where $\sum_{j \in S} \mathbb{1}_{r'_j \neq r^*_j} = m'$, then the function will be explained to contain at most $2^{m'}$ causal patterns. For the simplicity of notations, let $S = \{1, 2, ..., m\}$, and $r'_1 = r^*_1 + \epsilon_1, ..., r'_{m'} = r^*_{m'} + \epsilon_{m'}$, where $\epsilon_1, ..., \epsilon_{m'} \neq 0$. Let $\mathcal{T} = \{1, 2, ..., m'\}$. In this case, we have $(1) v_S(\mathbf{x}_S) \neq 0$

(1) ${}^{|\mathcal{S}(\boldsymbol{x}\mathcal{S}') \neq 0}$ (2) $\forall \mathcal{S}' \subsetneq \mathcal{S}, |\mathcal{S}'| < m - m', v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}'}) = w'_{\mathcal{S}} \prod_{j \in \mathcal{S}'} (x_j - r_j^*) \prod_{j \in \mathcal{S} \setminus \mathcal{S}'} (r'_j - r_j^*)$. Because $|\mathcal{S}| - |\mathcal{S}'| > m'$, there is at least one variable with ground-truth baseline value in $\mathcal{S} \setminus \mathcal{S}'$. Therefore, $v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}'}) = 0$. Furthermore, the causal effect of the pattern \mathcal{S} satisfies $w_{\mathcal{S}'} = \sum_{\mathcal{L} \subseteq \mathcal{S}'} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{L}}) = 0$ (3) $\forall \mathcal{S}' \subsetneq \mathcal{S}, |\mathcal{S}'| = k \ge m - m', v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}'}) = w'_{\mathcal{S}} \prod_{j \in \mathcal{S}'} (x_j - r_j^*) \prod_{j \in \mathcal{S} \setminus \mathcal{S}'} (r'_j - r_j^*)$. If $\mathcal{S} \setminus \mathcal{T} \subseteq \mathcal{S}'$, then $\mathcal{S} \setminus \mathcal{S}' \subseteq \mathcal{T}$ and $v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}'}) \neq 0$. Otherwise, $v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}'}) = 0$. Then,

$$\begin{split} w_{\mathcal{S}'} &= \sum_{\mathcal{L}\subseteq \mathcal{S}'} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{L}}) \\ &= \sum_{\mathcal{L}\subseteq \mathcal{S}', |\mathcal{L}| < m-m'} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{L}}) + \sum_{\mathcal{L}\subseteq \mathcal{S}', \mathcal{L} \ge m-m'} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{L}}) \\ &= 0 + \sum_{\mathcal{L}\subseteq \mathcal{S}', \mathcal{L} \ge m-m', \mathcal{L}\supseteq \mathcal{S} \setminus \mathcal{T}} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{L}}) + \sum_{\mathcal{L}\subseteq \mathcal{S}', \mathcal{L} \ge m-m', \mathcal{L}\supseteq \mathcal{S} \setminus \mathcal{T}} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{L}}) \\ &= \sum_{\mathcal{L}\subseteq \mathcal{S}', \mathcal{L} \ge m-m', \mathcal{L}\supseteq \mathcal{S} \setminus \mathcal{T}} (-1)^{|\mathcal{S}'| - |\mathcal{L}|} v_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{L}}) \end{split}$$

If the above $w_{\mathcal{S}'} = 0$, it indicates that $\mathcal{S} \setminus \mathcal{T} \not\subseteq \mathcal{S}'$. In this case, there is no subset $\mathcal{L} \subseteq \mathcal{S}'$ s.t. $\mathcal{S} \setminus \mathcal{T} \subseteq \mathcal{L}$. In other words, only if $\mathcal{S} \setminus \mathcal{T} \subseteq \mathcal{S}', w_{\mathcal{S}'} \neq 0$. In this way, a total of $\binom{m'}{k-(|\mathcal{S}|-m')}$ causal patterns of the *k*-th order emerge, where the order *k* of a causal pattern means that this causal pattern \mathcal{S}' contains $k = |\mathcal{S}'|$ variables. There are totally $\sum_{k=|\mathcal{S}|-m'}^{m} \binom{m'}{k-(|\mathcal{S}|-m')} = 2^{m'}$ causal patterns in \boldsymbol{x} . For example, if the input x is given as follows,

$$x_i = \begin{cases} r_i^* + 2\epsilon_i, & i \in \mathcal{T} = \{1, \dots, m'\}\\ r_i^* + \epsilon_i, & i \in \mathcal{S} \setminus \mathcal{T} = \{m' + 1, \dots, m\} \end{cases}$$

where $\epsilon_i \neq 0$ are arbitrary non-zero scalars. In this case, we have $\forall S' \subseteq T, w_{S' \cup \{m'+1,...,m\}} = \epsilon_1 \epsilon_2 \dots \epsilon_m \neq 0$. Besides, if $\{m'+1, \dots, m\} \not\subseteq S'$, we have $w_{S'} = 0$. In this way, there are totally $2^{m'}$ causal patterns in x.

F POTENTIAL ALTERNATIVE SETTINGS FOR BASELINE VALUES

This section discusses potential alternative settings for baseline values, as is mentioned in Section 3.2 of the main paper. Baseline values are used to represent absent states of variables in the computation of $v(\mathbf{x}_{S})$. To this end, many recent studies set baseline values from a heuristic perspective, as follows.

• Mean baseline values (Dabkowski & Gal, 2017). The baseline value of each input variable is set to the mean value of this variable over all samples, *i.e.* $\forall i \in \mathcal{N}, r_i = \mathbb{E}_{\mathbf{x}}[x_i]$.

• Zero baseline values (Ancona et al., 2019; Sundararajan et al., 2017). The baseline value of each input variable is set to zero, *i.e.* $\forall i \in \mathcal{N}, r_i = 0$.

• Blurring input samples. In the computation of $v(x_S)$, some studies (Fong & Vedaldi, 2017; Fong et al., 2019) removed variables in the input image by blurring the value of each input variable x_i $(i \in \mathcal{N} \setminus S)$ based on a Gaussian kernel.

However, it still remains an open problem to define optimal baseline values. As is discussed in Section E, the optimal baseline values provides a perspective that simplifies the explanation of a deep model, thereby boosting the conciseness of the explanation. Therefore, in this paper, we learn the optimal baseline values that enhance the conciseness of the explanation based on Eq. (6) of the main paper. Specifically, we initialize the baseline value r_i as the mean value of the variable *i* over all samples, and then we optimize r_i to minimize Eq. (6) in the main paper while constraining it within a relatively small range, *i.e.*, $||r_i - r_i^{initial}||^2 \le \tau$, to represent the absence state.

G SIMPLIFYING THE EXPLANATION USING THE MINIMUM DESCRIPTION LENGTH PRINCIPLE

In this section, we discuss the algorithm of extracting common coalitions to minimize the total description length in Eq. (8) of the main paper. Given an AOG g and input variables \mathcal{N} , let $\mathcal{M} = \mathcal{N} \cup \Omega^{\text{coalition}}$ denote the set of all terminal nodes and AND nodes in the bottom two layers (e.g. $\mathcal{M} = \mathcal{N} \cup \Omega^{\text{coalition}} = \{x_1, x_2, ..., x_6\} \cup \{\alpha, \beta\}$ in Fig. 1(d) of the main paper). The total description length $L(g, \mathcal{M})$ was given in Eq. (8) of the main paper.

In order to minimize $L(g, \mathcal{M})$, we used the greedy strategy to iteratively extract common coalitions of input variables. In each iteration, we chose the coalition $\alpha \subseteq \mathcal{N}$ which most efficiently decreased the total description length. Then we took this coalition as an AND node, and added it into $\Omega^{\text{coalition}}$ in the third layer of the AOG. The efficiency of a coalition α w.r.t. the decrease of the total description length was defined as follows.

$$\delta(\alpha) = \frac{\Delta L}{|\alpha|} = \frac{L(g, \mathcal{M} \cup \{\alpha\}) - L(g, \mathcal{M})}{|\alpha|},\tag{11}$$

where $L(g, \mathcal{M})$ denoted the total description length without using the newly added coalition α , and $L(g, \mathcal{M} \cup \{\alpha\})$ denoted the total description when we added the node α to further simplify the description of g. $|\alpha|$ denotes the number of input variables in α . We iteratively extracted the most efficient coalition α to minimize the total description length. The extracting process stopped when there was no new coalition α could further reduce the total description length (*i.e.* $\forall \alpha \notin \mathcal{M}, L(g, \mathcal{M} \cup \{\alpha\}) - L(g, \mathcal{M}) > 0$), or the most efficient α was not shared by multiple patterns. Algorithm 1 shows the pseudo-code of this algorithm.

	Algorithm 1: The greedy algorithm to minimize total description length $L(g, \mathcal{M})$								
	Input: The set of leaf nodes N, the set of causal patterns Ω , the causal effects of these patterns								
	$\{w_{\mathcal{S}}\}_{\mathcal{S}\in\Omega}$, the maximum iteration times T								
	Output: The set of nodes in the bottom two layers $\mathcal{M} = \mathcal{N} \cup \Omega^{\text{coalition}}$								
1	Initialize $\Omega^{\text{coalition}} = \emptyset$ and $\mathcal{M} = \mathcal{N} \cup \Omega^{\text{coalition}}$								
2	for iteration 1 to T do								
3	foreach <i>possible coalition</i> $\alpha \subseteq \mathcal{N}$ do								
4	Calculate the efficiency $\delta(\alpha)$ according to Eq. (11)								
5	end								
6	Select α as the coalition whose $\delta(\alpha)$ is the smallest								
7	if $\delta(\alpha) > 0$ or α co-appears in only one pattern then								
8	break								
9	end								
10	Update AND nodes, $\Omega^{\text{coalition}} \leftarrow \Omega^{\text{coalition}} \cup \{\alpha\}$								
11	Rewrite each pattern $S \in \Omega$ according to $\mathcal{M} = \mathcal{N} \cup \Omega^{\text{coalition}}$								
12	end								
13	return $\mathcal{M} = \mathcal{N} \cup \Omega^{coalition}$								

H MORE EXPERIMENTAL DETAILS, RESULTS, AND DISCUSSIONS

H.1 DATASETS AND MODELS

Datasets. We conducted experiments on both tasks of natural language processing and the classification/regression tasks based on tabular datasets. For natural language processing, we used the SST-2 dataset (Socher et al., 2013) for sentiment prediction and the CoLA dataset (Warstadt et al., 2019) for linguistic acceptability. For tabular datasets, we used the UCI census income dataset (*census*) (Dua & Graff, 2017), the UCI bike sharing dataset (*bike*) (Dua & Graff, 2017), and the UCI TV news channel commercial detection dataset (*TV news*) (Dua & Graff, 2017). We followed (Covert et al., 2020; Covert & Lee, 2021) to conduct data pre-processing for these tabular datasets. We also normalized data in each dataset to zero mean and unit variance.

Models. We trained LSTMs and CNNs based on NLP datasets. The LSTM was unidirectional and had two layers, with a hidden layer of size 100. The architecture of the CNN was the same as the network architecture in (Rakhlin, 2016). Besides, for tabular datasets, we followed (Covert et al., 2020; Covert & Lee, 2021) to train LightGBMs (Ke et al., 2017), XGBoost (Chen & Guestrin, 2016), and two-layer MLPs (namely *MLP-2*). We also trained five-layer MLPs (namely *MLP-5*) and five layer MLPs with skip-connections (namely *ResMLP-5*) on these datasets. For the ResMLP-5, we added a skip connection to each fully connected layer of the MLP-5. Figure 7 shows the architecture of the ResMLP-5. The hidden layers in MLP-5 and ResMLP-5 on each tabular dataset via adversarial training (Madry et al., 2018). During adversarial training, adversarial examples $x^{adv} = x + \delta$ were generated by the ℓ_{∞} PGD attack, where $\|\delta\|_{\infty} \leq 0.1$. The attack was iterated for 20 steps with the step size of 0.01.



Figure 7: The architecture of the ResMLP-5.

Accuracy of models. Table 4 reports the classification accuracy of models trained on the TV news dataset, the classification accuracy of models trained on the census dataset, and the mean squared error of models trained on the bike dataset. Table 5 reports the classification accuracy of models trained on the CoLA dataset and the SST-2 dataset.

Dotocot	M	LP-2	M	LP-5	Res	MLP-5	VCDaaat	LishtCDM	Dotocot	ISTM	CNN
Dataset	normal	adversarial	normal	adversarial	normal	adversarial	AGDOOSI	LIGHIGENI	Dataset	LSIM	CININ
TV news	83.11%	78.49%	79.86%	80.24%	79.01%	80.13%	84.48%	84.19%	CoLA	64.42%	65.79%
census	79.91%	75.77%	78.96%	77.79%	80.49%	77.99%	87.35%	87.54%	SST-2	86 83%	78 19%
bike	-	-	2161.47	3080.73	2149.43	2708.59	1623.71	-	551-2	00.05 //	70.1770

Table 4: Classification accuracy (on TV news and census dataset) and
mean squared error (on bike dataset) of different models.Table 5: Accuracy of models
trained on NLP datasets.

H.2 Settings of $v(\cdot)$ in experiments

In the computation of Harsanyi dividend, people can apply different confidence scores to implement $v(\cdot)$. For example, Lundberg & Lee (2017) directly set $v(x_S) = p(y = y^{truth} | x_S)$ in the classification task. Covert et al. (2020) computed $v(x_S)$ as the cross-entropy loss of the sample x_S . In this paper, we used $v(x_S) = \log \frac{p(y=1|x_S)}{1-p(y=1|x_S)}$ for models learned on binary classification tasks, where $p(y=1|x_S)$ denoted the output probability of the positive class. Specifically, y = 1 denoted the prediction of "is commercial", "income > 50k", "positive sentiment", and "grammatically correct" in the TV news dataset, the census dataset, the SST-2 dataset, and the CoLA dataset, respectively. For models learned on regression tasks, we directly computed $v(x_S)$ as the scalar output of the model on x_S .

H.3 DISCOVERING THE SPARSITY OF CAUSAL EFFECTS

This section provided more experimental results to verify the sparsity of causal effects, which is mentioned in Section 3.2. To this end, for each input sample, we computed causal effects w_S of all 2^n patterns $S \in 2^N$ encoded by a deep model. In the computation of causal effects, the baseline values of input variables were set according to Section 3.2. Just like experiments in Section 4.2, we computed the relative strength of causal effects $\frac{|w_S|}{\max_{S' \subseteq N} |w_{S'}|}$, which re-scaled the range of causal effects in each single sample. Fig. 8(a-c,right) show histograms of the relative strength of causal effects in each single sample. Fig. 8(a-c,left) show histograms of relative strength, which are averaged on different samples in each dataset. We found that the deep model usually encoded a very small number of causal patterns with significant causal effects w_S extracted from the deep model were usually very sparse.

H.4 MORE VISUALIZATION OF AOGS

This section provides the visualization of more AOGs generated by our method on various datasets.

For tabular datasets, Figure 15, Figure 16, Figure 17, Figure 18, and Figure 19 show examples of AOGs generated by our method on different models trained on the census, bike, and TV news dataset. The up-arrow(\uparrow) / down-arrow(\downarrow) labeled in terminal nodes indicated the actual value of the input variable was greater than / less than the baseline value.

For NLP datasets, Figure 20 and Figure 21 show examples of AOGs generated by our method on LSTMs and CNNs trained on the SST-2 dataset and the CoLA dataset. Furthermore, Figure 22 shows examples of AOGs for explaining incorrect predictions. Results show that the AOG explainer could reveal reasons why the model made incorrect predictions. For example, in the sentiment classification task, the local sentiment may significantly affect the inference on the entire sentence, such as words "originality" and "cleverness" in Figure 22(top), words "originality" and "delight" in Figure 22(middle), and words "painfully" and "bad" in Figure 22(bottom).

H.5 DETAILS OF EXPERIMENTS ON SYNTHESIZED FUNCTIONS AND DATASETS

This section provides more details of synthesized functions and datasets used in Section 4.1 of the main paper.

The Addition-Multiplication dataset (Zhang et al., 2021). This dataset contained 100 functions, which only consisted of addition and multiplication operations. For example, $v(x) = x_1 + x_2x_3 + x_3x_4x_5 + x_4x_6$. Each variable x_i was a binary variable, *i.e.* $x_i \in \{0, 1\}$.

The ground-truth causal patterns and there effects corresponding to these functions can be easily determined. For each term in these functions (*e.g.* the term $x_3x_4x_5$ in the function $v(x) = x_1 + x_2x_3 + x_3x_4x_5$)



(a) (left) The average histogram of relative strength of causal effects encoded by the MLP-5 network, for the *census* dataset. (right) Examples of histograms of relative strength of causal effects, for single samples in the *census* dataset.



(b) (left) The average histogram of relative strength of causal patterns encoded by the ResMLP-5 network, for the *TV news* dataset. (right) Examples of histograms of relative strength of causal effects, for single samples in the *TV news* dataset.



(c) (left) The average histogram of relative strength of causal effects encoded by the CNN network, for the *SST-2* dataset. (right) Examples of histograms of relative strength of causal effects, for single samples in the *SST-2* dataset.

Figure 8: Histograms of the relative strength of causal effects. The causal effects w_S extracted from the deep model are usually very sparse.

 $x_3x_4x_5 + x_4x_6$), only when variables contained by this term were all present (*e.g.* $x_3 = x_4 = x_5 = 1$), this term would contribute to the output. Therefore, we could consider input variables in each term formed a ground-truth causal pattern. In the above example function, given the input x = [1, 1, 1, 1, 1, 1], the ground-truth causal patterns were $\Omega^{\text{truth}} = \{\{x_1\}, \{x_2, x_3\}, \{x_3, x_4, x_5\}, \{x_4, x_6\}\}$. Given the input x = [1, 1, 0, 1, 1, 1], the ground-truth causal patterns were $\Omega^{\text{truth}} = \{\{x_1\}, \{x_2, x_3\}, \{x_4, x_5\}, \{x_4, x_6\}\}$.

In our experiments, we randomly generated 100 Addition-Multiplication functions. Each of them had 10 input variables, and had 10 to 100 terms. Then, we randomly generated 200 binary input samples for each of these functions. For each input sample, let $m = |\Omega^{\text{truth}}|$ denote the number of the labeled ground-truth patterns. For fair comparison, we computed causal effects I(S) and extracted the top-m salient patterns $\Omega^{\text{top-}m}$. Then, we averaged the value of $\text{IoU} = \frac{|\Omega^{\text{truth}}|}{|\Omega^{\text{top-}k} - \Omega^{\text{truth}}|}$ over all samples.

The dataset in (Ren et al., 2021). This dataset contained 100 functions, which consisted of addition, subtraction, multiplication, and the sigmoid operations. Just like the Addition-Multiplication dataset, the ground-truth causal patterns in this dataset could also be easily determined. Let us consider the function $v(x) = -x_1x_2x_3 - \text{sigmoid}(5x_4x_5 - 5x_6 - 2.5), x_i \in \{0, 1\}$ as an example. The term $x_1x_2x_3$ was activated (= 1) if and only if $x_1 = x_2 = x_3 = 1$. The term $\text{sigmoid}(5x_4x_5 - 5x_6 - 2.5)$ was activated (> 0.5) if and only if $x_4 = x_5 = 1$ and $x_6 = 0$. Thus, we could also consider this function contained two ground-truth causal pattern. In other words, for the above function, given the input x = [1, 1, 1, 1, 1, 0], the ground-truth causal patterns were $\Omega^{\text{truth}} = \{\{x_1, x_2, x_3\}, \{x_4, x_5, x_6\}\}$. Given the input x = [1, 1, 1, 1, 1, 1], the ground-truth causal patterns were $\Omega^{\text{truth}} = \{\{x_1, x_2, x_3\}, \{x_4, x_5, x_6\}\}$.

In our experiments, we followed (Ren et al., 2021) to randomly generated 100 functions. Each of them had 6 to 12 input variables. Then, we randomly generated 200 binary input samples for each of these functions. Just like the Addition-Multiplication dataset, we extracted the top-m ($m = |\Omega^{truth}|$)

salient patterns $\Omega^{\text{top-}m}$, and computed the average IoU between Ω^{truth} and $\Omega^{\text{top-}m}$ over all samples for comparison.

The manually labeled And-Or dataset. This dataset contained 10 functions with AND operations (denoted by &) and OR operations (denoted by ||). For example, the function $f(x) = (x_1 > 0)\&(x_2 > 0)|(x_2 > 0)\&(x_3 > 0)\&(x_4 > 0)|(x_3 > 0)\&(x_5 > 0)$. Each input variable is a scalar, *i.e.* $x_i \in \mathbb{R}$, and the output is binary, *i.e.* $f(x) \in \{0, 1\}$. For each And-Or function, we randomly generated 100,000 Gaussian noises with n = 8 variables as input samples, and labeled these samples following functions in the And-Or dataset, namely the *manually labeled And-Or dataset*.

The ground-truth causal patterns in this dataset could be determined as follows. For the above function, we could consider $\{x_1, x_2\}, \{x_2, x_3, x_4\}$, and $\{x_3, x_5\}$ as possible causal patterns. If any of these patterns was significantly activated, *i.e.* if all input variables in this pattern were greater than a threshold $\tau = 0.5$, then we consider this pattern is significant enough to be a valid ground-truth causal patterns. *I.e.* for the above function, given the input $\boldsymbol{x} = [1.0, 2.0, 1.5, 0.9, 0.8]$, the ground-truth causal patterns were $\Omega^{\text{truth}} = \{\{x_1, x_2\}, \{x_2, x_3, x_4\}, \{x_3, x_5\}\}$. Given the input $\boldsymbol{x} = [0.8, 1.5, 1.2, 0.1, 0.9]$, the ground-truth causal patterns were $\Omega^{\text{truth}} = \{\{x_1, x_2\}, \{x_2, x_3, x_4\}, \{x_3, x_5\}\}$.

In our experiments, we trained one MLP-5 and one ResMLP-5 networks for binary classification based on the manually labeled dataset generated based on each And-Or function. Just like the above experiments, for each well-trained model, we extracted the top-*m* salient patterns and computed the average IoU over 1000 training samples for comparison. Note that there was no principle to ensure that the model learned exactly the ground-truth causality between input variables for inference. Therefore, the average IoU on this dataset was less than 1.

An extended version of the Addition-Multiplication dataset. In order to evaluate the accuracy of the computed causal effects, we also extended the Addition-Multiplication dataset to generate functions with not only ground-truth causal patterns, but also ground-truth causal effects for evaluation. The extended Addition-Multiplication dataset also contained 100 functions, which consisted of addition and multiplication operations. Each variable x_i was a binary variable, *i.e.* $x_i \in \{0, 1\}$. Different from functions in the Addition-Multiplication dataset, there were different coefficients before each term in each function. For example, $v(\mathbf{x}) = 3x_1 - 2x_2x_3 - x_3x_4x_5 + 5x_4x_6$.

The ground-truth causal effects in these functions can be easily determined. Just like the original Addition-Multiplication dataset, each term was a ground-truth pattern. In this case, we could consider the causal effect of each pattern as the value of its coefficient. For the above function, given the input $\boldsymbol{x} = [1, 1, 1, 1, 1, 1]$, the ground-truth effects of causal patterns were $w_{\{x_1\}} = 3, w_{\{x_2, x_3\}} = -2, w_{\{x_3, x_4, x_5\}} = -1, w_{\{x_4, x_6\}} = 5$, and for other $S \subseteq \{x_1, ..., x_6\}, w_S = 0$. Given the input $\boldsymbol{x} = (1, 1, 0, 1, 1, 1)$, the ground-truth causal effects were $w_{\{x_1\}} = 3, w_{\{x_4, x_6\}} = 5$, and for other $S \subseteq \{x_1, ..., x_6\}, w_S = 0$.

In our experiments, we randomly generated 100 functions. Each of them had 10 input variables, and had 10 to 100 terms. Then, we randomly generated 200 binary input samples for each of these functions. For each input sample, we measured the Jaccard similarity coefficient $J = \frac{\sum_{S \subseteq N} \min(|w_S^{\text{ruth}}|,|w_S|)}{\sum_{S \subseteq N} \max(|w_S^{\text{ruth}}|,|w_S|)}$ between ground-truth causal effects w_S^{truth} (defined above) and causal effects w_S computed by our method. The average value of J over all samples was 1.00, indicating that our method based on Harsanyi dividends correctly extracted the causal effects in these functions.

H.6 MORE ANALYSIS ON THE FAITHFULNESS OF THE AOG EXPLAINER

In this section, we provide more discussions about the experiment in Section 4.1, where we evaluated whether an explanation method faithfully extracted causal effects encoded by deep models based on metric 2. To this end, we considered the SI value $I^{\text{Shapley}}(S)$ (Grabisch & Roubens, 1999) and the STI value $I^{\text{Shapley-Taylor}}(S)$ (Sundararajan et al., 2020) as numerical effects of different interactive patterns S on a DNN's inference. Besides, we could also consider that attribution-based explanations quantified the causal effect of each single variable *i* (*e.g.* the Shapley-Taylor interaction index, the Shapley value (Shapley, 1953), Input×Gradient (Shrikumar et al., 2016), LRP (Bach et al., 2015), Occlusion (Zeiler & Fergus, 2014)).

Specifically, the computation of the metric ρ^{unfaith} for each baseline method are discussed as follows.

Explanation methods		TV news		ce	ensus	bike	
		MLP-5	ResMLP-5	MLP-5	ResMLP-5	MLP-5	ResMLP-5
Input v Gradiant	original	738.68	2586.14	408.10	1325.09	1.4E+5	1.1E+5
input × Gradient	normalized	1892.92	6898.62	520.69	1966.58	5.9E+4	4.1E+4
LRP	original	317.56	9.4E+4	155.13	1.4E+4	1.4E+5	5.8E+8
	normalized	2542.52	1358.84	1.7E+04	219.79	5.9E+4	4.3E+4
Occlusion	original	1386.16	1117.46	638.75	287.39	6.2E+4	3.7E+4
	normalized	330.58	323.21	154.61	139.08	1.4E+6	2.2E+5
Ours		9.4E-12	1.1E-11	8.5E-12	8.5E-12	2.6E-9	1.9E-9

Table 6: Unfaithfulness (\downarrow) of attribution-based explanations when we used the normalized and original attributions, respectively.

Table 7: Unfaithfulness (\downarrow) of distillation-based explainer models. Our AOG exhibited the lowest unfaithfulness.

Explanation methods		TV	' news	Ce	ensus	bike	
		MLP-5	ResMLP-5	MLP-5	ResMLP-5	MLP-5	ResMLP-5
distillation-based explainer models	distill	115.00	139.51	62.19	78.75	8374.04	4773.01
	SDT	155.89	160.77	168.87	158.73	-	-
	GBT	188.58	168.32	87.84	90.73	10724.51	7786.60
ours		9.4E-12	1.1E-11	8.5E-12	8.5E-12	2.6E-9	1.9E-9

• For *interaction-based explanations*, given an input sample x, let $I^{\text{Shapley-Taylor}}(S)$, $I^{\text{Shapley-Taylor}}(S)$ denote the Shapley interaction (SI) value and the Shapley-Taylor interaction (STI) value of the interactive pattern S. According to the SCM in Eq. (2), the metric ρ^{unfaith} is defined as follows.

$$\rho_{\mathrm{SI}}^{\mathrm{unfaith}} = \mathbb{E}_{\mathcal{S} \subseteq \mathcal{N}} [v(\boldsymbol{x}_{\mathcal{S}}) - \sum_{\mathcal{S}' \subseteq \mathcal{S}} I^{\mathrm{Shapley}}(\mathcal{S}')]^2, \quad \rho_{\mathrm{STI}}^{\mathrm{unfaith}} = \mathbb{E}_{\mathcal{S} \subseteq \mathcal{N}} [v(\boldsymbol{x}_{\mathcal{S}}) - \sum_{\mathcal{S}' \subseteq \mathcal{S}} I^{\mathrm{Shapley-Taylor}}(\mathcal{S}')]^2$$
(12)

• For *attribution-based explainer models*, given the input sample x, let $\phi_{\text{Shapley}}(i)$, $\phi_{\text{IG}}(i)$, $\phi_{\text{LRP}}(i)$, $\phi_{\text{Occ}}(i)$ denote the attribution of the input variable *i* computed by the Shapley value, Input × Gradient, LRP, and Occlusion, respectively. As mentioned above, these attribution values quantify the causal effect of each single variable *i*. According to the SCM in Eq. (2), the unfaithfulness of these attribution-based explanations was similarly measured as follows.

$$\rho_{\text{Shapley}}^{\text{unfaith}} = \mathbb{E}_{\mathcal{S} \subseteq \mathcal{N}} [v(\boldsymbol{x}_{\mathcal{S}}) - \sum_{i \in \mathcal{S}} \phi^{\text{Shapley}}(i)]^2, \quad \rho_{\text{IG}}^{\text{unfaith}} = \mathbb{E}_{\mathcal{S} \subseteq \mathcal{N}} [v(\boldsymbol{x}_{\mathcal{S}}) - \sum_{i \in \mathcal{S}} \phi^{\text{IG}}(i)]^2, \\ \rho_{\text{LRP}}^{\text{unfaith}} = \mathbb{E}_{\mathcal{S} \subseteq \mathcal{N}} [v(\boldsymbol{x}_{\mathcal{S}}) - \sum_{i \in \mathcal{S}} \phi^{\text{LRP}}(i)]^2, \quad \rho_{\text{Occ}}^{\text{unfaith}} = \mathbb{E}_{\mathcal{S} \subseteq \mathcal{N}} [v(\boldsymbol{x}_{\mathcal{S}}) - \sum_{i \in \mathcal{S}} \phi^{\text{Occ}}(i)]^2$$
(13)

Then, we compared the unfaithfulness of the AOG explainer with the above six baseline explanation methods. Based on each tabular dataset, we computed the average ρ^{unfaith} over the training samples, *i.e.* $\mathbb{E}_{\boldsymbol{x}}[\rho^{\text{unfaith}}]_{\text{given }\boldsymbol{x}}$. Table 2 in the main paper shows that the AOG explainer exhibited significantly stronger faithfulness than other explanation methods.

Besides, for fair comparison, we also used normalized attribution values to quantify the causal effects of each single variable. *I.e.* $\tilde{\phi}_{IG}(i) = \frac{\phi_{IG}(i)}{\sum_{i \in N} \phi_{IG}(i)} \cdot v(\boldsymbol{x}), \quad \tilde{\phi}_{LRP}(i) = \frac{\phi_{LRP}(i)}{\sum_{i \in N} \phi_{LRP}(i)} \cdot v(\boldsymbol{x}), \text{ and } \tilde{\phi}_{Occ}(i) = \frac{\phi_{Occ}(i)}{\sum_{i \in N} \phi_{Occ}(i)} \cdot v(\boldsymbol{x}).$ Then, we computed the average ρ^{unfaith} over the training samples based on these normalized causal effects. Table 6 shows that the normalized attributions usually exhibited stronger faithfulness, compared with their original attributions. Nevertheless, the AOG explainer still exhibited much stronger faithfulness compared with the normalized attributions.

We also conducted experiments to verify that distilling knowledge from a deep model to an explainer model cannot faithfully explain the logic in the deep model, as mentioned in Section 2. To this end, we compared the faithfulness of the AOG explainer model with distillation-based explainer models, including GBT (Che et al., 2016), SDT (Frosst & Hinton, 2017), and knowledge distillation (Hinton et al., 2015). Similarly, a distillation-based explainer model is faithful if it can mimic the deep model's output on massive masked samples x_S . Accordingly, let $g(x_S)$ denote the output of the explainer model when given the masked input x_S . The unfaithfulness of distillation-based explainer models was defined as $\rho^{\text{unfaith}} = \mathbb{E}_{S \subseteq \mathcal{N}} [v(x_S) - g(x_S)]^2$. Table 7 shows that the AOG explainer exhibited stronger faithfulness than distillation-based explanation methods, indicating that distilling knowledge from a deep model to an explainer model cannot faithfully explain the logic in the deep model.

Besides, we also conducted another experiment to verify the above claim. Specifically, as Figure 9 shows, although knowledge distillation could ensure that the student model had similar outputs with



Figure 9: Knowledge distillation cannot ensure the faithfulness of the student model, because the student model and the teacher model use different image regions to compute features. The first row shows the Grad-CAM attention (Selvaraju et al., 2017) of the teacher model. The second row shows the Grad-CAM attention of the student model.



Figure 10: The relationship between R_{Ω} and the unfaithfulness ρ^{unfaith} of the AOG explainer.

the teacher model, the student model exhibited a different logic from the teacher model. We distilled the output before the softmax of a pre-trained VGG-11 (Simonyan & Zisserman, 2014) into another VGG-11 on the CIFAR-10 dataset (Krizhevsky et al., 2009). The result showed that the attention of the student model computed by Grad-CAM (Selvaraju et al., 2017) was significantly different from that of the teacher model. Therefore, knowledge distillation could not ensure the student model represented the same inference logic as the teacher model, making the student model an unfaithful explainer for the teacher model.

H.7 THE RELATIONSHIP BETWEEN RATIO OF THE EXPLAINED CAUSAL EFFECTS AND THE FAITHFULNESS

We further studied the relationship between the ratio of explained causal effects R_{Ω} in Eq. (8) of the main paper and the unfaithfulness ρ^{unfaith} of the explanation. We averaged the R_{Ω} - ρ^{unfaith} curve over the training samples in each dataset. Figure 10 shows that the faithfulness of the AOG explainer increased along with the increase of ratio of the explained causal effects R_{Ω} .

H.8 More experimental results on the ratio of the explained causal effects R_{Ω}

This section provides more experimental results on the relationship between the ratio of explained causal effects R_{Ω} and the AOG explainer.

Just like the experiment in Paragraph *Ratio of the explained causal effects*, Section 4.2 of the main paper, we used causal patterns in Ω to approximate the model output. Figure 11(a) and Figure 12(a) show the relationship between $|\Omega|$ and the ratio of explained causal effects R_{Ω} in different models, based on the census dataset and the bike dataset. We found that when we used a few causal patterns, we could explain most causal effects in the model output. Figure 11(b,c) and Figure 12(b,c) show that the node number and the edge number increased along with the increase of R_{Ω} .

Besides, Figure 11(a) and Figure 12 also show that compared with the normally trained model, we could use less causal patterns (smaller $|\Omega|$) to achieve the same ratio of the explained causal effects R_{Ω} in the adversarially trained model. Moreover, Figure 11(b,c) and Figure 12(b,c) also show that AOGs corresponding to adversarially trained models were less complex than AOGs corresponding to normally trained models. This indicated that *adversarial training made models encode more sparse causal patterns than normal training*.



Figure 11: (a) The relationship between the number of causal patterns $|\Omega|$ in the AOG and the ratio of the explained causal effects R_{Ω} , based on the census dataset. The relationship between R_{Ω} and (b) the number of nodes, and (c) the number of edges in the AOG, based on the census dataset.



Figure 12: (a) The relationship between the number of causal patterns $|\Omega|$ in the AOG and the ratio of the explained causal effects R_{Ω} , based on the bike dataset. The relationship between R_{Ω} and (b) the number of nodes, and (c) the number of edges in the AOG, based on the bike dataset.

H.9 ANOTHER METRIC FOR THE RATIO OF THE EXPLAINED CAUSAL EFFECTS

In this section, we provide another metric for the ratio of the explained causal effects. As is discussed in Section 3.2 of the main paper, we only used causal patterns in Ω , to approximately explain the output of the deep model. The ratio of the explained causal effects could also be quantified as follows.

$$Q_{\Omega} = \frac{\sum_{\mathcal{S} \in \Omega} |w_{\mathcal{S}}|}{\sum_{\mathcal{S} \subseteq \mathcal{N}} |w_{\mathcal{S}}|},\tag{14}$$

where we used $\sum_{S \in \Omega} |w_S|$ to quantify the explained causal effects in the AOG explainer, while $\sum_{S \subseteq N} |w_S|$ represented all causal effects encoded by the deep model.

We also studied the relationship between the number of causal patterns $|\Omega|$ and Q_{Ω} , the relationship between Q_{Ω} and the unfaithfulness ρ^{unfaith} of the AOG explainer, and the relationship between Q_{Ω} and the AOG complexity. Figure 13 show that the ratio of explained causal effects Q_{Ω} increased along with the increase of the number of causal patterns $|\Omega|$. Besides, the faithfulness of the AOG explainer was boosted along with the increase of Q_{Ω} . Moreover, the number of nodes and the number of edges in the AOG also increased along with the increase of Q_{Ω} .

H.10 MORE ANALYSIS ON THE EFFECTIVENESS OF THE LEARNED BASELINE VALUES

This section provides more experimental analysis on the effects of baseline values on the conciseness of explanations. Beyond experiments in the Paragraph *Effects of baseline values on the conciseness of explanations*, Section 4.2 of the main paper, in this section, we analyzed the effectiveness of the learned baseline values in terms of the AOG complexity from different perspectives. To this end, we first computed causal effects using baseline values obtained in different epochs during the learning phase. Then, based on the computed causal effects, we measured the number of causal patterns, the node number, and the edge number in the AOG at each learning epoch. For fair comparison, we selected the minimum number $|\Omega|$ of causal patterns such that the ratio of the explained causal effects Q_{Ω} exceeded 70%, to construct the AOG. Figure 14 shows the change of the AOG complexity during the learning process of baseline values, in terms of the number of causal patterns, the number of nodes, and the number of edges in the AOG. We found that the learning of baseline values significantly simplified the AOG, thus boosting the conciseness of explanations.



Figure 13: (1) The first column shows the relationship between the number of causal patterns $|\Omega|$ in the AOG and the ratio of the explained causal effects Q_{Ω} , based on different datasets. (2) The second column shows the relationship between Q_{Ω} and the unfaithfulness ρ^{unfaith} of the AOG explainer, based on different datasets. (3) The third column and the fourth column show the relationship between Q_{Ω} and the node number, and the edge number in the AOG, respectively.



Figure 14: The number of patterns (the first column), nodes (the second column), and edges (the third column) in the AOG, based on baseline values of different learning epochs. The learned baseline value significantly enhanced the conciseness of explanations.

I DISCUSSION ABOUT THE DIFFERENCE BETWEEN THE AOG EXPLAINER AND THE BOW MODEL

Do we explain a DNN as a linear model, such as a bag-of-words (BoW) model (Sivic & Zisserman, 2003; Csurka et al., 2004)? First, although the AOG explainer seems like a linear additive model, the AOG explainer does NOT simplify the non-linear deep model as a linear model. Instead, as mentioned in Section 3.1, the AOG explainer extracts different causal patterns from different input samples, rather than using the same set of causal patterns to explain different samples. It is because the deep model is non-linear and triggers different causal patterns to handle different samples. Therefore, unlike the BoW model that extracts the same set of features for each sample, the AOG explainer quantifies how the deep model triggers different causal patterns to handle different samples, thereby still being non-linear for different inputs. Second, the BoW model only considers the presence or absence of input variables, while the AOG explainer is sensitive to the spatial relationship of input variables. For example, Table 8 shows the causal effects w_S of the same sets of words S encoded by the deep model², given two sentences with the same words but different word positions. We found

²In this example, we explained the causal effects encoded by a two-layer LSTM model trained on the SST-2 dataset for sentiment classification. We set $v(\boldsymbol{x}_{S}) = p(y = \text{positive sentiment}|\boldsymbol{x}_{S})$.

that the deep model encoded significantly different causal effects between the same sets of words, which demonstrated that the AOG explainer was different from the BoW model.

Table 8: Given two sentences with the same words but different word positions, the causal effects of the same sets of words S encoded by the deep model were different. This demonstrated that the AOG explainer was sensitive to the spatial relationship of input variables, indicating a difference with the BoW model.

Sentence 1: it's just	not very smart.	Sentence 2: it's not just very smart.				
sets of words S	causal effects w_S	sets of words S	causal effects w_S			
{just, not, smart, .}	-1.616	{not, just, smart, .}	1.139			
{ <i>it, just, not, very</i> }	-1.510	{ <i>it, not, just, very</i> }	5.908			
{'s, just, not, very, smart}	-1.172	{'s, not, just, very, smart}	0.890			
{just, not, very, smart}	-0.715	{not, just, very, smart}	3.563			

Nevertheless, common and salient causal patterns shared by different input samples can also be considered as basic elementary concepts encoded by the deep model. For example, if two sentences contain the same set of words S in the same positions, then the deep model will encode the same causal effects $w_{S'}, \forall S' \subseteq S$. Table 9 shows that the deep model encoded the same causal effects within $S = \{not, very, smart\}$ in two different sentences. From this perspective, such common causal patterns can be roughly considered as typical "words" in a BoW model.

Table 9: Given two sentences containing the same set of words $S = \{not, very, smart\}$, the causal effects within the subset of words S encoded by the deep model were the same. The deep model encoded the same causal effects $w_{S'}, \forall S' \subseteq S$.

Sentence 1: it's just	t not very smart.	Sentence 3: he is just not very smart.			
sets of words $\mathcal{S}' \subseteq \mathcal{S}$	causal effect $w_{S'}$	sets of words $\mathcal{S}' \subseteq \mathcal{S}$	causal effect $w_{S'}$		
{not, smart}	-13.481	{not, smart}	-13.481		
$\{not, very\}$	-12.826	$\{not, very\}$	-12.826		
{smart}	6.568	{smart}	6.568		
{very, smart}	3.720	{very, smart}	3.720		
$\{not\}$	0.939	$\{not\}$	0.939		
{not, very, smart}	0.837	{not, very, smart}	0.837		
{very}	-0.197	{very}	-0.197		



Figure 15: An example of the AOG extracted from the MLP-5 network, trained on the census dataset. Red edges indicate the parse graph of the most salient causal pattern.



Figure 16: An example of the AOG extracted from the ResMLP-5 network, trained on the census dataset. Red edges indicate the parse graph of the most salient causal pattern.







(b) Examples of AOGs extracted from the MLP-5 network, adversarially trained on the census dataset.



(c) Examples of AOGs extracted from the ResMLP-5 network, adversarially trained on the census dataset.

Figure 17: Examples of AOGs extracted from models trained on the census dataset. Red edges indicate the parse graph of a specific causal pattern.



(a) Examples of AOGs extracted from the MLP-5 network, adversarially trained on the bike dataset.



(b) Examples of AOGs extracted from the ResMLP-5 network, adversarially trained on the bike dataset.

Figure 18: Examples of AOGs extracted from models trained on the bike dataset. Red edges indicate the parse graph of a specific causal pattern.







(b) Examples of AOGs extracted from the MLP-5 network, adversarially trained on the TV news dataset.



(c) Examples of AOGs extracted from the ResMLP-5 network, adversarially trained on the TV news dataset.

Figure 19: Examples of AOGs extracted from models trained on the TV news dataset. Red edges indicate the parse graph of a specific causal pattern.





(b) Examples of AOGs extracted from the LSTM network, trained on the SST-2 dataset.

Figure 20: Examples of AOGs extracted from models trained on the SST-2 dataset. Red edges indicate the parse graph of the most salient causal pattern.



(a) Examples of AOGs extracted from the CNN network, trained on the CoLA dataset.



(b) Examples of AOGs extracted from the LSTM network, trained on the CoLA dataset.

Figure 21: Examples of AOGs extracted from models trained on the CoLA dataset. Red edges indicate the parse graph of the most salient causal pattern.



label: negative sentiment | prediction: positive sentiment | $R_{\Omega}=99.31\%$

Figure 22: AOGs that explained incorrect predictions of the network model trained on the SST-2 dataset. Red edges indicated the parse graphs of causal patterns towards correct predictions, while blue edges indicated parse graphs of causal patterns towards wrong predictions.