Defining and Quantifying the Emergence of Sparse Concepts in DNNs: Supplementary Material

Jie Ren*, Mingjie Li*, Qirui Chen, Huiqi Deng, Quanshi Zhang†
Shanghai Jiao Tong University

A. Related works

Explanations for DNNs. Many methods have been proposed to explain DNNs, such as visualizing the features learned by the DNN [16, 46, 62, 63], and estimating the pixel-wise attribution/saliency of input samples [2, 19, 34, 41, 42, 66, 67]. [9] and [61] estimated the smallest subset of variables to mimic DNN's output. Some studies extracted logical rules as explanations [6, 27, 28, 36, 57]. Meanwhile, another direction is to distill a DNN into another interpretable symbolic model, for example, an additive model [53, 55], decision tree [1, 8, 20, 59], or graphical model [44, 65]. *However, most of these explainer models usually only consider the model's fitness to the network output, but whether their explanation can always faithfully reflect the logic in the DNN under various data transformations is still an open problem.* In this study, we find that the network outputs on an exponential number of randomly masked samples can always be explained by a causal graph, of which the faithfulness is theoretically proven.

Using causality to explain DNNs. The causality framework was originally proposed to study the causal structure of a set of observed variables [26, 38]. For example, [60] proposed a neural-causal model to identify and estimate causal relationships in data. Recently, several studies have explained DNNs based on causality. For example, some studies [21, 25, 56] proposed attribution methods based on manually defined causal relationships between input variables. Similarly, [3, 7, 23] explained the association between inputs and intermediate features/outputs using causal models. Instead of manually setting or assuming causal relationships, we quantify the exact interactive concepts encoded by the DNN as causal patterns for inference, whose faithfulness is both theoretically guaranteed and experimentally verified. Note that the SCM in Eq. (2) of the main paper does not explain the DNN as a linear model, such as a bag-of-words model [13, 48]. This is because given different 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 interaction in game theory. Similar to causal effects, interactions in game theory are widely used to quantify the numerical effects of interactive concepts between input variables on the DNN output [29, 30, 37, 47, 50]. In game theory, the Shapley interaction index [22] was used by [33] to analyze tree ensembles. [51, 54] proposed interaction metrics from different perspectives. [15] proved that DNNs were less likely to encode interactive concepts of intermediate complexity. Unlike previous studies, we find that we can use a few causal patterns (interactive concepts) to faithfully represent the inference logic of a DNN, which is experimentally verified.

B. Harsanyi dividend

This section revisits the definition of Harsanyi dividend [24], a typical metric in game theory. In this study, the causal effect $w_{\mathcal{S}}$ of each pattern $\mathcal{S}$ is quantified based on Harsanyi dividends. In game theory, a complex system (*e.g.,* a deep model) is usually considered a game. Each input variable represents a player in the game, and the output of this system is the reward obtained by a subset of players. Specifically, let us consider a deep model and an input sample $\boldsymbol{x}$ with n variables (*e.g.* a sentence with n words) $\mathcal{N} = \{1, 2, ..., n\}$. A deep model can be understood as a game $v(\cdot)$. In this game, the input variables in $\mathcal{N}$ do not individually contribute to the model output. Instead, they interact with each other to form concepts (causal patterns) for inference. Each concept $\mathcal{S} \subseteq \mathcal{N}$ has a certain causal effect on the model output. In this study, we prove in Theorem 1 that the Harsanyi dividend $w_{\mathcal{S}}$ is a unique faithful metric for quantifying such causal effects.

*These authors contributed equally to this work.
†Quanshi Zhang is the corresponding author. He is with the Department of Computer Science and Engineering, the John Hopcroft Center, at the Shanghai Jiao Tong University, China. zqs1022@sjtu.edu.cn.

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

where $v(\boldsymbol{x}_{\mathcal{S}})$ denotes the model output when only variables in the subset $\mathcal{S} \subseteq \mathcal{N}$ are given, and all other variables are masked using their baseline values.

We also prove that the Harsanyi dividend $w_{\mathcal{S}}$ satisfies seven desirable axioms, including the *efficiency, linearity, dummy, symmetry, anonymity, recursive* and *interaction distribution* axioms, which demonstrates its trustworthiness.

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

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

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

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

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

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

(7) *Interaction distribution axiom*. This axiom characterizes how causal effects are distributed for a class of "interaction functions" [51]. The interaction function $v_{\mathcal{T}}$ parameterized by a subset of variables $\mathcal{T}$ is defined as follows. $\forall \mathcal{S} \subseteq \mathcal{N}$, if $\mathcal{T} \subseteq \mathcal{S}$, $v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}}) = c$; otherwise, $v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}}) = 0$. The function $v_{\mathcal{T}}$ models the causal effect of the pattern $\mathcal{T}$, because only if all variables in $\mathcal{T}$ are present, will the output value be increased by c. The causal effects encoded in the function $v_{\mathcal{T}}$ satisfy $w_{\mathcal{T}} = c$, and $\forall \mathcal{S} \neq \mathcal{T}, w_{\mathcal{S}} = 0$.

More crucially, we also prove that causal effects $w_{\mathcal{S}}$ 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 [22]). *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 the causal effects inside $\mathcal{T}$ and the sub-environments of $\mathcal{S}$,* i.e. $\Delta v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}}) = \sum_{\mathcal{S}' \subseteq \mathcal{S}} w_{\mathcal{T}\cup\mathcal{S}'}$.

Theorem 2 (Connection to the Shapley value [43]). *Let $\phi(i)$ denote the Shapley value of input variable i. Then, the Shapley value $\phi(i)$ can be explained as the result of uniformly assigning causal effects to each involved variable i,* i.e., $\phi(i) = \sum_{\mathcal{S} \subseteq \mathcal{N}\backslash\{i\}} \frac{1}{|\mathcal{S}|+1} w_{\mathcal{S}\cup\{i\}}$. This theorem also proves that the Shapley value is a fair assignment of attributions from the perspective of causal effects.

Theorem 3 (Connection to the Shapley interaction index [22]). *Given a subset of input variables $\mathcal{T} \subseteq \mathcal{N}$, the Shapley interaction index $I^{\text{Shapley}}(\mathcal{T})$ can be represented as $I^{\text{Shapley}}(\mathcal{T}) = \sum_{\mathcal{S} \subseteq \mathcal{N}\backslash\mathcal{T}} \frac{1}{|\mathcal{S}|+1} w_{\mathcal{S}\cup\mathcal{T}}$. 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.*

Theorem 4 (Connection to the Shapley Taylor interaction index [51]). *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 $|\mathcal{T}| < k$; $I^{\text{Shapley-Taylor}}(\mathcal{T}) = \sum_{\mathcal{S} \subseteq \mathcal{N}\backslash\mathcal{T}} \binom{|\mathcal{S}|+k}{k}^{-1} w_{\mathcal{S}\cup\mathcal{T}}$ if $|\mathcal{T}| = k$; and $I^{\text{Shapley-Taylor}}(\mathcal{T}) = 0$ if $|\mathcal{T}| > k$.*

C. The proof of Theorem 1 in the main paper

Theorem 1. *Given a certain input $\boldsymbol{x}$, let the causal graph in Fig. 1 (in the main paper) encode 2^n causal patterns,* i.e., *$\Omega = 2^{\mathcal{N}} = \{\mathcal{S} : \mathcal{S} \subseteq \mathcal{N}\}$. If the causal effect $w_{\mathcal{S}}$ of each causal pattern $\mathcal{S} \in \Omega$ is measured by the Harsanyi dividend [24],* i.e. *$w_{\mathcal{S}} \triangleq \sum_{\mathcal{S}' \subseteq \mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} \cdot v(\boldsymbol{x}_{\mathcal{S}'})$, then the causal graph faithfully encodes the inference logic of the DNN, as follows.*

$$\forall \mathcal{S} \subseteq \mathcal{N},\ \ Y(\boldsymbol{x}_{\mathcal{S}}) = v(\boldsymbol{x}_{\mathcal{S}}) \tag{2}$$

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 $w_{\mathcal{S}}$ satisfies the faithfulness requirement $\forall \mathcal{S} \subseteq \mathcal{N}, Y(\boldsymbol{x}_{\mathcal{S}})=v(\boldsymbol{x}_{\mathcal{S}})$. (2) Sufficiency: if there exists another metric $\tilde{w}_{\mathcal{S}}$ that also satisfies the faithfulness requirement, then, it is equivalent to the Harsanyi dividend, *i.e.* $\forall \mathcal{S} \subseteq \mathcal{N}, \tilde{w}_{\mathcal{S}} = w_{\mathcal{S}}$.

According to the SCM in Eq. (2) of the main paper, 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 $\forall \mathcal{S} \subseteq \mathcal{N}$,

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

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 $\tilde{w}_{\mathcal{S}} = w_{\mathcal{S}}$ holds for any $\mathcal{S}$ with $|\mathcal{S}| = s \geq 2$. Then, for $|\mathcal{S}| = s+1$, we have

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

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 $w_{\mathcal{S}}$ 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.* $v(\boldsymbol{x}) = \sum_{\mathcal{S}\subseteq\mathcal{N}} w_{\mathcal{S}}$.

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

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

(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.* $\forall \mathcal{S}\subseteq\mathcal{N},\ v(\boldsymbol{x}_{\mathcal{S}}) = t(\boldsymbol{x}_{\mathcal{S}}) + u(\boldsymbol{x}_{\mathcal{S}})$, then the corresponding causal effects $w^t_{\mathcal{S}}$ and $w^u_{\mathcal{S}}$ can also be merged as $\forall \mathcal{S}\subseteq\mathcal{N}, w^v_{\mathcal{S}} = w^t_{\mathcal{S}} + w^u_{\mathcal{S}}$.

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

$$
\begin{aligned}
w^v_{\mathcal{S}} &= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}}) \\
&= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} [t(\boldsymbol{x}_{\mathcal{S}}) + u(\boldsymbol{x}_{\mathcal{S}})] \\
&= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} t(\boldsymbol{x}_{\mathcal{S}}) + \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} u(\boldsymbol{x}_{\mathcal{S}}) \\
&= w^t_{\mathcal{S}} + w^u_{\mathcal{S}}.
\end{aligned}
$$

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

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

$$
\begin{aligned}
w_{\mathcal{S}\cup\{i\}} &= \sum_{\mathcal{S}'\subseteq\mathcal{S}\cup\{i\}} (-1)^{|\mathcal{S}|+1-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'}) \\
&= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|+1-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'}) + \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'\cup\{i\}}) \\
&= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|+1-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'}) + \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} [v(\boldsymbol{x}_{\mathcal{S}}) + v(\boldsymbol{x}_{\{i\}})] \\
&= \Big[\sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|}\Big] \cdot v(\boldsymbol{x}_{\{i\}}) \\
&= 0.
\end{aligned}
$$

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

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

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

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

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

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

$$
\begin{aligned}
w^{\pi v}_{\pi\mathcal{S}} &= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} (\pi v)(\boldsymbol{x}_{\pi\mathcal{S}'}) \\
&= \sum_{\mathcal{S}'\subseteq\mathcal{S}} (-1)^{|\mathcal{S}|-|\mathcal{S}'|} v(\boldsymbol{x}_{\mathcal{S}'}) \\
&= w^v_{\mathcal{S}}.
\end{aligned}
$$

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

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

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

(7) Interaction distribution axiom. This axiom characterizes how causal effects are distributed for a class of "interaction functions" [51]. An interaction function $v_{\mathcal{T}}$ parameterized by a subset of variables $\mathcal{T}$ is defined as follows. $\forall\mathcal{S} \subseteq \mathcal{N}$, if $\mathcal{T} \subseteq \mathcal{S}$, $v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}}) = c$; otherwise, $v_{\mathcal{T}}(\boldsymbol{x}_{\mathcal{S}}) = 0$. 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 $\forall\mathcal{S} \neq \mathcal{T}$, $w_{\mathcal{S}} = 0$.

• *Proof:* If $\mathcal{S} \subsetneq \mathcal{T}$, 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 $\mathcal{S} = \mathcal{T}$, we have

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

If $\mathcal{S} \supsetneq \mathcal{T}$, we have

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

$$=c\cdot\sum_{\substack{\mathcal{S}'\subseteq\mathcal{S}\\ \mathcal{S}'\supseteq\mathcal{T}}}(-1)^{|\mathcal{S}|-|\mathcal{S}'|}$$
$$=c\cdot\sum_{m=0}^{|\mathcal{S}|-|\mathcal{T}|}\binom{|\mathcal{S}|-|\mathcal{T}|}{m}(-1)^m=0.$$

D.2. Proofs of theorems

In this section, we prove connections between the Harsanyi dividend $w_\mathcal{S}$ 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.* $\Delta v_\mathcal{T}(\boldsymbol{x}_\mathcal{S})=\sum_{\mathcal{S}'\subseteq\mathcal{S}}w_{\mathcal{T}\cup\mathcal{S}'}$.

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

$$\begin{aligned}
\Delta v_\mathcal{T}(\boldsymbol{x}_\mathcal{S}) &= \sum_{\mathcal{L}\subseteq\mathcal{T}}(-1)^{|\mathcal{T}|-|\mathcal{L}|}v(\boldsymbol{x}_{\mathcal{L}\cup\mathcal{S}})\\
&= \sum_{\mathcal{L}\subseteq\mathcal{T}}(-1)^{|\mathcal{T}|-|\mathcal{L}|}\sum_{\mathcal{K}\subseteq\mathcal{L}\cup\mathcal{S}}w_\mathcal{K} \quad \text{// by Theorem 1}\\
&= \sum_{\mathcal{L}\subseteq\mathcal{T}}(-1)^{|\mathcal{T}|-|\mathcal{L}|}\sum_{\mathcal{L}'\subseteq\mathcal{L}}\sum_{\mathcal{S}'\subseteq\mathcal{S}}w_{\mathcal{L}'\cup\mathcal{S}'} \quad \text{// since } \mathcal{L}\cap\mathcal{S}=\emptyset\\
&= \sum_{\mathcal{S}'\subseteq\mathcal{S}}\left[\sum_{\mathcal{L}\subseteq\mathcal{T}}(-1)^{|\mathcal{T}|-|\mathcal{L}|}\sum_{\mathcal{L}'\subseteq\mathcal{L}}w_{\mathcal{L}'\cup\mathcal{S}'}\right]\\
&= \sum_{\mathcal{S}'\subseteq\mathcal{S}}\left[\sum_{\mathcal{L}'\subseteq\mathcal{T}}\sum_{\substack{\mathcal{L}\subseteq\mathcal{T}\\ \mathcal{L}\supseteq\mathcal{L}'}}(-1)^{|\mathcal{T}|-|\mathcal{L}|}w_{\mathcal{L}'\cup\mathcal{S}'}\right]\\
&= \sum_{\mathcal{S}'\subseteq\mathcal{S}}\left[\underbrace{w_{\mathcal{S}'\cup\mathcal{T}}}_{\mathcal{L}'=\mathcal{T}}+\underbrace{\sum_{\mathcal{L}'\subsetneq\mathcal{T}}\left(\sum_{l=|\mathcal{L}'|}^{|\mathcal{T}|}\binom{|\mathcal{T}|-|\mathcal{L}'|}{l-|\mathcal{L}'|}(-1)^{|\mathcal{T}|-|\mathcal{L}|}w_{\mathcal{L}'\cup\mathcal{S}'}\right)}_{L'\subsetneq\mathcal{T}}\right]\\
&= \sum_{\mathcal{S}'\subseteq\mathcal{S}}\left[w_{\mathcal{S}'\cup\mathcal{T}}+\sum_{\mathcal{L}'\subsetneq\mathcal{T}}\left(w_{\mathcal{L}'\cup\mathcal{S}'}\cdot\underbrace{\sum_{l=|\mathcal{L}'|}^{|\mathcal{T}|}\binom{|\mathcal{T}|-|\mathcal{L}'|}{l-|\mathcal{L}'|}(-1)^{|\mathcal{T}|-|\mathcal{L}|}}_{=0}\right)\right]\\
&= \sum_{\mathcal{S}'\subseteq\mathcal{S}}w_{\mathcal{S}'\cup\mathcal{T}} \quad \square
\end{aligned}$$

In particular, if $\mathcal{T}$ is a singleton set, *i.e.* $\mathcal{T}=\{i\}$, we can obtain a similar conclusion to [40] 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 [43] 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.*, $\phi(i)=\sum_{\mathcal{S}\subseteq\mathcal{N}\setminus\{i\}}\frac{1}{|\mathcal{S}|+1}w_{\mathcal{S}\cup\{i\}}$. *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

$$\begin{aligned}
\phi(i) &= \mathop{\mathbb{E}}_{m}\mathop{\mathbb{E}}_{\substack{\mathcal{S}\subseteq\mathcal{N}\setminus\{i\}\\ |\mathcal{S}|=m}}[v(\boldsymbol{x}_{\mathcal{S}\cup\{i\}})-v(\boldsymbol{x}_\mathcal{S})]\\
&= \frac{1}{|\mathcal{N}|}\sum_{m=0}^{|\mathcal{N}|-1}\frac{1}{\binom{|\mathcal{N}|-1}{m}}\sum_{\substack{\mathcal{S}\subseteq\mathcal{N}\setminus\{i\}\\ |\mathcal{S}|=m}}\Big[v(\boldsymbol{x}_{\mathcal{S}\cup\{i\}})-v(\boldsymbol{x}_\mathcal{S})\Big]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{|\mathcal{N}|} \sum_{m=0}^{|\mathcal{N}|-1} \frac{1}{\binom{|\mathcal{N}|-1}{m}} \sum_{\substack{\mathcal{S}\subseteq\mathcal{N}\setminus\{i\}\\|\mathcal{S}|=m}} \Delta v_{\{i\}}(\boldsymbol{x}_{\mathcal{S}}) \\
&= \frac{1}{|\mathcal{N}|} \sum_{m=0}^{|\mathcal{N}|-1} \frac{1}{\binom{|\mathcal{N}|-1}{m}} \sum_{\substack{\mathcal{S}\subseteq\mathcal{N}\setminus\{i\}\\|\mathcal{S}|=m}} \left[\sum_{\mathcal{L}\subseteq\mathcal{S}} w_{\mathcal{L}\cup\{i\}}\right] \quad \text{// by Theorem 5} \\
&= \frac{1}{|\mathcal{N}|} \sum_{\mathcal{L}\subseteq\mathcal{N}\setminus\{i\}} \sum_{m=0}^{|\mathcal{N}|-1} \frac{1}{\binom{|\mathcal{N}|-1}{m}} \sum_{\substack{\mathcal{S}\subseteq\mathcal{N}\setminus\{i\}\\|\mathcal{S}|=m\\\mathcal{S}\supseteq\mathcal{L}}} w_{\mathcal{L}\cup\{i\}} \\
&= \frac{1}{|\mathcal{N}|} \sum_{\mathcal{L}\subseteq\mathcal{N}\setminus\{i\}} \sum_{m=|\mathcal{L}|}^{|\mathcal{N}|-1} \frac{1}{\binom{|\mathcal{N}|-1}{m}} \sum_{\substack{\mathcal{S}\subseteq\mathcal{N}\setminus\{i\}\\|\mathcal{S}|=m\\\mathcal{S}\supseteq\mathcal{L}}} w_{\mathcal{L}\cup\{i\}} \quad \text{// since } \mathcal{S}\supseteq\mathcal{L},\ |\mathcal{S}|=m\geq|\mathcal{L}|. \\
&= \frac{1}{|\mathcal{N}|} \sum_{\mathcal{L}\subseteq\mathcal{N}\setminus\{i\}} \sum_{m=|\mathcal{L}|}^{|\mathcal{N}|-1} \frac{1}{\binom{|\mathcal{N}|-1}{m}} \cdot \binom{|\mathcal{N}|-|\mathcal{L}|-1}{m-|\mathcal{L}|} \cdot w_{\mathcal{L}\cup\{i\}} \\
&= \frac{1}{|\mathcal{N}|} \sum_{\mathcal{L}\subseteq\mathcal{N}\setminus\{i\}} w_{\mathcal{L}\cup\{i\}} \underbrace{\sum_{k=0}^{|\mathcal{N}|-|\mathcal{L}|-1} \frac{1}{\binom{|\mathcal{N}|-1}{|\mathcal{L}|+k}} \cdot \binom{|\mathcal{N}|-|\mathcal{L}|-1}{k}}_{\alpha_{\mathcal{L}}}
\end{aligned}
$$

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 $\binom{n}{m} = \frac{1}{m\cdot B(n-m+1,m)}$.

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

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

$$
\begin{aligned}
â‘ &= \int_0^1 |\mathcal{L}| \sum_{k=0}^{|\mathcal{N}|-|\mathcal{L}|-1} \binom{|\mathcal{N}|-|\mathcal{L}|-1}{k} \cdot x^{|\mathcal{N}|-|\mathcal{L}|-k-1} \cdot (1-x)^{|\mathcal{L}|+k-1}\, dx \\
&= \int_0^1 |\mathcal{L}| \cdot \underbrace{\left[\sum_{k=0}^{|\mathcal{N}|-|\mathcal{L}|-1} \binom{|\mathcal{N}|-|\mathcal{L}|-1}{k} \cdot x^{|\mathcal{N}|-|\mathcal{L}|-k-1} \cdot (1-x)^{k}\right]}_{=1} \cdot (1-x)^{|\mathcal{L}|-1}\, dx \\
&= \int_0^1 |\mathcal{L}|(1-x)^{|\mathcal{L}|-1}\, dx = 1
\end{aligned}
$$

For ②, we have

$$
\begin{aligned}
② = & \sum_{k=1}^{|\mathcal{N}|-|\mathcal{L}|-1} (|\mathcal{N}|-|\mathcal{L}|-1) \cdot \binom{|\mathcal{N}|-|\mathcal{L}|-2}{k-1} \cdot B(|\mathcal{N}|-|\mathcal{L}|-k, |\mathcal{L}|+k) \\
= & (|\mathcal{N}|-|\mathcal{L}|-1) \sum_{k'=0}^{|\mathcal{N}|-|\mathcal{L}|-2} \binom{|\mathcal{N}|-|\mathcal{L}|-2}{k'} \cdot B(|\mathcal{N}|-|\mathcal{L}|-k'-1, |\mathcal{L}|+k'+1) \\
= & (|\mathcal{N}|-|\mathcal{L}|-1) \int_0^1 \sum_{k'=0}^{|\mathcal{N}|-|\mathcal{L}|-2} \binom{|\mathcal{N}|-|\mathcal{L}|-2}{k'} \cdot x^{|\mathcal{N}|-|\mathcal{L}|-k'-2} \cdot (1-x)^{|\mathcal{L}|+k'} \, dx \\
= & (|\mathcal{N}|-|\mathcal{L}|-1) \int_0^1 \underbrace{\left[\sum_{k'=0}^{|\mathcal{N}|-|\mathcal{L}|-2} \binom{|\mathcal{N}|-|\mathcal{L}|-2}{k'} \cdot x^{|\mathcal{N}|-|\mathcal{L}|-k'-2} \cdot (1-x)^{k'} \right]}_{=1} \cdot (1-x)^{|\mathcal{L}|} \, dx \\
= & (|\mathcal{N}|-|\mathcal{L}|-1) \int_0^1 (1-x)^{|\mathcal{L}|} \, dx = \frac{|\mathcal{N}|-|\mathcal{L}|-1}{|\mathcal{L}|+1}
\end{aligned}
$$

Hence, we have

$$\alpha_{\mathcal{L}} = ① + ② = 1 + \frac{|\mathcal{N}|-|\mathcal{L}|-1}{|\mathcal{L}|+1} = \frac{|\mathcal{N}|}{|\mathcal{L}|+1}$$

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

Theorem 3 (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 [22] 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:*

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

Just like the proof of Theorem 2, we leverage the properties of combinitorial numbers and the Beta function to simplify α_L.

$$
\begin{aligned}
\alpha_{\mathcal{L}} &= \sum_{k=0}^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|} \frac{1}{\binom{|\mathcal{N}|-|\mathcal{T}|}{|\mathcal{L}|+k}} \binom{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|}{k} \\
&= \sum_{k=0}^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|} \binom{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|}{k} \cdot \Big(|\mathcal{L}|+k\Big) \cdot B\Big(|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-k+1, |\mathcal{L}|+k\Big) \\
&= \sum_{k=0}^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|} |\mathcal{L}| \cdot \binom{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|}{k} \cdot B\Big(|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-k+1, |\mathcal{L}|+k\Big) \qquad \cdots ①\\
&\quad + \sum_{k=0}^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|} k \cdot \binom{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|}{k} \cdot B\Big(|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-k+1, |\mathcal{L}|+k\Big) \qquad \cdots ②
\end{aligned}
$$

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

$$
\begin{aligned}
① &= \int_0^1 |\mathcal{L}| \sum_{k=0}^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|} \binom{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|}{k} \cdot x^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-k} \cdot (1-x)^{|\mathcal{L}|+k-1}\, dx \\
&= \int_0^1 |\mathcal{L}| \cdot \underbrace{\left[\sum_{k=0}^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|} \binom{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|}{k} \cdot x^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-k} \cdot (1-x)^{k} \right]}_{=1} \cdot (1-x)^{|\mathcal{L}|-1}\, dx \\
&= \int_0^1 |\mathcal{L}| \cdot (1-x)^{|\mathcal{L}|-1}\, dx = 1
\end{aligned}
$$

For ②, we have

$$
\begin{aligned}
② &= \sum_{k=1}^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|} (|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|) \binom{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-1}{k-1} \cdot B\Big(|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-k+1, |\mathcal{L}|+k\Big) \\
&= (|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|) \sum_{k'=0}^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-1} \binom{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-1}{k'} \cdot B\Big(|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-k', |\mathcal{L}|+k'+1\Big) \\
&= (|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|) \int_0^1 \sum_{k'=0}^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-1} \binom{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-1}{k'} \cdot x^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-k'-1} \cdot (1-x)^{|\mathcal{L}|+k'}\, dx \\
&= (|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|) \int_0^1 \underbrace{\left[\sum_{k'=0}^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-1} \binom{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-1}{k'} \cdot x^{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|-k'-1} \cdot (1-x)^{k'} \right]}_{=1} \cdot (1-x)^{|\mathcal{L}|}\, dx \\
&= (|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|) \int_0^1 (1-x)^{|\mathcal{L}|}\, dx = \frac{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|}{|\mathcal{L}|+1}
\end{aligned}
$$

Hence, we have

$$
\alpha_{\mathcal{L}} = ① + ② = 1 + \frac{|\mathcal{N}|-|\mathcal{L}|-|\mathcal{T}|}{|\mathcal{L}|+1} = \frac{|\mathcal{N}|-|\mathcal{T}|+1}{|\mathcal{L}|+1}
$$

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 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 $I^{\text{Shapley-Taylor}}(\mathcal{T})$ 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

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

Just like the proof of Theorem 2, we leverage the properties of combinatorial numbers and the Beta function to simplify α_L.

$$\begin{aligned}
\alpha_{\mathcal{L}} =& \sum_{m=0}^{|\mathcal{N}|-|\mathcal{L}|-k} \frac{1}{\binom{|\mathcal{N}|-1}{|\mathcal{L}|+m}} \binom{|\mathcal{N}|-|\mathcal{L}|-k}{m} \\
=& \sum_{m=0}^{|\mathcal{N}|-|\mathcal{L}|-k} \binom{|\mathcal{N}|-|\mathcal{L}|-k}{m} \cdot \Big(|\mathcal{L}|+m\Big) \cdot B\Big(|\mathcal{N}|-|\mathcal{L}|-m, |\mathcal{L}|+m\Big) \\
=& \sum_{m=0}^{|\mathcal{N}|-|\mathcal{L}|-k} |\mathcal{L}| \cdot \binom{|\mathcal{N}|-|\mathcal{L}|-k}{m} \cdot B\Big(|\mathcal{N}|-|\mathcal{L}|-m, |\mathcal{L}|+m\Big) \qquad \cdots ① \\
&+ \sum_{m=0}^{|\mathcal{N}|-|\mathcal{L}|-k} m \cdot \binom{|\mathcal{N}|-|\mathcal{L}|-k}{m} \cdot B\Big(|\mathcal{N}|-|\mathcal{L}|-m, |\mathcal{L}|+m\Big) \qquad \cdots ②
\end{aligned}$$

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

$$\begin{aligned}
① =& \int_0^1 |\mathcal{L}| \cdot \sum_{m=0}^{|\mathcal{N}|-|\mathcal{L}|-k} \binom{|\mathcal{N}|-|\mathcal{L}|-k}{m} \cdot x^{|\mathcal{N}|-|\mathcal{L}|-m-1} \cdot (1-x)^{|\mathcal{L}|+m-1} \, dx \\
=& \int_0^1 |\mathcal{L}| \cdot \underbrace{\left[\sum_{m=0}^{|\mathcal{N}|-|\mathcal{L}|-k} \binom{|\mathcal{N}|-|\mathcal{L}|-k}{m} \cdot x^{|\mathcal{N}|-|\mathcal{L}|-m-k} \cdot (1-x)^m \right]}_{=1} \cdot x^{k-1} \cdot (1-x)^{|\mathcal{L}|-1} \, dx \\
=& \int_0^1 |\mathcal{L}| \cdot x^{k-1} \cdot (1-x)^{|\mathcal{L}|-1} \, dx = |\mathcal{L}| \cdot B(k, |\mathcal{L}|) = \frac{1}{\binom{|\mathcal{L}|+k-1}{k-1}}
\end{aligned}$$

For ②, we have

$$
\begin{aligned}
② &= \sum_{m=1}^{|\mathcal{N}|-|\mathcal{L}|-k} (|\mathcal{N}|-|\mathcal{L}|-k) \cdot \binom{|\mathcal{N}|-|\mathcal{L}|-k-1}{m-1} \cdot B\Big(|\mathcal{N}|-|\mathcal{L}|-m, |\mathcal{L}|+m\Big) \\
&= \sum_{m'=0}^{|\mathcal{N}|-|\mathcal{L}|-k-1} (|\mathcal{N}|-|\mathcal{L}|-k) \cdot \binom{|\mathcal{N}|-|\mathcal{L}|-k-1}{m'} \cdot B\Big(|\mathcal{N}|-|\mathcal{L}|-m'-1, |\mathcal{L}|+m'+1\Big) \\
&= \int_0^1 (|\mathcal{N}|-|\mathcal{L}|-k) \sum_{m'=0}^{|\mathcal{N}|-|\mathcal{L}|-k-1} \binom{|\mathcal{N}|-|\mathcal{L}|-k-1}{m'} \cdot x^{|\mathcal{N}|-|\mathcal{L}|-m'-2} \cdot (1-x)^{|\mathcal{L}|+m'} \, dx \\
&= \int_0^1 (|\mathcal{N}|-|\mathcal{L}|-k) \underbrace{\left[\sum_{m'=0}^{|\mathcal{N}|-|\mathcal{L}|-k-1} \binom{|\mathcal{N}|-|\mathcal{L}|-k-1}{m'} \cdot x^{|\mathcal{N}|-|\mathcal{L}|-m'-k-1} \cdot (1-x)^{m'} \right]}_{=1} \cdot x^{k-1} \cdot (1-x)^{|\mathcal{L}|} \, dx \\
&= \int_0^1 (|\mathcal{N}|-|\mathcal{L}|-k) \cdot x^{k-1} \cdot (1-x)^{|\mathcal{L}|} \, dx = (|\mathcal{N}|-|\mathcal{L}|-k) \cdot B(k, |\mathcal{L}|+1) \\
&= \frac{|\mathcal{N}|-|\mathcal{L}|-k}{(|\mathcal{L}|+1)\binom{|\mathcal{L}|+k}{k-1}}
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
\alpha_{\mathcal{L}} &= ① + ② = \frac{1}{\binom{|\mathcal{L}|+k-1}{k-1}} + \frac{|\mathcal{N}|-|\mathcal{L}|-k}{(|\mathcal{L}|+1)\binom{|\mathcal{L}|+k}{k-1}} \\
&= \frac{|\mathcal{L}|! \cdot (k-1)!}{(|\mathcal{L}|+k-1)!} + \frac{|\mathcal{N}|-|\mathcal{L}|-k}{|\mathcal{L}|+1} \cdot \frac{(|\mathcal{L}|+1)! \cdot (k-1)!}{(|\mathcal{L}|+k)!} \\
&= \frac{|\mathcal{L}|! \cdot (k-1)!}{(|\mathcal{L}|+k-1)!} + \frac{|\mathcal{N}|-|\mathcal{L}|-k}{|\mathcal{L}|+k} \cdot \frac{|\mathcal{L}|! \cdot (k-1)!}{(|\mathcal{L}|+k-1)!} \\
&= \left[1 + \frac{|\mathcal{N}|-|\mathcal{L}|-k}{|\mathcal{L}|+k}\right] \cdot \frac{|\mathcal{L}|! \cdot (k-1)!}{(|\mathcal{L}|+k-1)!} \\
&= \frac{|\mathcal{N}|}{|\mathcal{L}|+k} \cdot \frac{|\mathcal{L}|! \cdot (k-1)!}{(|\mathcal{L}|+k-1)!} \\
&= \frac{|\mathcal{N}|}{k} \cdot \frac{|\mathcal{L}|! \cdot k!}{(|\mathcal{L}|+k)!} \\
&= \frac{|\mathcal{N}|}{k} \cdot \frac{1}{\binom{|\mathcal{L}|+k}{k}}
\end{aligned}
$$

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. Potential alternative settings for baseline values

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

- *Mean baseline values [14].* The baseline value of each input variable is set to the mean value of this variable over all samples, *i.e.* $\forall i \in \mathcal{N}, r_i = \mathbb{E}_{\boldsymbol{x}}[x_i]$.
- *Zero baseline values [4, 52].* 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 $v(\boldsymbol{x}_{\mathcal{S}})$, some studies [18, 19] removed variables from the input image by blurring the value of each input variable x_i $(i \in \mathcal{N} \setminus \mathcal{S})$ based on a Gaussian kernel.

However, defining optimal baseline values remains an open problem. Therefore, in this study, 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 for the tabular and NLP datasets. For the MNIST dataset, we initialize r_i to zero (*i.e.* black pixels) for each input variable i. Then, we optimize r_i to minimize Eq. (6) in the main paper while constraining it within a relatively small range, *i.e.*, $\|r_i - r_i^{\text{initial}}\|^2 \leq \tau$, to represent the absence state.

F. Simplifying the explanation using the minimum description length principle

In this section, we discuss the algorithm for 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 $L(g, \mathcal{M})$ is given in Eq. (8) of the main paper.

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

$$\delta(\alpha) = \frac{\Delta L}{|\alpha|} = \frac{L(g, \mathcal{M} \cup \{\alpha\}) - L(g, \mathcal{M})}{|\alpha|}, \tag{3}$$

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 extraction process stopped when there was no new coalition α that 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 when the most efficient α was not shared by multiple patterns.

G. More experimental details, results, and discussions

G.1. Datasets and models

Datasets. We conducted experiments on both natural language processing tasks and the classification/regression tasks based on tabular datasets. For natural language processing, we used the SST-2 dataset [49] for sentiment prediction and the CoLA dataset [58] for linguistic acceptability. For tabular datasets, we used the UCI census income dataset (*census*) [17], the UCI bike sharing dataset (*bike*) [17], and the UCI TV news channel commercial detection dataset (*TV news*) [17]. We followed [11, 12] to pre-process data for these tabular datasets. We also normalized the data in each dataset to a zero mean and unit variance.

Models. We trained the 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 architecture in [39]. In addition, for tabular datasets, we followed [11, 12] to train LightGBMs [31], XGBoost [10], and two-layer MLPs (*MLP-2*). We also trained five-layer MLPs (*MLP-5*) and five layer MLPs with skip-connections (*ResMLP-5*) on these datasets. For the ResMLP-5, we added a skip connection to each fully connected layer of the MLP-5. Figure 1 shows the architecture of the ResMLP-5. The hidden layers in MLP-5 and ResMLP-5 had the same width of 100. In our experiment, we also learned MLP-2, MLP-5, and ResMLP-5 on each tabular dataset via adversarial training [35]. During adversarial training, adversarial examples $x^{\text{adv}} = x + \delta$ 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.

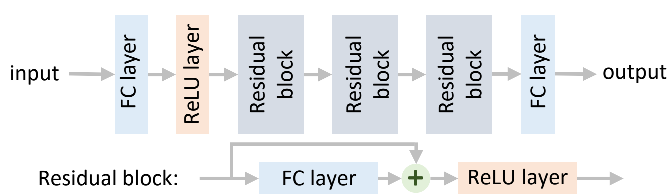


Figure 1. The architecture of the ResMLP-5.

Accuracy of models. Table 1 reports the classification accuracy of models trained on the TV news and census datasets, and the mean squared error of models trained on the bike dataset. Table 2 reports the classification accuracy of the models trained on the CoLA and SST-2 datasets. Table 3 reports the classification accuracy of the models trained on the MNIST dataset.

G.2. More visualization of AOGs

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

For tabular data, Figures 10, 11, 12, 13, and 14 show examples of AOGs generated by our method on different models trained on the census, bike, and TV news datasets. The up-arrow(↑) / down-arrow(↓) labeled in the terminal nodes indicated that the actual value of the input variable was greater than or less than the baseline value.

Table 1. Classification accuracy (on TV news and census dataset) and mean squared error (on bike dataset) of different models.

Dataset	MLP-2		MLP-5		ResMLP-5		XGBoost	LightGBM
	normal	adversarial	normal	adversarial	normal	adversarial		
TV news	83.11%	78.49%	79.86%	80.24%	79.01%	80.13%	84.48%	84.19%
census	79.91%	75.77%	78.96%	77.79%	80.49%	77.99%	87.35%	87.54%
bike	-	-	2161.47	3080.73	2149.43	2708.59	1623.71	-

Table 2. Accuracy of models trained on NLP datasets.

Dataset	LSTM	CNN
CoLA	64.42%	65.79%
SST-2	86.83%	78.19%

Table 3. Classification accuracy of models trained on the MNIST dataset.

Dataset	ResNet-20	ResNet-32	ResNet-44	VGG-16
MNIST	99.45%	99.57%	99.47%	99.68%

For the image data, Figure 6 shows an example of the AOG generated by our method on ResNet-18 trained on the CelebA dataset. The ResNet-18 was trained to classify the *eyeglasses* attribute. We manually segmented the facial parts and used these parts as input variables to construct the AOG. We found that salient patterns usually fitted human cognition. Figures 7, 8, and 9 show examples of the AOGs generated using our method on ResNet-32/44 and VGG-16 trained on the MNIST dataset, respectively. We manually segmented the digits in the MNIST dataset into eight connected parts, as the eight corresponding input variables of each DNN. We observed that the AOGs extracted meaningful digit shapes used by the DNN for inference.

For NLP data, Figures 15 and 16 show examples of the AOGs generated by our method on LSTMs and CNNs trained on the SST-2 and CoLA datasets. Furthermore, Figure 17 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 17(top), words "originality" and "delight" in Figure 17(middle), and words "painfully" and "bad" in Figure 17(bottom).

G.3. Details of experiments on synthesized functions and datasets

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

The Addition-Multiplication dataset [64]. This dataset contained 100 functions consisting of only addition and multiplication operations. For example, $v(\boldsymbol{x}) = x_1 + x_2x_3 + x_3x_4x_5 + x_4x_6$. Each variable x_i was a binary variable, *i.e.* $x_i \in \{0,1\}$.

The ground-truth causal patterns and there corresponding effects can be easily determined. For each term in these functions (*e.g.* the term $x_3x_4x_5$ in the function $v(\boldsymbol{x}) = x_1 + x_2x_3 + x_3x_4x_5 + x_4x_6$), 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 to form a ground-truth causal pattern. In the example function above, given the input $\boldsymbol{x} = [1,1,1,1,1,1]$, 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 $\boldsymbol{x} = [1,1,0,1,1,1]$, the ground-truth causal patterns were $\Omega^{\text{truth}} = \{\{x_1\},\{x_4,x_6\}\}$.

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

The dataset in [40]. This dataset contained 100 functions consisting of addition, subtraction, multiplication, and sigmoid operations. Similar to the Addition-Multiplication dataset, the ground-truth causal patterns in this dataset could also be easily determined. Let us consider the function $v(\boldsymbol{x}) = -x_1x_2x_3 - \text{sigmoid}(5x_4x_5 - 5x_6 - 2.5), x_i \in \{0,1\}$ 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 that this function contained two ground-truth causal patterns. In other words, for the above function, given the input $\boldsymbol{x} = [1,1,1,1,1,0]$, the ground-truth causal patterns were $\Omega^{\text{truth}} = \{\{x_1,x_2,x_3\},\{x_4,x_5,x_6\}\}$. Given the input $\boldsymbol{x} = [1,1,1,1,1,1]$, the ground-truth causal patterns were $\Omega^{\text{truth}} = \{\{x_1,x_2,x_3\}\}$.

In our experiments, we followed [40] to randomly generated 100 functions. Each of them had 6-12 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 ($m = |\Omega^{\text{truth}}|$) 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, let us consider the function $f(\boldsymbol{x}) = (x_1 > 0) \& (x_2 > 0) \,|\, (x_2 > 0) \& (x_3 > 0) \& (x_4 > 0) \,|\, (x_3 > 0) \& (x_5 > 0)$. Each input variable is a scalar, *i.e.* $x_i \in \mathbb{R}$, and the output is binary, *i.e.* $f(\boldsymbol{x}) \in \{0, 1\}$. 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 to be significant enough to be a valid ground-truth causal pattern. *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 $\Omega^{\text{truth}} = \{\{x_1, x_2\}, \{x_3, x_5\}\}$.

In our experiments, we trained one MLP-5 network and one ResMLP-5 network for binary classification using the manually labeled dataset generated based on each And-Or function. Similar to 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 the exact 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, $v(\boldsymbol{x}) = 3x_1 - 2x_2x_3 - x_3x_4x_5 + 5x_4x_6$.

The ground-truth causal effects in these functions can be easily determined. Similar to 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 $\mathcal{S} \subseteq \{x_1, ..., x_6\}$, $w_{\mathcal{S}} = 0$. 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 $\mathcal{S} \subseteq \{x_1, ..., x_6\}$, $w_{\mathcal{S}} = 0$.

In our experiments, we randomly generated 100 functions. Each of them had 10 input variables, and had 10-100 terms. Subsequently, 200 binary input samples were randomly generated for each function. For each input sample, we measured the Jaccard similarity coefficient $J = \frac{\sum_{\mathcal{S} \subseteq \mathcal{N}} \min(|w_{\mathcal{S}}^{\text{truth}}|, |w_{\mathcal{S}}|)}{\sum_{\mathcal{S} \subseteq \mathcal{N}} \max(|w_{\mathcal{S}}^{\text{truth}}|, |w_{\mathcal{S}}|)}$ between ground-truth causal effects $w_{\mathcal{S}}^{\text{truth}}$ (defined above) and causal effects $w_{\mathcal{S}}$ computed using 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.

G.4. More experimental results on the faithfulness of the AOG explainer

This section presents the results of the faithfulness of the AOG explainer on NLP and vision tasks. For NLP tasks, we used the SST-2 dataset. For the vision tasks, we used the MNIST and CelebA datasets. We computed the unfaithfulness metric ρ^{unfaith} to evaluate whether the explanation method faithfully extracted the causal effects encoded by the DNNs. Table 4 compares the extracted causal effects in the AOG with SI values, STI values, and attribution-based explanations (including the Shapley value [43], Input×Gradient [45], LRP [5], and Occlusion [63]). Our AOG explainer exhibited significantly lower ρ^{unfaith} values than the baseline methods.

Table 4. Unfaithfulness (↓) of different explanation methods on the NLP and vision tasks.

Dataset		DNN	Shapley	I×G	LRP	OCC	SI	STI (k=2)	STI (k=3)	Ours
NLP	SST-2	LSTM	15.8	1.0E+3	258	65.9	166	4.05	2.50	**1.4E-12**
		CNN	27.4	38.5	210	577	234	4.06	1.12	**6.7E-12**
Vision	MNIST	RN-20	22.6	303	349	21.6	234	3.44	0.47	**9.1E-14**
	CelebA	RN-18	1.57	5.1E+5	358	290	13.88	0.42	4.5E-2	**2.1E-13**

G.5. More analysis on the faithfulness of the AOG explainer

In this section, we discuss the experiment in Section 4.1 of the main paper, in which 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 $I^{\text{Shapley}}(\mathcal{S})$ [22] and the STI value $I^{\text{Shapley-Taylor}}(\mathcal{S})$ [51] as the numerical effects of different interactive patterns $\mathcal{S}$ on a DNN's

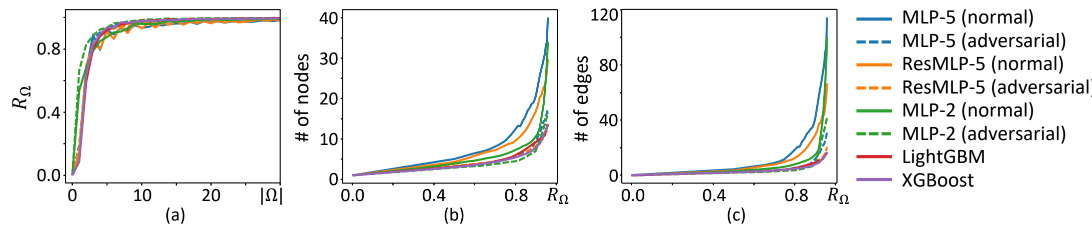


Figure 2. (a) The relationship between the number of causal patterns $|\Omega|$ in the AOG and the ratio of the explained causal effects R_Ω, based on the census dataset. The relationship between R_Ω and (b) the number of nodes, and (c) the number of edges in the AOG, based on the census dataset.

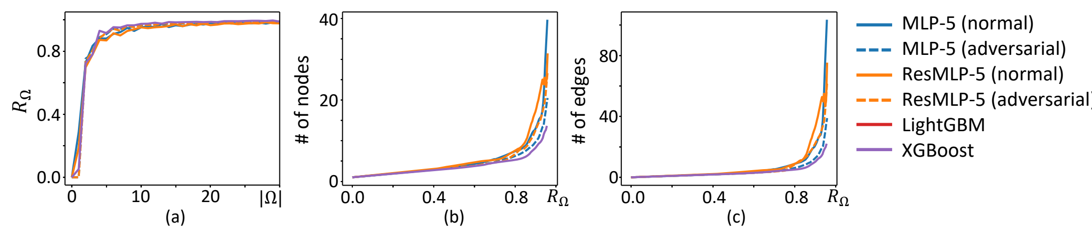


Figure 3. (a) The relationship between the number of causal patterns $|\Omega|$ in the AOG and the ratio of the explained causal effects R_Ω, based on the bike dataset. The relationship between R_Ω and (b) the number of nodes, and (c) the number of edges in the AOG, based on the bike dataset.

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 [43], Input×Gradient [45], LRP [5], Occlusion [63]).

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

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

$$\rho^{\text{unfaith}}_{\text{SI}} = \mathbb{E}_{\mathcal{S}\subseteq\mathcal{N}}[v(\boldsymbol{x}_\mathcal{S}) - \sum\nolimits_{\mathcal{S}'\subseteq\mathcal{S}} I^{\text{Shapley}}(\mathcal{S}')]^2, \quad \rho^{\text{unfaith}}_{\text{STI}} = \mathbb{E}_{\mathcal{S}\subseteq\mathcal{N}}[v(\boldsymbol{x}_\mathcal{S}) - \sum\nolimits_{\mathcal{S}'\subseteq\mathcal{S}} I^{\text{Shapley-Taylor}}(\mathcal{S}')]^2 \tag{4}$$

• For *attribution-based explainer models*, given the input sample $\boldsymbol{x}$, 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 using the Shapley value, Input × Gradient, LRP, and Occlusion, respectively. As previously mentioned, these attribution values quantify the causal effects of each variable i. Based on the SCM in Eq. (2) of the main paper, the unfaithfulness of these attribution-based explanations was similarly measured as follows.

$$\begin{aligned}\rho^{\text{unfaith}}_{\text{Shapley}} &= \mathbb{E}_{\mathcal{S}\subseteq\mathcal{N}}[v(\boldsymbol{x}_\mathcal{S}) - \sum\nolimits_{i\in\mathcal{S}} \phi^{\text{Shapley}}(i)]^2, \quad \rho^{\text{unfaith}}_{\text{IG}} = \mathbb{E}_{\mathcal{S}\subseteq\mathcal{N}}[v(\boldsymbol{x}_\mathcal{S}) - \sum\nolimits_{i\in\mathcal{S}} \phi^{\text{IG}}(i)]^2, \\ \rho^{\text{unfaith}}_{\text{LRP}} &= \mathbb{E}_{\mathcal{S}\subseteq\mathcal{N}}[v(\boldsymbol{x}_\mathcal{S}) - \sum\nolimits_{i\in\mathcal{S}} \phi^{\text{LRP}}(i)]^2, \quad \rho^{\text{unfaith}}_{\text{Occ}} = \mathbb{E}_{\mathcal{S}\subseteq\mathcal{N}}[v(\boldsymbol{x}_\mathcal{S}) - \sum\nolimits_{i\in\mathcal{S}} \phi^{\text{Occ}}(i)]^2\end{aligned} \tag{5}$$

Then, we compared the unfaithfulness of the AOG explainers using 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.

G.6. 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.

Similar to the experiment in the 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 2(a) and Figure 3(a) show the relationship between $|\Omega|$ and the ratio of explained causal effects R_Ω in different models, based on the census and bike datasets. We found that when we used a few causal patterns, we could explain most of the causal effects in the model output. Figure 2(b,c) and Figure 3(b,c) show that the node number and edge number increased with the increase in R_Ω.

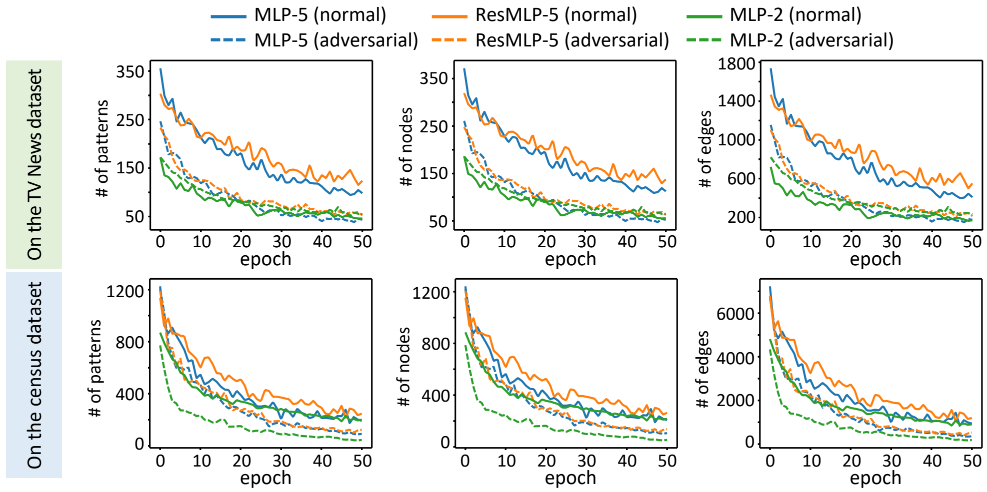


Figure 4. The number of patterns (the first column), nodes (the second column), and edges (the third column) in the AOG, based on baseline values of different learning epochs. The learned baseline value significantly enhanced the conciseness of explanations.

Besides, Figure 2(a) and Figure 3 also show that compared with the normally trained model, we could use fewer causal patterns (smaller $|\Omega|$) to achieve the same ratio of the explained causal effects R_Ω in the adversarially trained model. Moreover, Figure 2(b,c) and Figure 3(b,c) also show that the AOGs corresponding to adversarially trained models were less complex than the AOGs corresponding to normally trained models. This indicated that *adversarial training made models encode sparser causal patterns than normal training*.

G.7. More analysis on the effectiveness of the learned baseline values

This section provides experimental analysis of the effects of baseline values on the conciseness of explanations. In addition to the experiments in the Paragraph *Effects of baseline values on the conciseness of explanations* in Section 4.2 of the main paper, in this section, we analyze 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 the baseline values obtained in different epochs during the learning phase. Then, based on the computed causal effects, we measured the numbers of causal patterns, nodes, and edges in the AOG at each learning epoch. For a 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 4 shows the change in the AOG complexity during the learning process of baseline values, in terms of the number of causal patterns, nodes, and edges in the AOG. We found that learning the baseline values significantly simplified the AOG, thus boosting the conciseness of the explanations.

G.8. Comparing the complexity of AOGs and the complexity of DNNs

In this subsection, we compare the complexity of AOGs and the complexity of DNNs. We trained ResMLP networks with different numbers of layers on the Add-Mul and census datasets, and we explained these DNNs using AOGs. Figure 5 shows a comparison of the node number (complexity) of the AOG with the depth and parameter number (complexity) of the DNN. We found that a more complex DNN did not necessarily encode more complex features and thereby did not always obtaining a more complex AOG.

H. Discussion about the running time of the AOG explainer

In this section, we conducted an experiment to measure the running time of the methods in Table 2, Section 4.1 of the main paper. Specifically, we measured the average running time to compute the explanation of a single sample for MLP-5 trained on the census dataset. The running time was averaged over 20 different input samples. Table 5 shows that the proposed AOG explainer was comparable to the existing methods in terms of time complexity. For the implementation, we implemented the Harsanyi dividend, the Shapley value [43], the Shapley interaction index [22], and the Shapley Taylor interaction index [51] by ourselves, and implemented the other three methods (Input×Gradient [45], LRP [5], and Occlusion [63]) based on the Captum [32] package. All the computation was conducted using an NVIDIA GeForce RTX 2080 Ti GPU.

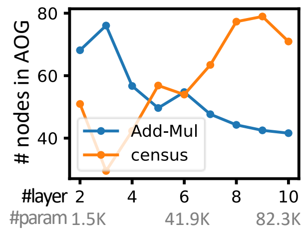


Figure 5. Comparing the complexity (the node number) of the AOG and the complexity (the parameter number) of the DNN.

Table 5. The average running time to compute the explanation of a single sample, based on different methods.

Method	SI	STI ($k=2$)	STI ($k=3$)	Shapley	IxG	LRP	Occ	Ours
Running time (s)	0.0179±0.0013	0.0176±1.6 × 10^{-5}	0.0176±3.9 × 10^{-5}	0.0179±0.0019	0.0045±0.0014	0.0170±0.0007	0.0302±0.0018	0.0182±0.0010

For high-dimensional inputs such as images, there are many techniques to solve the dimension problem and reduce the time cost. For example, we can manually segment an input into multiple parts, and use these parts as input variables to construct the AOG. In this way, the running time required to compute an AOG on the CelebA dataset was reduced to 4.03 s. Besides, we can also ignore casual patterns between distant parts to accelerate the computation.

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 [13, 48]? First, although the AOG explainer appears to be 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 of the main paper, the AOG explainer extracts different causal patterns from different input samples, instead of 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, which extracts the same set of features for each sample, the AOG explainer quantifies the manner in which the deep model triggers different causal patterns to handle different samples, thereby remaining non-linear for different inputs. Second, the BoW model considers only the presence or absence of input variables, whereas the AOG explainer is sensitive to the spatial relationships of input variables. For example, Table 6 shows the causal effects $w_{\mathcal{S}}$ of the same sets of words $\mathcal{S}$ encoded by the deep model[1], given two sentences with the same words but different word positions. We found that the deep model encoded significantly different causal effects between the same sets of words, demonstrating that the AOG explainer differs from the BoW model.

Table 6. Given two sentences with the same words but different word positions, the causal effects of the same sets of words $\mathcal{S}$ 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 $\mathcal{S}$	causal effects $w_{\mathcal{S}}$	sets of words $\mathcal{S}$	causal effects $w_{\mathcal{S}}$
{*just, not, smart, .*}	-1.616	{*not, just, smart, .*}	1.139
{*it, just, not, very*}	-1.510	{*it, not, just, very*}	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 the basic elementary concepts encoded by the deep model. For example, if two sentences contain the same set of words $\mathcal{S}$ in the same position, then the deep model encodes the same causal effects $w_{\mathcal{S}'}, \forall \mathcal{S}' \subseteq \mathcal{S}$. Table 7 shows that the deep model encoded the same causal effects within $\mathcal{S} = \{$*not, very, smart*$\}$ for two different sentences. From this perspective, such common causal patterns can be roughly considered as typical "words" in a BoW model.

[1] In 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 $v(\boldsymbol{x}_{\mathcal{S}}) = p(y = \text{positive sentiment}|\boldsymbol{x}_{\mathcal{S}})$.

Table 7. Given two sentences containing the same set of words $\mathcal{S} = \{not, very, smart\}$, the causal effects within the subset of words $\mathcal{S}$ encoded by the deep model were the same. The deep model encoded the same causal effects $w_{\mathcal{S}'}, \forall \mathcal{S}' \subseteq \mathcal{S}$.

Sentence 1: it's just not very smart.		Sentence 3: he is just not very smart.	
sets of words $\mathcal{S}' \subseteq \mathcal{S}$	causal effect $w_{\mathcal{S}'}$	sets of words $\mathcal{S}' \subseteq \mathcal{S}$	causal effect $w_{\mathcal{S}'}$
{*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

References

[1] Mahed Abroshan, Saumitra Mishra, and Mohammad Mahdi Khalili. Symbolic metamodels for interpreting black-boxes using primitive functions. *arXiv preprint arXiv:2302.04791*, 2023. 1

[2] Kamil Adamczewski, Frederik Harder, and Mijung Park. Bayesian importance of features (bif). *arXiv preprint arXiv:2010.13872*, 2010. 1

[3] David Alvarez-Melis and Tommi S Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *EMNLP*, 2017. 1

[4] 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*, pages 272–281. PMLR, 2019. 11

[5] 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. 14, 15, 16

[6] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Pietro Lió, Marco Gori, and Stefano Melacci. Entropy-based logic explanations of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6046–6054, 2022. 1

[7] Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective. In *ICML*, 2019. 1

[8] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable deep models for icu outcome prediction. In *AMIA annual symposium proceedings*, volume 2016, page 371. American Medical Informatics Association, 2016. 1

[9] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 882–891. PMLR, 2018. 1

[10] 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*, pages 785–794, 2016. 12

[11] Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3457–3465. PMLR, 2021. 12

[12] 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. 12

[13] 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. 1, 17

[14] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *arXiv preprint arXiv:1705.07857*, 2017. 11

[15] Huiqi Deng, Qihan Ren, Hao Zhang, and Quanshi Zhang. Discovering and explaining the representation bottleneck of dnns. In *International Conference on Learning Representations*, 2022. 1

[16] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4829–4837, 2016. 1

[17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. 12

[18] 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*, pages 2950–2958, 2019. 11

[19] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017. 1, 11

[20] Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017. 1

[21] 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. 1

[22] 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. 1, 2, 8, 14, 16

[23] Michael Harradon, Jeff Druce, and Brian Ruttenberg. Causal learning and explanation of deep neural networks via autoencoded activations. *arXiv preprint arXiv:1802.00541*, 2018. 1

[24] John C Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963. 1, 2

[25] 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. 1

[26] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, Bernhard Schölkopf, et al. Nonlinear causal discovery with additive noise models. In *NIPS*, 2008. 1

[27] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *AAAI*, 2019. 1

[28] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. *Advances in Neural Information Processing Systems*, 32:15883–15893, 2019. 1

[29] Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *arXiv preprint arXiv:2002.04138*, 2020. 1

[30] 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. 1

[31] 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. 12

[32] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020. 16

[33] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018. 1

[34] 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*, pages 4768–4777, 2017. 1

[35] 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. 12

[36] Joao Marques-Silva, Thomas Gerspacher, Martin C Cooper, Alexey Ignatiev, and Nina Narodytska. Explanations for monotonic classifiers. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7469–7479. PMLR, 18–24 Jul 2021. 1

[37] 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. 1

[38] Judea Pearl. *Causality*. Cambridge university press, 2009. 1

[39] A Rakhlin. Convolutional neural networks for sentence classification. *GitHub*, 2016. 12

[40] Jie Ren, Zhanpeng Zhou, Qirui Chen, and Quanshi Zhang. Can we faithfully represent masked states to compute shapley values on a dnn? In *The eleventh International Conference on Learning Representations, ICLR 2023, Kigali Rwanda, May 1-5, 2023*, 2023. 6, 13

[41] 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*, pages 1135–1144, 2016. 1

[42] 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*, pages 618–626, 2017. 1

[43] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953. 2, 6, 14, 15, 16

[44] Andy Shih, Arthur Choi, and Adnan Darwiche. Compiling bayesian network classifiers into decision graphs. In *AAAI*, 2019. 1

[45] 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. 14, 15, 16

[46] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1

[47] Chandan Singh, W James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*, 2018. 1

[48] 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, pages 1470–1470. IEEE Computer Society, 2003. 1, 17

[49] 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*, pages 1631–1642, 2013. 12

[50] 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*, pages 1000–1007, 2008. 1

[51] Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International Conference on Machine Learning*, pages 9259–9268. PMLR, 2020. 1, 2, 5, 14, 16

[52] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328, 2017. 11

[53] 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. 1

[54] Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-shap: The faithful shapley interaction index. *arXiv preprint arXiv:2203.00870*, 2022. 1

[55] 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. 1

[56] 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*, pages 721–729. PMLR, 2021. 1

[57] Zhuo Wang, Wei Zhang, Ning Liu, and Jianyong Wang. Scalable rule-based representation learning for interpretable classification. *Advances in Neural Information Processing Systems*, 34:30479–30491, 2021. 1

[58] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. 12

[59] 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. 1

[60] Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34:10823–10836, 2021. 1

[61] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. INVASE: instance-wise variable selection using neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. 1

[62] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. 1

[63] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1, 14, 15, 16

[64] Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. Interpreting multivariate shapley interactions in dnns. In *AAAI*, 2021. 13

[65] 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. 1

[66] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *In ICLR*, 2015. 1

[67] 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*, pages 2921–2929, 2016. 1

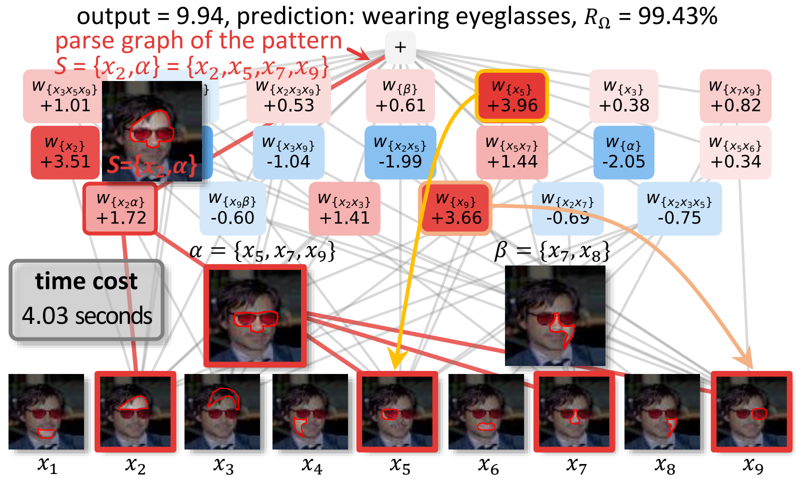


Figure 6. An examples of AOGs extracted from the ResNet-18 network, trained on the CelebA dataset.

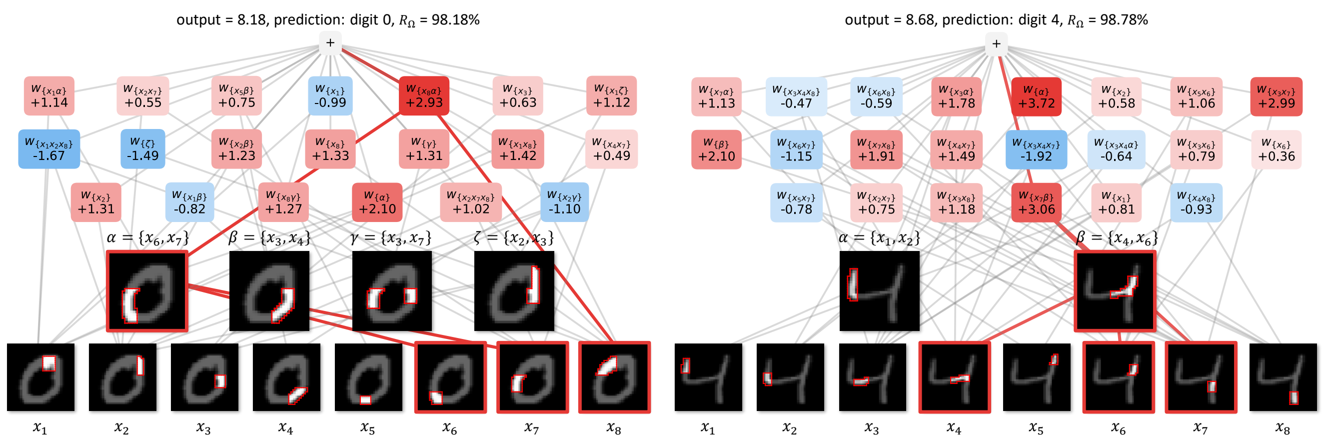


Figure 7. Examples of AOGs extracted from the ResNet-32 network, trained on the MNIST dataset.

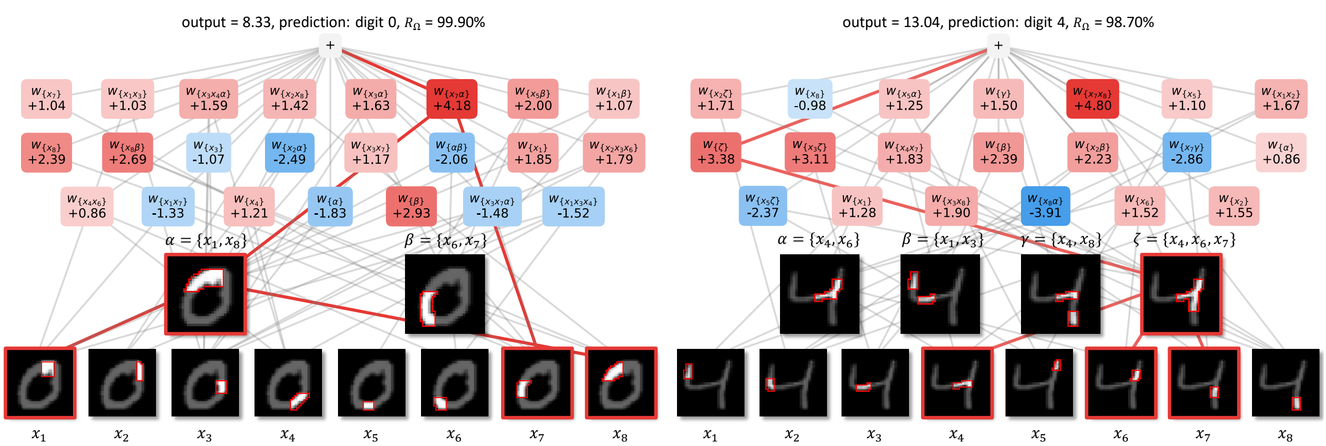


Figure 8. Examples of AOGs extracted from the ResNet-44 network, trained on the MNIST dataset.

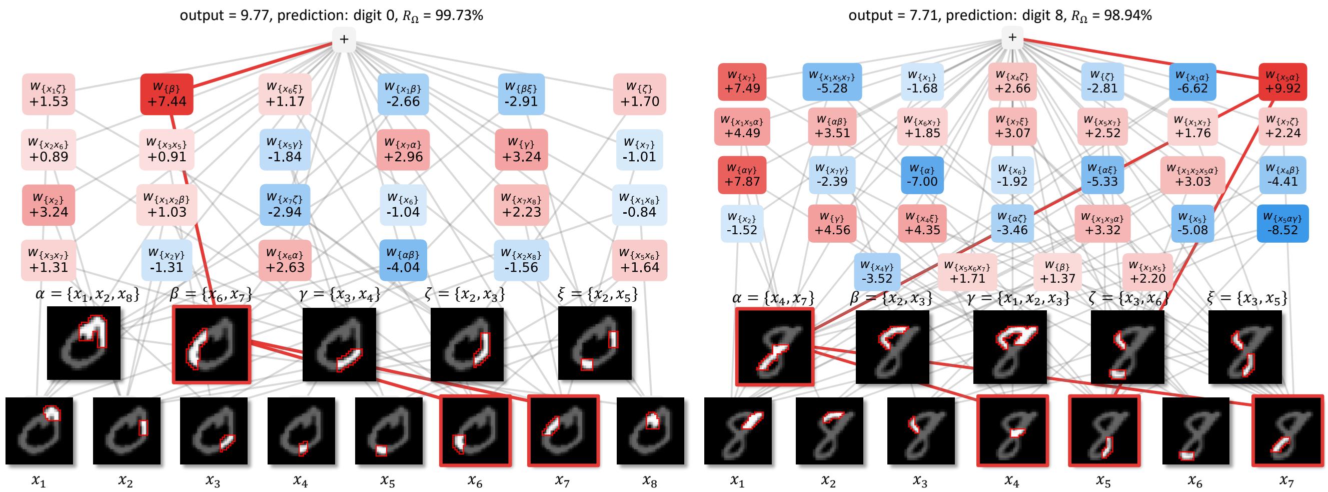


Figure 9. Examples of AOGs extracted from the VGG-16 network, trained on the MNIST dataset.

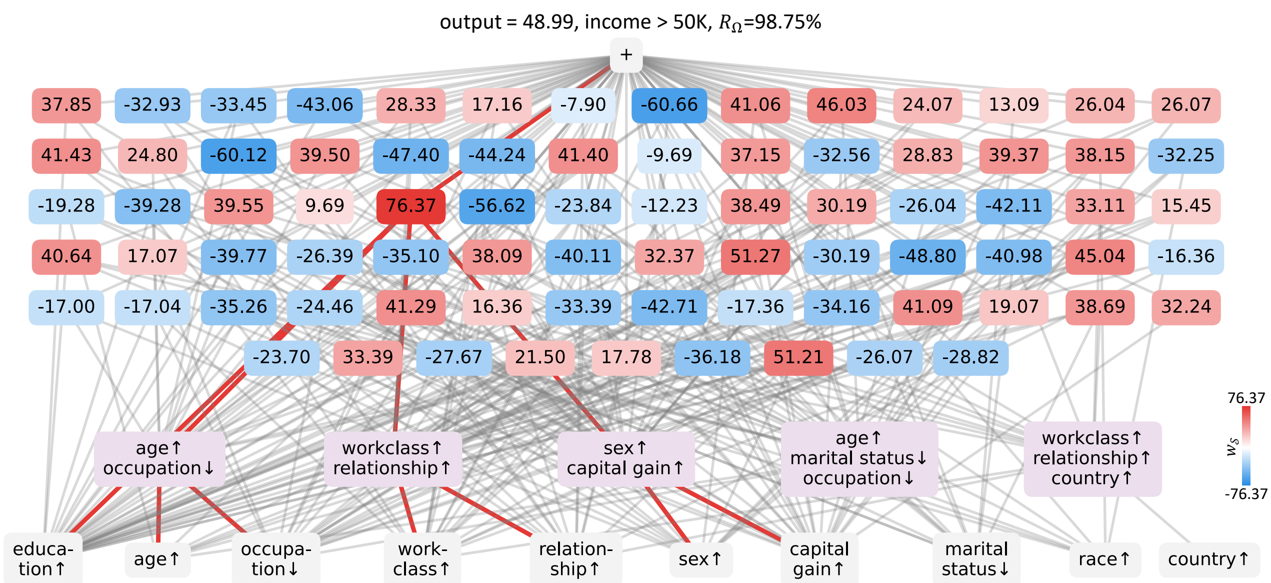


Figure 10. An example of the AOG extracted from the MLP-5 network, trained on the census dataset. Red edges indicate the parse graph of the most salient causal pattern.

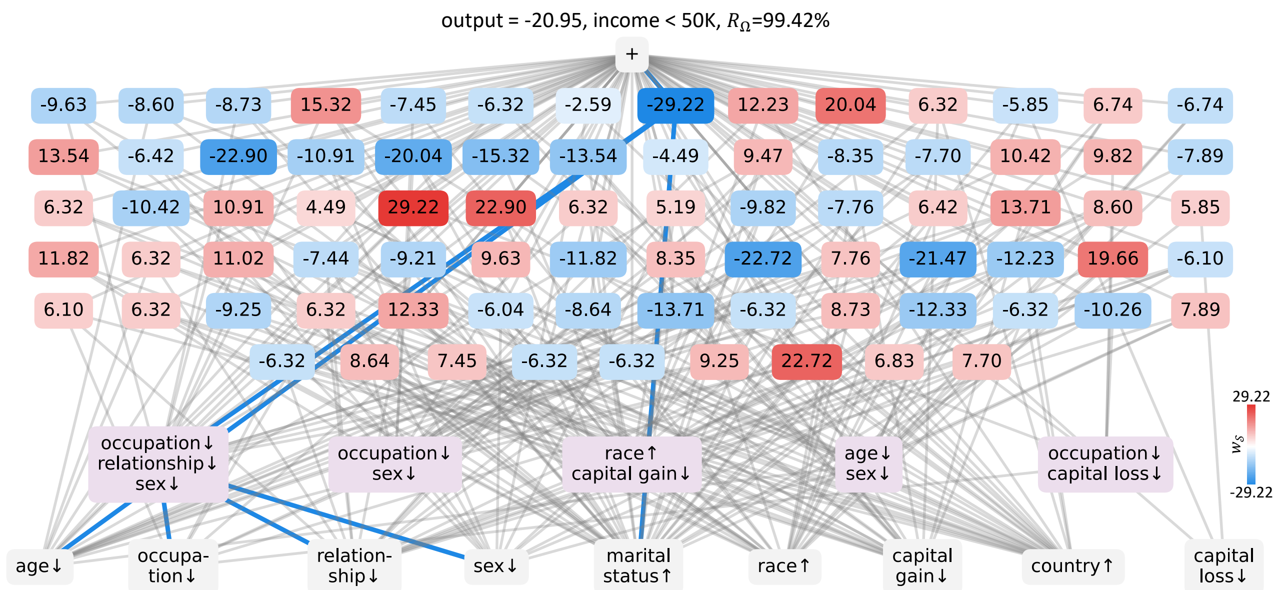
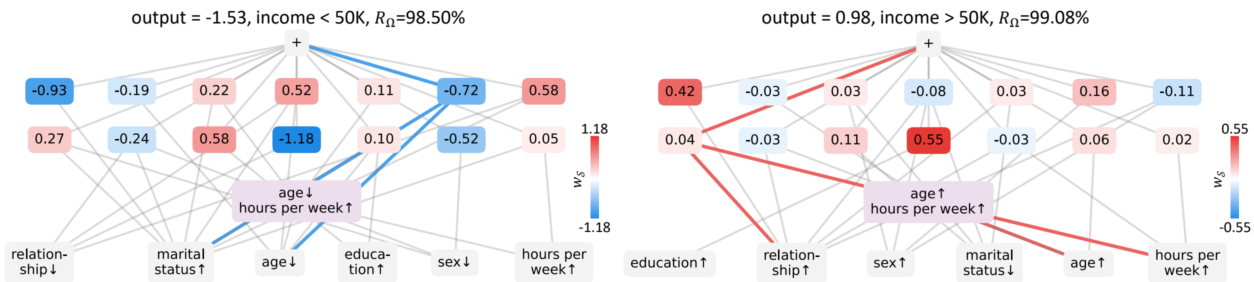
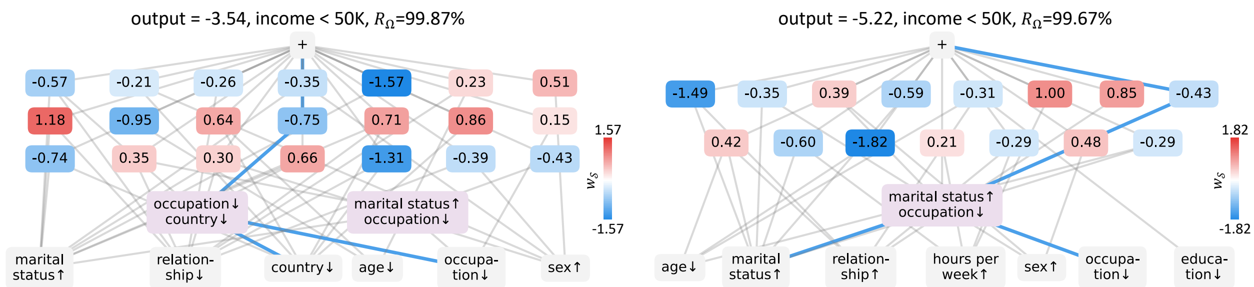


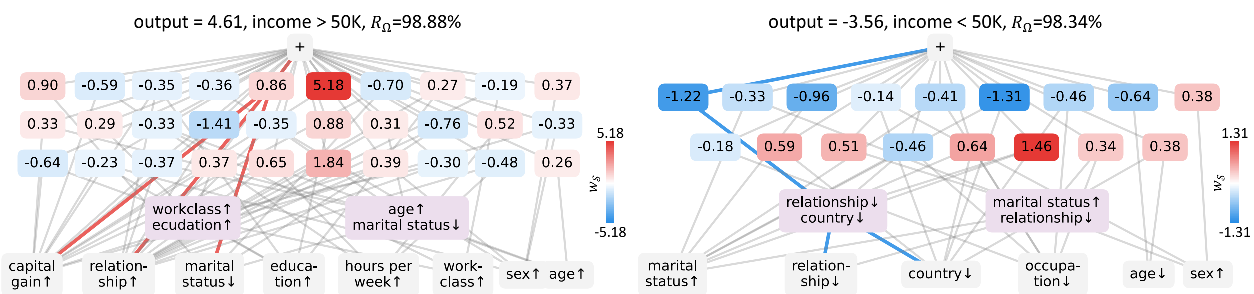
Figure 11. An example of the AOG extracted from the ResMLP-5 network, trained on the census dataset. Red edges indicate the parse graph of the most salient causal pattern.



(a) Examples of AOGs extracted from the MLP-2 network, adversarially trained on the census dataset.

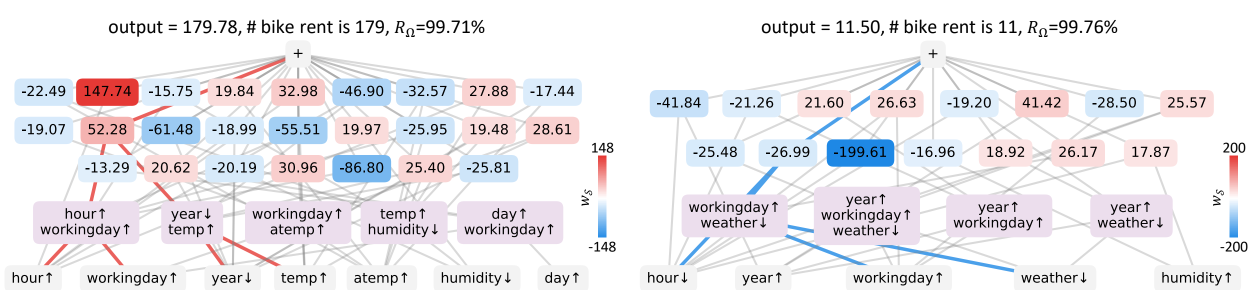


(b) Examples of AOGs extracted from the MLP-5 network, adversarially trained on the census dataset.

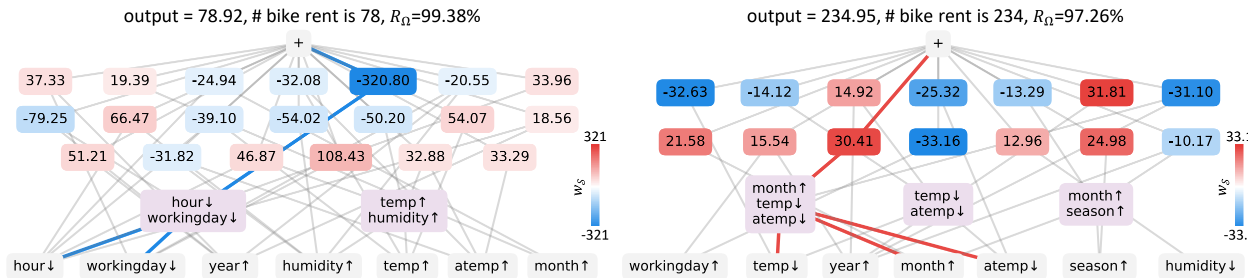


(c) Examples of AOGs extracted from the ResMLP-5 network, adversarially trained on the census dataset.

Figure 12. Examples of AOGs extracted from models trained on the census dataset. Red edges indicate the parse graph of a specific causal pattern.

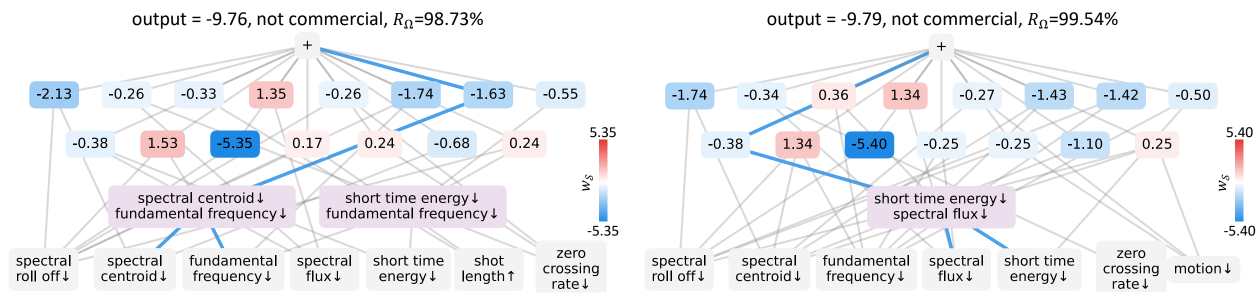


(a) Examples of AOGs extracted from the MLP-5 network, adversarially trained on the bike dataset.

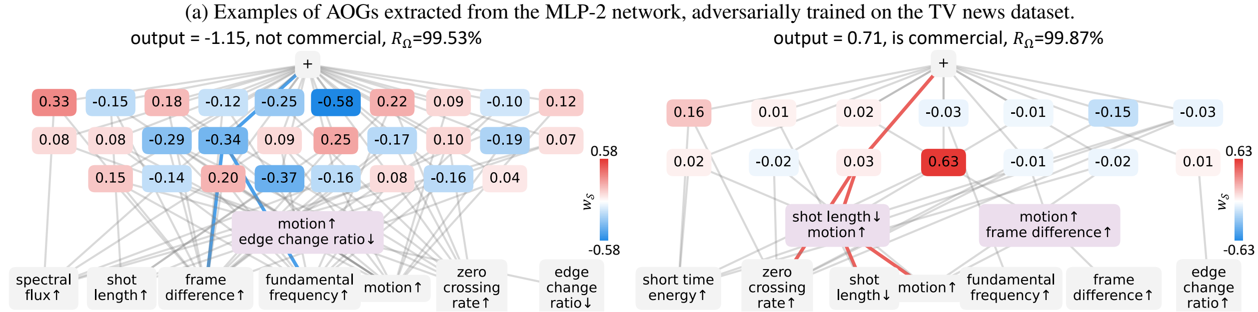


(b) Examples of AOGs extracted from the ResMLP-5 network, adversarially trained on the bike dataset.

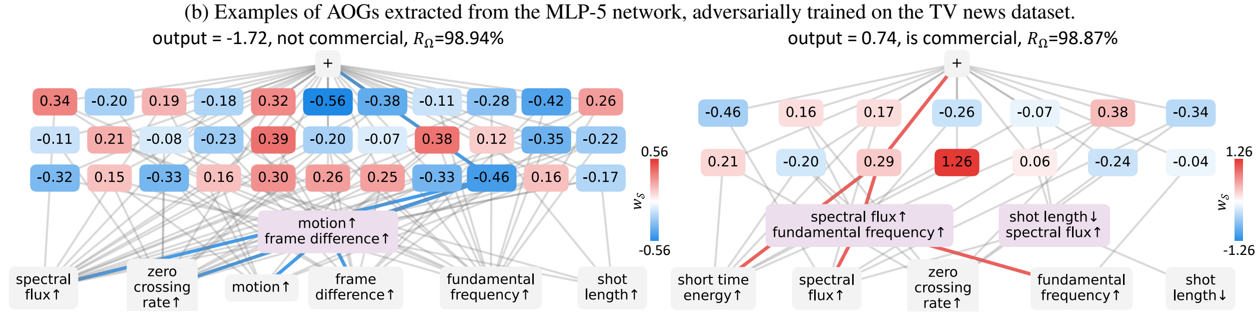
Figure 13. Examples of AOGs extracted from models trained on the bike dataset. Red edges indicate the parse graph of a specific causal pattern.



(a) Examples of AOGs extracted from the MLP-2 network, adversarially trained on the TV news dataset.

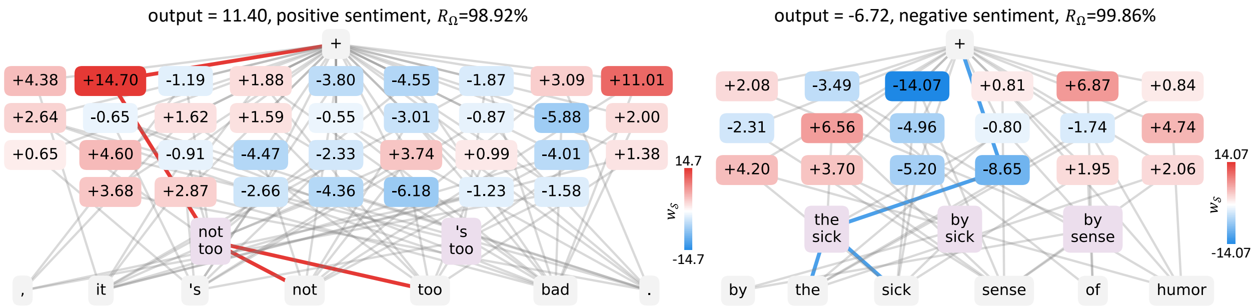


(b) Examples of AOGs extracted from the MLP-5 network, adversarially trained on the TV news dataset.

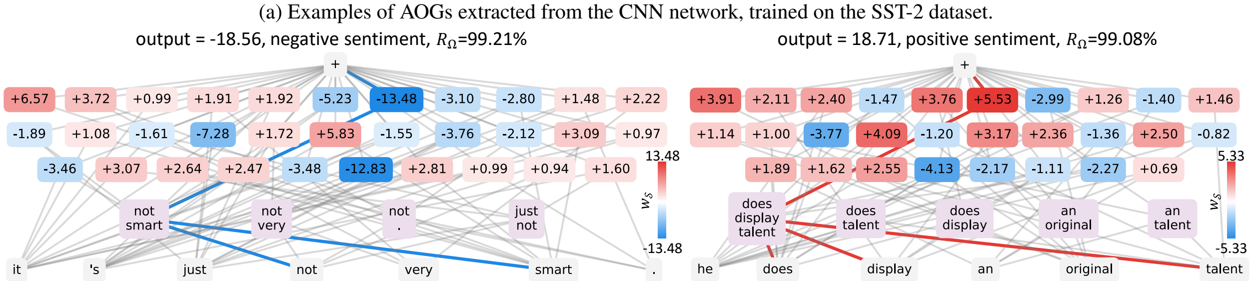


(c) Examples of AOGs extracted from the ResMLP-5 network, adversarially trained on the TV news dataset.

Figure 14. Examples of AOGs extracted from models trained on the TV news dataset. Red edges indicate the parse graph of a specific causal pattern.

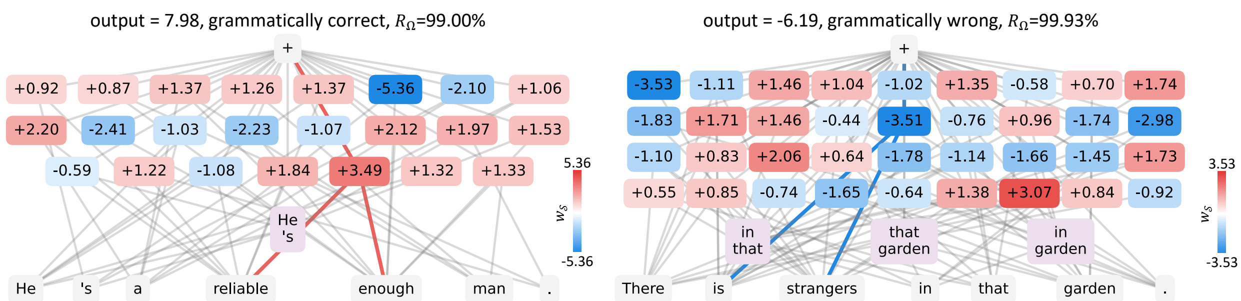


(a) Examples of AOGs extracted from the CNN network, trained on the SST-2 dataset.

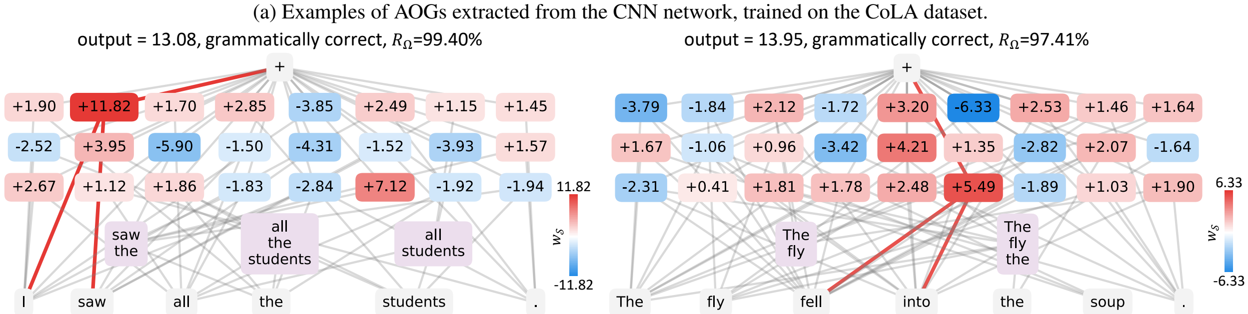


(b) Examples of AOGs extracted from the LSTM network, trained on the SST-2 dataset.

Figure 15. 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 16. Examples of AOGs extracted from models trained on the CoLA dataset. Red edges indicate the parse graph of the most salient causal pattern.

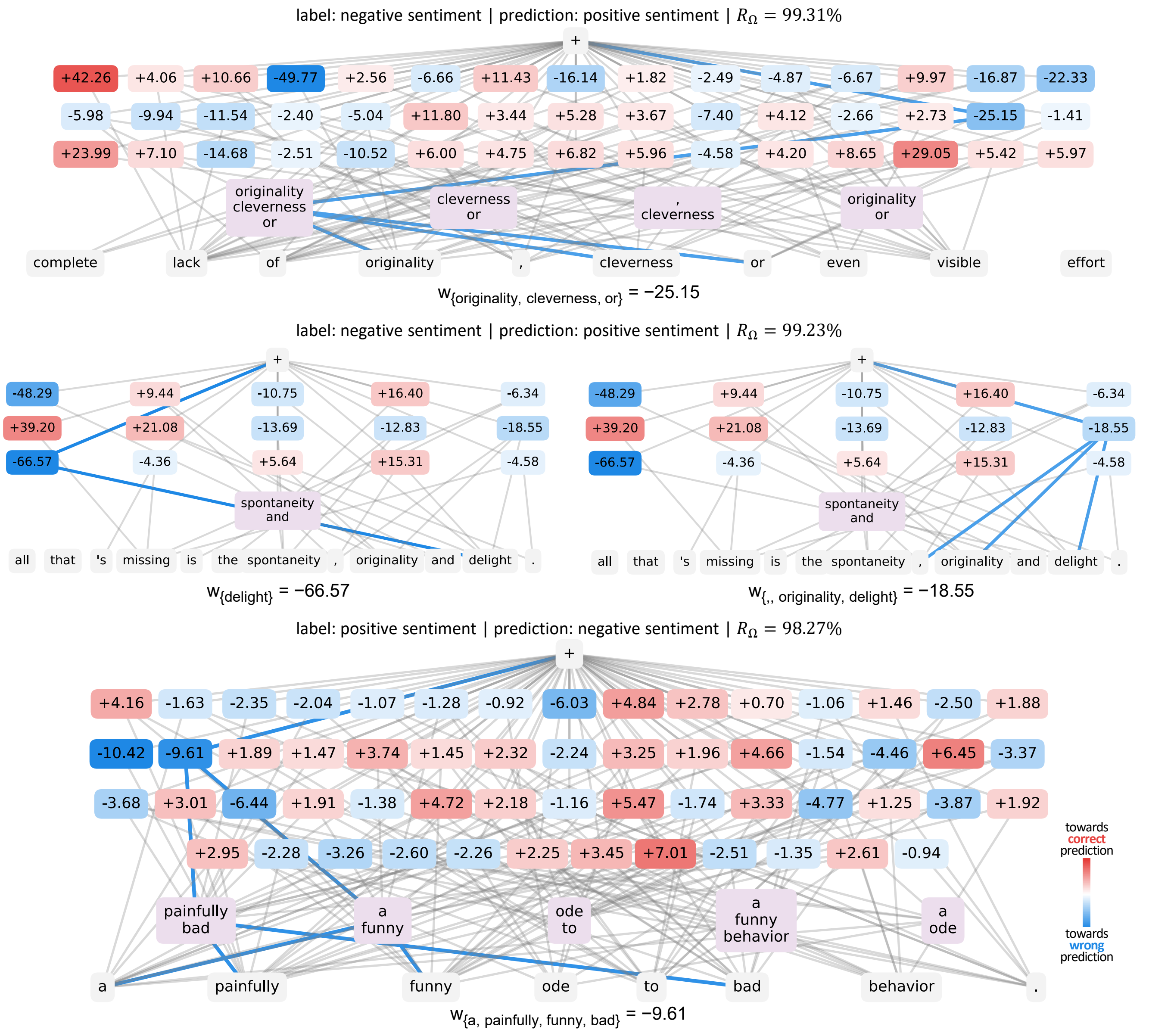


Figure 17. AOGs that explained incorrect predictions of the network model trained on the SST-2 dataset. Red edges indicated the parse graphs of causal patterns towards correct predictions, while blue edges indicated parse graphs of causal patterns towards wrong predictions.