AUDITING HIGHER-ORDER FEATURE INTERACTIONS VIA INTERVENTIONAL DECOMPOSITION

Anonymous authors
Paper under double-blind review

Signed pairwise interaction scores are widely used in explainable AI to classify feature cooperation as synergy (if positive) or redundancy (if negative). However, when interactions involve more than two features, pairwise scores fail: they either vanish entirely or flip sign across inputs, mislabeling synergy as redundancy. A single scalar summary is fundamentally insufficient to disentangle uniqueness (U), redundancy (R), and synergy (S). To overcome this limitation, we introduce INDE-X: Interventional Decomposition for eXplainability. While classical predictability decompositions study theoretical data capacity by retraining models on feature subsets, our method audits the internal causal pathways of a fixed, deployed network via post-hoc interventional masked inference, decomposing marginal predictive gains into per-feature U/R/S profiles over the Boolean lattice. We provide finite-sample concentration bounds, strict variance reduction via coupled diamond sampling, and uniform convergence guarantees over finite coalition vocabularies. INDE-X recovers ground-truth roles on tabular SCMs (overcoming up to $411\times$ baseline conflation), surfaces synergistic and backup circuits in GPT-2 missed by activation patching, and identifies multi-region anatomical co-dependencies on ChestX-ray14 with superior deletion faithfulness. Code and data are available at INDE-X repository.

1 INTRODUCTION

Modern deep networks route information through complex, non-additive pathways, raising a fundamental question: is joint feature dependence *unique*, *redundant*, or truly *synergistic*? In explainable AI (XAI), feature cooperation is standardly quantified through signed pairwise interaction scores (Muschalik et al., 2024; Hedström et al., 2023), interpreting positive values as synergy and negative values as redundancy. While higher-order formulations exist in principle (Sundararajan et al., 2020), evaluating interactions beyond pairs is computationally prohibitive for high-dimensional models, forcing practical XAI to rely on order-2 truncations. We show that under higher-order cooperation, this pairwise truncation faces an inescapable representational dilemma. On a minimal 3-way XOR structural causal model ($\oplus\text{SCM}_3$), faithful interaction indices (e.g., Shapley-Taylor at full order (Sundararajan et al., 2020), Faith-SHAP (Tsai et al., 2023)) strictly vanish at the pair level, erasing the cooperative signal. Conversely, projective indices truncated at order 2 (e.g., standard SII (Grabisch & Roubens, 1999), k-SII, kADD-SHAP (Muschalik et al., 2024)) project triplet interactions into spurious pairwise coefficients of $\pm 1/4$ whose sign flips with the input's local parity (Proposition 1). Thus, identical cooperative mechanisms are arbitrarily misclassified as synergy in one instance and redundancy in another. A single scalar summary cannot isolate how much of a feature's role is unique (U), redundant (R), or synergistic (S). Existing alternative paradigms fail to address this auditing challenge. Classical variance decompositions (functional ANOVA (Hoeffding, 1948; Hooker, 2004)) and information-theoretic Partial Information Decompositions (PID (Williams & Beer, 2010)) operate on observational data distributions, which describe statistical dependencies rather than causal reliance. Predictability frameworks like Hi-Fi (Ontivero-Ortega et al., 2025) operationalize the U/R/S triad by retraining a combinatorial suite of 2^n models on feature subsets. For auditing deployed networks, retraining is computationally demanding and conceptually divergent: it forces the architecture to learn alternative decision rules, answering what a hypothetical model could learn from data rather than what the frozen predictor f actually computes.

To bridge this gap, we formulate INDE-X (**In**terventional **De**composition for e**X**plainability), lifting the U/R/S codomain to the post-hoc auditing of fixed deep networks. Instead of retraining, INDE-X measures marginal loss variations under interventional feature masking against a reference background distribution p_{bg} (Janzing et al., 2020), isolating the model's true internal dependencies while preserving its learned representations (see sensitivity ablations in Appendix H). Computationally, INDE-X operates over the Boolean lattice $(\mathcal{P}([n]), \subseteq)$, maintaining a shared coalition-loss vocabulary updated online via Welford's algorithm. To reduce Monte Carlo estimation variance under an explicit

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Taxonomy of interaction explanation frameworks. INDE-X provides a mechanism-disentangled U/R/S profile for deployed models using a training-free interventional baseline masking.

Framework / Methods	Target Estimand	Output Codomain	Interventional	Retraining-Free
Shapley Indices (SII, SHAP-IQ, TN-SHAP)	Game-theoretic payoffs	Scalar signed score	No (Observational)	Yes
Functional ANOVA (Sobol, STRIDE)	Observational variance	Additive terms	No (Observational)	Yes
Masking / Perturbation (RISE)	Deletion AUC	Scalar score	No (Perturbation)	Yes
Predictability Decomp. (Hi-Fi)	Retrained predictivity	U/R/S profile	Yes (Counterfactual)	No (Retrains)
PID (Williams–Beer)	Information decomposition	U/R/S profile	No (Observational)	No
Higher-Order Interactions (Archipelago, HO-IG, T-NID, H-Sets)	Interaction attribution	Scalar / Tensor	No (Observational)	Yes
INDE-X (Ours)	**Interventional LOCO loss**	**Per-feature U/R/S**	**Yes (Baseline imputation)**	**Yes**

covariance condition (Theorem 2), we introduce coupled "diamond" queries that evaluate adjacent coalitions $(C, C \cup \{i\}, C \cup \{j\}, C \cup \{i, j\})$ on identical background batches. Our contributions are:

1. **Pairwise truncation dilemma.** We prove that truncated scalar interaction indices under higher-order cooperation either vanish or flip signs with input parity, arbitrarily conflating synergy and redundancy (Proposition 1).
2. **Retraining-free interventional estimator.** We formulate an operational U/R/S decomposition over the Boolean lattice with finite-sample bounds, strict variance reduction via diamond coupling (Theorem 2), and uniform convergence.
3. **Downstream causal validation.** Across tabular SCMs, INDE-X overcomes $411\times$ baseline conflation. On GPT-2 IOI (Wang et al., 2023), predicted interactions accurately track double-knockout excess ($r = -0.893$) and guide near-optimal circuit pruning (93.9% at $k = 6$). On ChestX-ray14 (Wang et al., 2017), it uncovers distal anatomical co-dependencies while improving deletion faithfulness ($0.332 \to 0.269$).

2 RELATED WORK

We organize prior interaction literature along two foundational axes: the *output codomain* (scalar vs. U/R/S-disentangled) and the *target estimand* (observational/retraining vs. post-hoc interventional loss); see Table 1. ***Output codomain.*** The dominant explainability paradigm compresses feature cooperation into signed scalar scores. Even richer functional forms, such as Archipelago (Tsang et al., 2020), Higher-Order IG (Janizek et al., 2021), T-NID (Lerman et al., 2021), and H-Sets (Mehrotra et al., 2026), or graph-restricted variants (Myerson, 1977; Bui et al., 2024) preserve a *non-disentangled* codomain (scalar set scores or interaction tensors) that cannot factorize into U/R/S channels. As formalized in Proposition 1, this limitation is structural: projecting higher-order interactions down to signed scalars inevitably erases synergy or misdiagnoses it as redundancy. ***Target estimand.*** Alternative decomposition traditions address parts of this gap but target fundamentally different estimands. Variance-based methods (Hoeffding (1948), Sobol' (2001), STRIDE (Ko, 2025)) decompose observational variance rather than interventional loss. Conversely, Partial Information Decomposition (PID (Williams & Beer, 2010)) provides the principled U/R/S codomain (Bertschinger et al., 2014; Kolchinsky, 2022), but predictive operationalizations (Hi-Fi (Ontivero-Ortega et al., 2025), Diffusion-PID (Dewan et al., 2024)) require combinatorial model retraining. While this approach characterizes what hypothetical models could learn from data distributions, it cannot audit the fixed computational pathways of a deployed predictor f.

To audit frozen networks, INDE-X causalizes leave-one-covariate-out loss (LOCO (Lei et al., 2018)) via post-hoc masked inference. We fuse PID's U/R/S codomain with the interventional attribution framework of Janzing et al. (2020), separating it from observational occlusion methods (Petsiuk et al., 2018; Fong & Vedaldi, 2017; Zeiler & Fergus, 2014) and generalizing activation-patching and causal-scrubbing workflows (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023; Conmy et al., 2023) to higher-order interactions. The background distribution p_{bg} defines the estimand (Sundararajan & Najmi, 2020): following Janzing et al. (2020), we use marginal background interventions to assess internal model reliance without requiring a learned covariate DAG (Heskes et al., 2020). Finally, admitting signed uniqueness under interventional contrasts matches relaxed PID formulations (Pakman et al., 2021; James et al., 2019), recovering classical Shannon PID under log-loss.

3 INTERVENTIONAL DECOMPOSITION AND THEORETICAL GUARANTEES

To resolve the limits of scalar interactions, we introduce INDE-X (**In**terventional **De**composition for e**X**plainability). We formalize the interventional U/R/S decomposition on the Boolean lattice, establish online vocabulary estimation with finite-sample bounds, prove variance reduction via diamond coupling, and derive uniform convergence under budgeted inference.

3.1 INTERVENTIONAL LOSS AND LOCO U/R/S DECOMPOSITION

Interventional coalition loss. Let $f : \mathcal{X} \to \mathcal{Y}$ be a fixed predictor, $x \in \mathcal{X}$ an input with features indexed by $[n] \coloneqq \{1, \ldots, n\}$, $y \in \mathcal{Y}$ its target, and $S \subseteq [n]$ a coalition with indicator mask $m_S \in \{0,1\}^n$ where $(m_S)_i = 1$ if $i \in S$ and 0 otherwise. Given background distribution $Z \sim p_{\mathrm{bg}}$ (Appendix H.5) and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we define the interventional coalition loss:

$$L(S; x, y) = \mathbb{E}_{Z \sim p_{\mathrm{bg}}} \left[\ell \left(f(m_S \odot x + (1 - m_S) \odot Z), y \right) \right], \tag{1}$$

where $\odot$ denotes the element-wise Hadamard product. Equation 1 audits a deployed, frozen predictor via atomic marginal intervention (baseline imputation) (Janzing et al., 2020), directly isolating the model's functional reliance on the retained features while avoiding observational confounders and retraining artifacts (Heskes et al., 2020). Under log-loss $\ell(\hat{y}, y) = -\log p(y \mid \hat{y})$, $L(S)$ corresponds to conditional cross-entropy, naturally recovering classical Shannon information measures.

LOCO U/R/S decomposition. For a target feature $i \in [n]$ and context coalition $C \subseteq [n] \setminus \{i\}$, we define the leave-one-covariate-out (LOCO (Lei et al., 2018)) gain as:

$$\mathrm{LOCO}(i \mid C) \coloneqq L(C; x, y) - L(C \cup \{i\}; x, y). \tag{2}$$

Evaluating this marginal gain across inclusion contexts determines the bounds of the feature's predictive utility:

$$U_i = \min_{C \subseteq [n] \setminus \{i\}} \mathrm{LOCO}(i \mid C), \qquad L_i^{\max} = \max_{C \subseteq [n] \setminus \{i\}} \mathrm{LOCO}(i \mid C), \qquad \pi_i = \mathrm{LOCO}(i \mid \emptyset), \tag{3}$$

where U_i represents guaranteed utility in the least favorable context, and π_i captures the first-order standalone effect. Maximum predictive capacity $L_i^{\max}$ decomposes exactly into redundancy R_i and synergy S_i:

$$R_i = \pi_i - U_i, \qquad S_i = L_i^{\max} - \pi_i, \qquad \text{yielding} \quad L_i^{\max} = U_i + R_i + S_i. \tag{4}$$

Lattice grounding: Dedekind vs. Boolean mask lattice. Classical Partial Information Decomposition (PID) (Williams & Beer, 2010) constructs a redundancy lattice over antichains of subsets. However, this formulation is computationally and statistically intractable for deep network auditing: the Dedekind cardinality $M(n)$ scales super-exponentially ($M(5) = 7{,}581$, $M(6) \approx 7.8 \times 10^6$), and evaluating continuous mutual information requires intractable density estimation. INDE-X circumvents both limitations by grounding the decomposition on the Boolean mask lattice $(\mathcal{P}([n]), \subseteq)$ of size 2^n. On this lattice, our $\min / \max$ LOCO operators provide the operational counterpart to modern variational PID frameworks (Bertschinger et al., 2014; Kolchinsky, 2022), directly auditing the loss landscape without density estimators.

Antagonistic uniqueness ($U_i < 0$) in loss space. Unlike Shannon mutual information, which is strictly non-negative, loss-space LOCO gains can take both positive and negative values. A negative uniqueness ($U_i < 0 \iff \min_C \mathrm{LOCO}(i \mid C) < 0$) has a precise operational meaning: it proves the existence of an adverse context $C^* \subseteq [n] \setminus \{i\}$ where incorporating feature i actively increases model loss ($L(C^* \cup \{i\}) > L(C^*)$), even if i delivers positive standalone ($\pi_i > 0$) or synergistic ($S_i > 0$) utility in other contexts. Rather than an estimator artifact, signed atoms are well-established in continuous and relaxed information decompositions (James et al., 2019; Pakman et al., 2021). In mechanistic interpretability, preserving the sign of U_i is indispensable for identifying inhibitory pathways, destructive interference, and backup mechanisms, such as the negative name-mover suppression exhibited by head 9.6 in GPT-2 IOI (Section 4).

Representational bottleneck of pairwise truncation. Why is this 3-channel profile necessary? While game-theoretic interactions can in principle extend to higher orders (Sundararajan et al., 2020; Tsai et al., 2023), evaluating all k-way subsets scales combinatorially as $\mathcal{O}(n^k)$, compelling practical XAI to rely on order-2 truncations (Muschalik et al., 2024; Hedström et al., 2023). Under higher-order cooperation, this truncation induces an inescapable representational failure.

Proposition 1. ***Instability of Truncated Pairwise Interactions under Parity.*** *Let $Y = \oplus(X_1, X_2, X_3)$ with $X_1, X_2, X_3 \sim$ iid Bernoulli$(1/2)$, evaluated under uniform binary background p_{bg}, squared loss $\ell(\hat{y}, y) = (\hat{y} - y)^2$, and oracle predictor $f = Y$. For every active pair $\{i, j\} \subset \{1, 2, 3\}$, faithful interaction indices evaluated at order two strictly vanish (0), erasing the cooperative signal. Conversely, projective indices truncated at order two produce instance-dependent values in $\{-1/4, +1/4\}$ whose sign inverts with input parity. In contrast, the interventional decomposition of Equations 3 and 4 invariantly isolates the cooperative effect into the synergistic channel, yielding $(U_i, R_i, S_i) = (0, 0, 1/2)$ for each active feature across all instances (Proof in Appendix B).*

Remark. ***Algebraic mechanism and representation theory context.*** *Full-order STII ($k = 3$) captures the parity triplet by construction (Sundararajan et al., 2020). However, practical XAI relies on order-2 truncations, interpreting positive signs as synergy and negative signs as redundancy. The failure formalized in Proposition 1 is algebraic: because marginalizing any XOR parent yields constant expectation $1/2$, all order-1 and order-2 Möbius coefficients identically vanish ($m(\{i\}) = 0, m(\{i, j\}) = 0$). Faithful indices (preserving Möbius order purity) must output 0. Conversely, projective indices (e.g., standard SII) distribute the pure order-3 coefficient $m(\{1, 2, 3\}) = \pm 1/2$ across lower-order pairs via Grabisch–Roubens weights (Grabisch & Roubens, 1999), yielding $\pm 1/4$. Thus, identical symmetric cooperative mechanisms are arbitrarily misdiagnosed as redundant depending on local instance parity. Following canonical practice in representation theory, the 3-way parity SCM serves as the minimal counterexample of pure higher-order synergy with vanishing lower-order Möbius mass; extending this analysis to continuous, mixed, and deep neural architectures (where lower-order mass is non-zero) is empirically validated in Section 4.*

3.2 Coalition Vocabulary and Interventional Monte Carlo Estimation

Evaluating all 2^n coalitions is intractable for large n. InDE-X maintains an online coalition vocabulary $\mathcal{V} : S \mapsto (\widehat{L}(S), \widehat{\sigma}^2(S), K(S))$ tracking the running empirical mean loss $\widehat{L}(S)$, unbiased sample variance $\widehat{\sigma}^2(S)$, and visit count $K(S)$, updated in streaming passes via Welford's algorithm (Welford, 1962). For a visited coalition S, the empirical loss estimator using $K(S)$ Monte Carlo background samples is:

$$\widehat{L}(S) = \frac{1}{K(S)} \sum_{k=1}^{K(S)} \ell\left(f(m_S \odot x + (1 - m_S) \odot Z^{(k)}), y\right). \tag{5}$$

Non-converged entries are explored via an ε-greedy softmin schedule: with probability ε, a coalition is drawn uniformly from $\{0, 1\}^n$, guaranteeing that every coalition has positive exploration mass $p(S) \geq \varepsilon 2^{-n} > 0$ so that the sample count condition $K(S) \geq K$ of Theorem 3 (Section 3.4) is met in finite time almost surely. With probability $1 - \varepsilon$, a coalition is drawn from the non-converged set $\mathcal{V}_{\text{open}}$ with probability proportional to $\exp(-\beta K(S))$ with temperature $\beta > 0$. Sampled coalitions are dynamically expanded into diamond quadruples (Section 3.3). A coalition is marked converged once its 95% Student-t half-width $w(S) \coloneqq t_{\alpha/2, K(S)-1} \widehat{\sigma}(S)/\sqrt{K(S)}$ drops below threshold w^*. Amortizing vocabularies across instances introduces cross-input bias; we therefore maintain dedicated per-input vocabularies.

Theorem 1. ***Interventional Monte Carlo Bound.*** *Assume bounded loss $\ell \in [\ell_{\min}, \ell_{\max}]$ with range $B = \ell_{\max} - \ell_{\min}$. For any coalition S, the empirical mean $\widehat{L}(S)$ in Equation 5 over $K(S)$ iid background samples is unbiased ($\mathbb{E}[\widehat{L}(S)] = L(S)$) and satisfies:*

$$\text{Var}(\widehat{L}(S)) = \frac{\sigma^2(S)}{K(S)}, \qquad \Pr(|\widehat{L}(S) - L(S)| > \epsilon) \leq 2 \exp\left(-\frac{2K(S)\epsilon^2}{B^2}\right), \tag{6}$$

where $\sigma^2(S) \coloneqq \text{Var}_{Z \sim p_{\text{bg}}}(\ell(f(m_S \odot x + (1 - m_S) \odot Z), y))$ is the coalition loss variance, and the probability $\Pr$ is taken over the product measure of the $K(S)$ background draws.

Loss boundedness holds by construction in tabular regression. For vision and language models, we estimate the effective range B using empirical quantiles with data-justified clipping (Section 4), complemented by empirical-Bernstein bounds (Maurer & Pontil, 2009) (Appendix C).

3.3 Coupled Diamond Sampling and Strict Variance Reduction

Pairwise interactions depend on the second-order mixed difference for context $C \subseteq [n] \setminus \{i, j\}$:

$$\Delta_{ij}(C) = L(C \cup \{i\}) + L(C \cup \{j\}) - L(C \cup \{i, j\}) - L(C). \tag{7}$$

Estimating these four terms with independent background draws incurs high variance. To mitigate this, INDE-X employs *diamond sampling* (named for the diamond shape these four subsets form in the Boolean lattice). We evaluate $\{C, C \cup \{i\}, C \cup \{j\}, C \cup \{i, j\}\}$ simultaneously using the *exact same* background batch $\{Z^{(k)}\}_{k=1}^{K}$; by coupling the noise across adjacent evaluations, the positive covariance cancels out estimation variance.

Theorem 2. ***Diamond Variance Reduction.*** *Let context C be drawn from a distribution over $2^{[n]\setminus\{i,j\}}$ and background $Z \sim p_{\text{bg}}$, with joint covariances over (C, Z). Let (a, b, c, d) denote losses at $(C \cup \{i, j\}, C \cup \{i\}, C \cup \{j\}, C)$ on shared (C, Z). We define the adjacent covariances Σ_{adj}, between coalitions differing by one feature, and diagonal covariances Σ_{diag}:*

$$\Sigma_{\text{adj}} \coloneqq \text{Cov}(a, b) + \text{Cov}(c, d) + \text{Cov}(a, c) + \text{Cov}(b, d), \qquad \Sigma_{\text{diag}} \coloneqq \text{Cov}(a, d) + \text{Cov}(b, c).$$

Let $\widehat{\Delta}_{ij}^{\text{diam}}$ denote the diamond estimator averaging across M sampled contexts on coupled background batches, and let $\widehat{\Delta}_{ij}^{\text{unif}}$ denote the independent estimator evaluating each of the four coalitions with M independent context and background draws. Letting $V_u \coloneqq \text{Var}(u)$ for $u \in \{a, b, c, d\}$, under the Adjacency-Dominant Covariance condition $\Sigma_{\text{adj}} > \Sigma_{\text{diag}}$ (see Assumption 1 in Appendix D), the diamond estimator achieves strictly lower variance:

$$\text{Var}(\widehat{\Delta}_{ij}^{\text{diam}}) = \text{Var}(\widehat{\Delta}_{ij}^{\text{unif}}) - \frac{2}{M}(\Sigma_{\text{adj}} - \Sigma_{\text{diag}}) < \text{Var}(\widehat{\Delta}_{ij}^{\text{unif}}), \tag{8}$$

yielding variance ratio $\frac{\text{Var}(\widehat{\Delta}_{ij}^{\text{unif}})}{\text{Var}(\widehat{\Delta}_{ij}^{\text{diam}})} = 1 + \frac{2(\Sigma_{\text{adj}} - \Sigma_{\text{diag}})}{(V_a + V_b + V_c + V_d) - 2(\Sigma_{\text{adj}} - \Sigma_{\text{diag}})} > 1.$

The variance reduction scales monotonically with the dominance gap $(\Sigma_{\text{adj}} - \Sigma_{\text{diag}})$. In deep networks, adjacent coalitions (differing by a single feature) produce strongly correlated loss responses ($\Sigma_{\text{adj}} \gg 0$), whereas diagonal coalitions (differing by two features) decorrelate rapidly ($\Sigma_{\text{diag}} \approx 0$). While pure parity violates this condition due to bit-flip decorrelation (serving as an analytical boundary test; Appendix F), Assumption 1 holds robustly across all empirical neural network benchmarks in Section 4, yielding variance reductions exceeding $10^3\times$.

Pair-level interaction intensities. We define pair-level synergy S_{ij} and redundancy R_{ij} as the non-negative extrema of mixed differences across sampled contexts:

$$S_{ij} = \max_{C \subseteq [n]\setminus\{i,j\}} [\,\Delta_{ij}(C)\,]_+, \qquad R_{ij} = \max_{C \subseteq [n]\setminus\{i,j\}} [\,-\Delta_{ij}(C)\,]_+, \tag{9}$$

where $[z]_+ \coloneqq \max(z, 0)$. Because extrema are evaluated across different contexts, a feature pair can exhibit synergy in cooperative regimes ($S_{ij} > 0$) and redundancy in substitution regimes ($R_{ij} > 0$) exceeding estimation variance in the empirical analysis (Appendix H.3). Theorems 1 and 3 operate on coalition losses $L(S)$, while Theorem 2 operates on the signed discrete derivative $\Delta_{ij}(C)$; the rectification $[\cdot]_+$ in Equation 9 is merely an operational convention for reporting intensities.

3.4 Budgeted Complexity and Uniform Convergence

Unlike retraining methods scaling as $\mathcal{O}(2^n \cdot N_{\text{train}} \cdot |\theta|)$ in weights and computation, INDE-X audits deployed models post-hoc under two explicit operational regimes:

1. **Exhaustive regime ($n \leq 12$):** Evaluates the full 2^n lattice with $\mathcal{O}(K \cdot 2^n)$ forward passes, yielding exact population profiles without lattice truncation bias. In experiment E3 ($n = 10$ attention heads in GPT-2 IOI), exhaustive evaluation finishes in 21 CPU minutes with zero lattice bias.
2. **Budgeted regime ($n \gg 12$):** Forward evaluations are strictly user-bounded by $B_{\text{eval}} = 4TK_{\text{batch}}$ via coupled diamonds. Phase 2 LOCO extraction iterates exclusively over the set of valid sampled contexts $\mathcal{C}_{\mathcal{V}}(i) \coloneqq \{C \subseteq [n] \setminus \{i\} \mid C \in \mathcal{V},\, C \cup \{i\} \in \mathcal{V}\}$ in $\mathcal{O}(|\mathcal{C}_{\mathcal{V}}(i)|) \leq \mathcal{O}(4T)$ CPU lookups. Evaluating over a partial vocabulary $\mathcal{V} \subset \mathcal{P}([n])$ with baseline and singletons evaluated by default ($\{\emptyset\} \cup \{\{i\}\}_{i=1}^{n} \subset \mathcal{V}$) defines the sub-lattice population quantities:

$$U_i^{\mathcal{V}} \coloneqq \min_{C \in \mathcal{C}_{\mathcal{V}}(i)} \text{LOCO}(i \mid C) \geq U_i, \quad S_i^{\mathcal{V}} \coloneqq \max_{C \in \mathcal{C}_{\mathcal{V}}(i)} \text{LOCO}(i \mid C) - \pi_i \leq S_i, \quad R_i^{\mathcal{V}} \coloneqq \pi_i - U_i^{\mathcal{V}} \leq R_i. \tag{10}$$

Table 2: Theoretical contract of the method: guarantees, formal assumptions, applicability, and boundary/failure regimes for each analytical result.

Result	Requires	Applies to	Fails when
Prop. 1 (Instability)	Fixed p_{bg}, higher-order SCM	Signed pairwise indices	Outside SCM family
Thm 1 (MC bound)	Bounded loss ℓ, iid draws	Coalition loss $\widehat{L}(S)$	Unbounded loss ($B = \infty$)
Thm 2 (Variance)	A1 ($\Sigma_{\mathrm{adj}} > \Sigma_{\mathrm{diag}}$)	Mixed diff. Δ_{ij}	A1 violated (e.g., parity)
Thm 3 (Vocabulary)	Finite $\mathcal{V}$, samples K	Uniform error on $\mathcal{V}$	Infinite/unvisited $\mathcal{V}$
Prop. 2 (Extrema CI)	Uniform error bound $\leq \epsilon$	Sub-lattice $\mathcal{C}_{\mathcal{V}}(i)$	Error $> \epsilon$ (bias on global U_i)

Lattice truncation introduces a unidirectional structural bias: $U_i \leq U_i^{\mathcal{V}}$, $S_i \geq S_i^{\mathcal{V}}$, and $R_i \geq R_i^{\mathcal{V}}$. Similarly, pair-level intensities evaluated over sampled diamond contexts $\mathcal{C}_{\mathcal{V}}(i,j)$ yield conservative lower bounds: $S_{ij}^{\mathcal{V}} \leq S_{ij}$ and $R_{ij}^{\mathcal{V}} \leq R_{ij}$.

Proposition 2. ***Perturbation Stability and Interval Propagation.*** *Assume uniform vocabulary error* $\max_{S\in\mathcal{V}} |\widehat{L}(S) - L(S)| \leq \epsilon$. *For any feature i evaluated over its sampled context set $\mathcal{C}_{\mathcal{V}}(i)$, empirical LOCO gains* $\widehat{G}_i(C) \coloneqq \widehat{L}(C) - \widehat{L}(C \cup \{i\})$ *satisfy* $|\widehat{G}_i(C) - \mathrm{LOCO}(i \mid C)| \leq 2\epsilon$ *uniformly in* $C \in \mathcal{C}_{\mathcal{V}}(i)$. *Deterministic perturbation stability bounds the sub-lattice extrema:*

$$|\widehat{U}_i^{\mathcal{V}} - U_i^{\mathcal{V}}| \leq 2\epsilon, \quad |\widehat{L}_i^{\max,\mathcal{V}} - L_i^{\max,\mathcal{V}}| \leq 2\epsilon, \quad |\widehat{\pi}_i - \pi_i| \leq 2\epsilon, \quad |\widehat{R}_i^{\mathcal{V}} - R_i^{\mathcal{V}}| \leq 4\epsilon, \quad |\widehat{S}_i^{\mathcal{V}} - S_i^{\mathcal{V}}| \leq 4\epsilon. \tag{11}$$

When $\mathcal{C}_{\mathcal{V}}(i)$ covers all 2^{n-1} inclusion contexts (exhaustive regime), sub-lattice quantities coincide with the true global channels ($U_i^{\mathcal{V}} = U_i, S_i^{\mathcal{V}} = S_i, R_i^{\mathcal{V}} = R_i$).

Proposition 2 enables propagating coalition-level 95% Student-t confidence intervals into conservative uncertainty bounds for each feature's U/R/S profile (Appendix E).

Theorem 3. ***Finite-Vocabulary Uniform Convergence.*** *Under the bounded-loss assumption of Theorem 1 ($\ell \in [\ell_{\min}, \ell_{\max}]$, $B \coloneqq \ell_{\max} - \ell_{\min}$), for any finite vocabulary $\mathcal{V} \subseteq \mathcal{P}([n])$, if each $S \in \mathcal{V}$ is evaluated with sample size $K \geq \frac{B^2}{2\epsilon^2} \log \frac{2|\mathcal{V}|}{\alpha}$, then with probability at least $1 - \alpha$:* $\max_{S\in\mathcal{V}} |\widehat{L}(S) - L(S)| \leq \epsilon$.

Directional conservatism and theoretical guarantees under sampling. Combining the lattice monotonicity of Equation 10 with the concentration of Proposition 2 yields the directional bounds:

$$U_i \leq U_i^{\mathcal{V}} \leq \widehat{U}_i^{\mathcal{V}} + 2\epsilon, \qquad S_i \geq S_i^{\mathcal{V}} \geq \widehat{S}_i^{\mathcal{V}} - 4\epsilon, \qquad R_i \geq R_i^{\mathcal{V}} \geq \widehat{R}_i^{\mathcal{V}} - 4\epsilon. \tag{12}$$

Thus, $\widehat{U}_i^{\mathcal{V}} + 2\epsilon$ serves as a conservative upper bound on true uniqueness, while $\widehat{S}_i^{\mathcal{V}} - 4\epsilon$ and $\widehat{R}_i^{\mathcal{V}} - 4\epsilon$ serve as conservative lower bounds on true synergy and redundancy. Concretely, a positive $\widehat{U}_i^{\mathcal{V}}$ under partial budgets does not guarantee true uniqueness, since the true minimizing context may remain unvisited (as in ChestX-ray14 experiments with $n = 196$); observing $\widehat{U}_i^{\mathcal{V}} + 2\epsilon < 0$ (or a statistically significant negative value via Welford confidence intervals), however, definitively certifies an antagonistic suppressor. Streaming Welford updates with bitset keys bound host RAM to < 1 MB even for $n = 196$ (Table 21, Appendix I). Table 2 summarizes the formal requirements, scope of applicability, and boundary failure regimes for all analytical guarantees.

4 Experiments

Protocol. Following established XAI benchmarking practice (Hooker et al., 2019; Hedström et al., 2023; Agarwal et al., 2022), our evaluation addresses three falsification targets: **E1** (tabular SCMs) evaluates ground-truth cooperative role recovery and benchmarks scalar baseline failure on trained networks (Proposition 1); **E2** (NIH ChestX-ray14) audits multi-region patch cooperation under spatial localization vs. causal deletion; and **E3** (GPT-2 IOI) evaluates higher-order circuit discovery and downstream pruning beyond first-order patching. Loss boundedness (Theorem 1) holds by construction in E1 and is operationalized via empirical quantiles with data-supported clipping safeguards in E2–E3. We report mean $\pm$ std across 5 random seeds, paired-Wilcoxon p-values (Wilcoxon, 1945) with Bonferroni correction (Bonferroni, 1936), and Hodges-Lehmann effect sizes (Hodges & Lehmann, 2012).

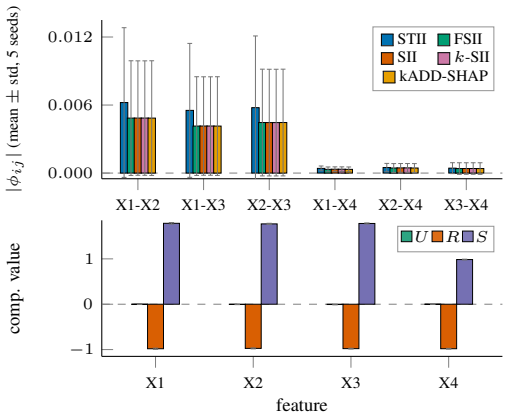


Figure 1: **Top:** Pair-level scalar indices on the 3-way XOR SCM. **Bottom:** Per-feature U/R/S profile from INDE-X (mean ± std, 5 seeds): active features are synergy-dominated; spectator is isolated.

Table 3: E1 disambiguation and recovery. **Top:** conflation ratios on XOR SCMs show scalar baselines conflating mechanisms vs. INDE-X. **Bottom:** on Synth3, INDE-X accurately recovers uniqueness and pair synergy.

Baseline	$\oplus$**3**	$\oplus$**+**$\wedge$
STII	319.51×	102.55×
Faith-SHAP	411.88×	110.68×
SII	411.14×	110.66×
k-SII	411.14×	110.66×
kADD-SHAP	411.88×	110.68×

Target	**Pearson** r	**Spearman** ρ
U	$0.978_{\pm 0.02}$	$0.577_{\pm 0.0}$
S_{pair}	$0.912_{\pm 0.06}$	$0.548_{\pm 0.1}$

4.1 TABULAR SCMS: SCALAR PAIR REPORTS CONFLATE MECHANISMS

We test Proposition 1 on trained two-layer MLPs across three structural causal models: a 3-way XOR SCM (three synergistic features, one spectator), an XOR+AND SCM, and an 8-feature continuous dataset (Synth3) with ground-truth U/R/S structure (Appendix H). Baselines are canonical Shapley interaction indices (STII, Faith-SHAP, SII, k-SII, kADD-SHAP).

Empirical confirmation of Proposition 1. Figure 1 and Table 3 confirm our theoretical prediction on trained MLPs. On XOR3, faithful baselines (STII, Faith-SHAP) vanish ($|\phi_{ij}| \leq 0.006$), while projective indices (SII, k-SII, kADD-SHAP) scatter the triplet into spurious signed coefficients ($\pm 1/4$). Conversely, INDE-X concentrates over 98% of active capacity in the synergy coordinate ($S = 0.50 \pm 0.01$) while setting spectator $X_4 \approx 0$. By projecting higher-order synergy into scalar pairs, baselines underestimate the true mechanism magnitude by up to $411.88\times$ (see Table 3).

Continuous recovery and higher-order baselines. This limitation extends beyond parity. On continuous Synth3 (containing ground-truth unique, redundant, and synergistic roles; Appendix H), INDE-X accurately recovers feature uniqueness (Pearson $r = 0.978 \pm 0.022$) and pairwise synergy (Pearson $r = 0.912 \pm 0.058$). Moreover, higher-order scalar methods (T-NID (Lerman et al., 2021), Higher-Order IG (Janizek et al., 2021), H-Sets (Mehrotra et al., 2026)) inherit Proposition 1's bottleneck: because they collapse set interactions into single signed scalars, all exhibit severe conflation on XOR3 ($\rho_m \in [14.4\times, 108.0\times]$, $p < 0.005$; Table 9 in Appendix H.1). Widening set sizes without U/R/S channels cannot resolve cooperation.

Robustness and error attribution. This separation reflects true causal routing rather than MLP approximation or Monte Carlo error (Theorem 1): synergy dominates across all 5 seeds ($S > 95\%$) with non-overlapping confidence intervals. Furthermore, role rankings are robust to background choice p_{bg}, with median relative drift $< 10\%$ under empirical resampling ($|\Delta U| = 0.023, |\Delta S| = 0.105$; Appendix H.5). Finally, diamond sampling (Assumption 1) holds on Synth3, delivering a $3.58\times$ empirical variance reduction (Appendix F).

4.2 CLINICAL EVALUATION: DISCOVERING MULTI-REGION CO-DEPENDENCIES

Clinical diagnosis frequently hinges on *multi-region co-dependencies*, such as assessing the cardiothoracic ratio in Cardiomegaly across opposing lateral cardiac borders. Standard scalar heatmaps (e.g., GradCAM (Selvaraju et al., 2017) or pixel gradients) collapse these distributed landmarks into a single diffuse heatmap. We evaluate INDE-X on NIH ChestX-ray14 ($n = 880$ images, 984 bounding boxes) (Wang et al., 2017) with DenseNet121 (Huang et al., 2017) using a 14×14 patch

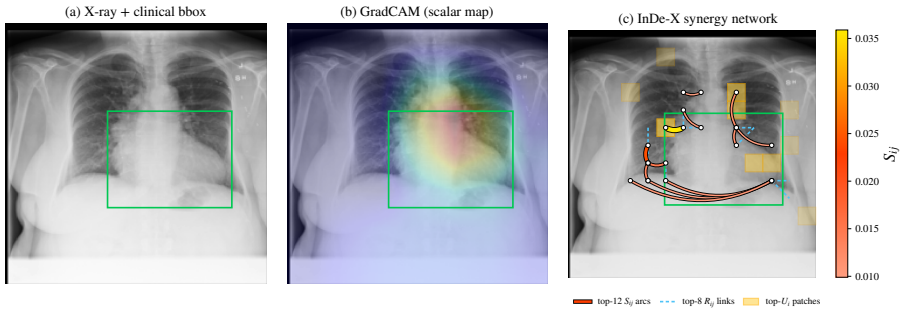


Figure 2: Interaction network on NIH ChestX-ray14 (Cardiomegaly). (a) Radiograph with clinical ground-truth bounding box. While GradCAM (b) collapses into a diffuse central blob over the sternum, INDE-X (c) resolves distal bilateral synergy arcs (S_{ij}, red-yellow, $d \geq 6$) bridging opposing cardiac borders alongside local contour redundancy (R_{ij}, cyan).

grid (196 patches). To break circularity in deletion evaluation, removed patches are imputed with an independent 11×11 Gaussian blur baseline rather than mean or zero ablation.

Bilateral synergy across anatomical landmarks. While redundancy clusters locally along tissue contours (43.1% at distance $d \leq 2$), nearly one-third (32.7%) of top synergistic pairs span distal regions ($d \geq 4$), crossing the thoracic midline (spatial analysis in Appendix H). In Cardiomegaly (Figure 2), while GradCAM peaks uninformatively over the central sternum, INDE-X uncovers distal synergy arcs ($d \in \{6.0, 8.0\}$) bridging opposing cardiac margins, directly capturing the bilateral landmarks required for clinical cardiothoracic assessment.

Causal deletion faithfulness vs. spatial localization. We benchmarked spatial localization (Pointing Game, IoU) alongside interventional deletion across all 880 images (full baseline comparison in Appendix H). Crucially, edge-tracing heuristics do not imply causal fidelity: while high-resolution pixel gradients identify localized visual contours, INDE-X achieves significantly superior Deletion AUC under independent Gaussian blur (0.269 vs. 0.332 for GradCAM, $\Delta = -0.063, p = 1.61 \times 10^{-78}$). This confirms that INDE-X faithfully audits the distributed, multi-region causal pathways driving diagnostic classification rather than merely detecting high-frequency edges.

4.3 TRANSFORMER CIRCUIT: FIRST-ORDER PATCHING MISSES SYNERGISTIC HEADS

We evaluate INDE-X on the Indirect Object Identification (IOI) circuit in GPT-2 small (Wang et al., 2023). We audit 10 attention heads spanning five functional roles (Name-Mover, S-Inhibition, Induction, Duplicate-Token, Previous-Token) by enumerating the full $2^{10} = 1{,}024$ coalition lattice under mean ablation across 5 prompt seeds (26-head circuit extension in Appendix H).

Mechanistic discovery of latent synergy. As a foundational consistency check, standalone attribution $\pi(h) = v(\emptyset) - v(\{h\})$ matches standard single-head activation patching algebraically (Spearman$(\pi, \text{patch}) \equiv 1.000$), ensuring strict compatibility with first-order baselines. The pairwise interaction map (Figure 3a) separates substitution blocks (Duplicate-Token + Induction, $\bar{R} = 0.226$) from cooperative chains (Induction + S-Inhibition, $\bar{S} = 0.383$). In the high-synergy, low-singleton quadrant (Figure 3b), heads 10.0, 9.9, and 4.11 exhibit negligible standalone impact ($\pi \leq 0$) despite exerting substantial pair-level synergy ($S > 0.45$). Scalar patching overlooks these critical circuit components because their function is strictly cooperative.

Actionable causal validation: double knockout and circuit pruning. To demonstrate that U/R/S profiles provide actionable causal guidance, we conduct two downstream interventional experiments evaluating joint model dynamics. ***First***, we test whether predicted interaction $S_{ij} - R_{ij}$ governs empirical degradation under *double head knockout*. Defining knockout excess as $\Gamma_{ij} := \Delta_{ij} - (\Delta_i + \Delta_j)$, where Δ represents logit-difference degradation from the full model, Γ_{ij} correlates exceptionally strongly with predicted interaction across all 45 pairs: Pearson $r = -0.893$ ($p < 10^{-15}$, Figure 3c). This negative sign reflects exact causal mechanics: synergistic pairs ($S \gg 0$) act as logical AND-gates

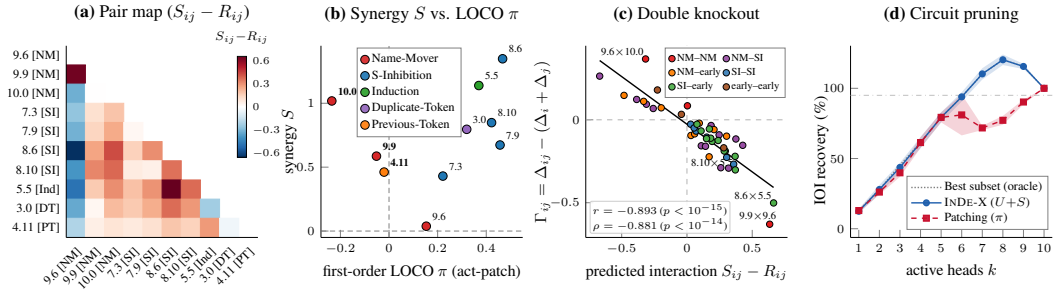


Figure 3: Auditing and downstream causal validation on the GPT-2 IOI circuit. **(a)** Pairwise interaction map $S_{ij} - R_{ij}$. **(b)** Synergy S vs. standalone LOCO π: heads with near-zero singleton impact ($\pi \leq 0$) carry substantial synergy. **(c)** Double knockout (45 pairs): predicted interaction tracks empirical knockout excess Γ_{ij} ($r = -0.893, p < 10^{-15}$). **(d)** Circuit pruning: INDE-X ($U + S$ ranking) matches the theoretical Oracle upper bound and outperforms activation patching (π).

that saturate under knockout ($\Gamma_{ij} < 0$), as an individual knockout already disables the joint circuit; conversely, redundant backup pairs ($R \gg 0$) exhibit super-additive collapse ($\Gamma_{ij} > 0$) as both heads must be knocked out to overcome mutual compensation. Symmetrically, mirror-image knock-in from the empty model is strongly super-additive for synergistic pairs ($r = +0.700, p = 8.5 \times 10^{-8}$). ***Second***, we evaluate *circuit pruning* by progressively retaining top-k heads ranked by activation patching $\pi(h)$ versus INDE-X cooperative ranking ($U + S$). While activation patching degrades sharply ($81.0\% \pm 14.6\%$ recovery at $k = 6$, dropping to $77.2\% \pm 3.2\%$ at $k = 8$ due to omitting synergistic heads 10.0 and 9.9), INDE-X achieves $93.9\% \pm 3.3\%$ at $k = 6$ and reaches $120.3\% \pm 3.8\%$ at $k = 8$ (see Figure 3d and Table 17 in Appendix H.3). Recovery exceeds 100% because INDE-X successfully prunes the antagonistic suppressor head 9.6 ($U = -1.214$), matching the combinatorial Oracle upper bound across $k \in \{4, 6, 8\}$.

5 Discussion and Conclusion

The prevailing paradigm in feature interaction analysis fundamentally fails when cooperation extends beyond two variables: projecting higher-order dynamics down to signed pairwise scalars arbitrarily conflates synergy with redundancy, erasing genuine cooperative mechanisms or inverting their signs across instances. INDE-X overcomes this representational dilemma by operationalizing the Partial Information Decomposition triad—Uniqueness, Redundancy, and Synergy—for post-hoc auditing of deployed neural networks without combinatorial retraining. By measuring marginal loss variations under interventional baseline masking over the Boolean lattice and strictly bounding estimation variance via coupled diamond sampling, our framework replaces heuristic scalar summaries with a structured, causally grounded characterization of a model's true internal dependencies.

Operational boundaries and real-world impact. This methodology introduces explicit trade-offs that govern its practical deployment. Because the estimand is interventional, attribution profiles are inherently conditioned on the chosen reference background distribution, and budgeted exploration over high-dimensional lattices necessarily yields conservative directional bounds rather than exact global channels. Nevertheless, these theoretical guarantees provide critical diagnostic value in safety-critical domains: they definitively certify true antagonistic suppressors, resolve distal multi-region anatomical co-dependencies that diffuse attribution maps collapse, and expose latent cooperative circuits in transformer architectures that remain invisible to standard first-order activation patching.

Conclusion: toward multi-channel mechanistic cartography. Faithfully auditing modern deep networks requires abandoning the assumption that complex feature cooperation can be compressed into a single scalar score. By unifying interventional leave-one-out loss with finite-sample statistical guarantees, INDE-X demonstrates that disentangling uniqueness, redundancy, and synergy is not merely an optional refinement, but a necessary foundation for faithful model auditing. Transitioning from one-dimensional feature attribution to multi-channel interventional cartography marks a foundational step toward rigorous, verifiable alignment and trustworthy artificial intelligence.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools only for grammar and spelling checks. We have not used generative AI tools for conceiving core research hypotheses, designing the theoretical framework, deriving mathematical proofs, or analyzing experimental data, and all other tasks with required disclosure are not applicable to this work. All algorithmic pipelines and evaluation scripts were independently written, audited, and tested for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work focuses on foundational algorithmic and theoretical methodology for auditing feature interactions in deployed deep neural networks, with the broader objective of enhancing algorithmic transparency and safety. Our research does not involve human participant recruitment, the collection of private personal data, or actionable bio-behavioral interventions. All empirical evaluations are conducted strictly on publicly accessible, fully de-identified open-source benchmarks (including NIH ChestX-ray14) and standard public language model tasks, adhering fully to their respective release terms and institutional data guidelines. As a post-hoc explanatory framework, our method presents no foreseeable direct negative societal, environmental, or security risks, and we anticipate that providing mathematically rigorous tools for higher-order model auditing will contribute positively to verifiable AI alignment and accountability.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

To guarantee complete mathematical and empirical reproducibility, we provide comprehensive theoretical proofs, algorithmic specifications, experimental testbeds, and computational environments via open-source code. Full mathematical proofs for all analytical claims, including the parity instability of pairwise truncations, interventional semantics, and non-asymptotic concentration bounds, diamond variance reduction under Adjacency Dominance, and finite-vocabulary uniform convergence with perturbation stability, are rigorously established in Appendices B, C, D, and E. The end-to-end operational engine, encompassing streaming Welford moment accumulation, ε-greedy softmin exploration, and localized Phase 2 Boolean lattice extraction, is formally detailed as complete pseudocode in Appendix G.6. All experimental testbeds, synthetic data generation procedures (3-way XOR, XOR+AND, and the 8-feature continuous Synth3), clinical patch discretizations for NIH ChestX-ray14, and prompt configurations for GPT-2 IOI circuits are thoroughly documented in Appendix H, alongside systematic background-distribution (p_{bg}) sensitivity analyses and spatial grid evaluations. All experiments were conducted across 5 fixed random seeds in standardized compute environments (NVIDIA RTX A6000 and RTX 3090) alongside formal time complexities, empirical wall-clock runtimes, and host RAM verification confirming a < 1 MB streaming footprint. Finally, an anonymized, self-contained reproduction repository containing complete source code, evaluation harnesses, synthetic benchmark generators, and pre-computed coalition vocabularies is included in the supplementary material and will be fully open-sourced upon publication.

REFERENCES

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/65398a0eba88c9b4a1c38ae405b125ef-Abstract-Datasets_and_Benchmarks.html.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. doi: 10.3390/E16042161. URL https://doi.org/10.3390/e16042161.

C.E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber, 1936. URL https://books.google.it/books?id=3CY-HQAACAAJ.

Ngoc Bui, Hieu Trung Nguyen, Viet Anh Nguyen, and Rex Ying. Explaining graph neural networks via structure-aware interaction index. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pp. 4854–4883. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/bui24b.html.

Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Comput. Oper. Res.*, 36(5):1726–1730, 2009. doi: 10.1016/J.COR.2008.04.004. URL https://doi.org/10.1016/j.cor.2008.04.004.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Ian Covert, Scott M. Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c7bf0b7c1a86d5eb3be2c722cf2cf746-Abstract.html.

Shaurya Dewan, Rushikesh Zawar, Prakanshul Saxena, Yingshan Chang, Andrew Luo, and Yonatan Bisk. Diffusionpid: interpreting diffusion via partial information decomposition. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11423–11434, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/95e1533eb1b20a97777749fb94fdb944-Abstract.html.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 3449–3457. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.371. URL https://doi.org/10.1109/ICCV.2017.371.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *Int. J. Game Theory*, 28(4):547–565, 1999. doi: 10.1007/S001820050125. URL https://doi.org/10.1007/s001820050125.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *J. Mach. Learn. Res.*, 24:34:1–34:11, 2023. URL https://jmlr.org/papers/v24/22-0142.html.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html.

J. L. Hodges and E. L. Lehmann. *Estimates of Location Based on Rank Tests*, pp. 287–300. Springer US, Boston, MA, 2012. ISBN 978-1-4614-1412-4. doi: 10.1007/978-1-4614-1412-4_25. URL https://doi.org/10.1007/978-1-4614-1412-4_25.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. ISSN 00034851. URL http://www.jstor.org/stable/2235637.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459, 1537274X.

Giles Hooker. Discovering additive structure in black box functions. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel (eds.), *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pp. 575–580. ACM, 2004. doi: 10.1145/1014052.1014122. URL https://doi.org/10.1145/1014052.1014122.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 9734–9745, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2261–2269. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.243. URL https://doi.org/10.1109/CVPR.2017.243.

Ryan G. James, Jeffrey Emenheiser, and James P. Crutchfield. Unique information and secret key agreement. *Entropy*, 21(1):12, 2019. doi: 10.3390/E21010012. URL https://doi.org/10.3390/e21010012.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *J. Mach. Learn. Res.*, 22:104:1–104:54, 2021. URL https://jmlr.org/papers/v22/20-1223.html.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable AI: A causal problem. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, Proceedings of Machine Learning Research, pp. 2907–2916. PMLR, 2020. URL http://proceedings.mlr.press/v108/janzing20a.html.

Chaeyun Ko. STRIDE: subset-free functional decomposition for XAI in tabular settings. *CoRR*, abs/2509.09070, 2025. doi: 10.48550/ARXIV.2509.09070. URL https://doi.org/10.48550/arXiv.2509.09070.

Artemy Kolchinsky. A novel approach to the partial information decomposition. *Entropy*, 24(3):403, 2022. doi: 10.3390/E24030403. URL https://doi.org/10.3390/e24030403.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://doi.org/10.1080/01621459.2017.1307116.

Samuel Lerman, Charles Venuto, Henry A. Kautz, and Chenliang Xu. Explaining local, global, and higher-order interactions in deep learning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 1204–1213. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00126. URL https://doi.org/10.1109/ICCV48922.2021.00126.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL http://www.cs.mcgill.ca/%7Ecolt2009/papers/012.pdf#page=1.

Ayushi Mehrotra, Dipkamal Bhusal, Michael Clifford, and Nidhi Rastogi. H-sets: Hessian-guided discovery of set-level feature interactions in image classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17067–17076, June 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. shapiq: Shapley interactions for machine learning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/eb3a9313405e2d4175a5a3cfcd49999b-Abstract-Datasets_and_Benchmarks_Track.html.

Roger B. Myerson. Graphs and cooperation in games. *Math. Oper. Res.*, 2(3):225–229, 1977. doi: 10.1287/MOOR.2.3.225. URL https://doi.org/10.1287/moor.2.3.225.

Marlis Ontivero-Ortega, Luca Faes, Jesus M. Cortes, Daniele Marinazzo, and Sebastiano Stramaglia. Assessing high-order effects in feature importance via predictability decomposition. *Phys. Rev. E*, 111:L033301, Mar 2025. doi: 10.1103/PhysRevE.111.L033301. URL https://link.aps.org/doi/10.1103/PhysRevE.111.L033301.

Ari Pakman, Amin Nejatbakhsh, Dar Gilboa, Abdullah Makkeh, Luca Mazzucato, Michael Wibral, and Elad Schneidman. Estimating the unique information of continuous variables. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 20295–20307, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a9a1d5317a33ae8cef33961c34144f84-Abstract.html.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, pp. 151. BMVA Press, 2018. URL http://bmvc2018.org/contents/papers/1064.pdf.

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 *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 618–626. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.74. URL https://doi.org/10.1109/ICCV.2017.74.

I.M Sobol'. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1):271–280, 2001. ISSN 0378-4754. doi: https://doi.org/10.1016/S0378-4754(00)00270-6. The Second IMACS Seminar on Monte Carlo Methods.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pp. 9269–9278. PMLR, 2020. URL http://proceedings.mlr.press/v119/sundararajan20b.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, pp. 3319–3328. PMLR, 2017. URL http://proceedings.mlr.press/v70/sundararajan17a.html.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pp. 9259–9268. PMLR, 2020. URL http://proceedings.mlr.press/v119/sundararajan20a.html.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-shap: The faithful shapley interaction index. *J. Mach. Learn. Res.*, 24:94:1–94:42, 2023. URL https://jmlr.org/papers/v24/22-0202.html.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? interpretable attribution for feature interactions. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/443dec3062d0286986e21dc0631734c9-Abstract.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Investigating gender bias in language models using causal mediation analysis. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 3462–3471. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.369. URL https://doi.org/10.1109/CVPR.2017.369.

B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962. doi: 10.1080/00401706.1962.10490022.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 00994987. URL http://www.jstor.org/stable/3001968.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010. URL http://arxiv.org/abs/1004.2515.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, Lecture Notes in Computer Science, pp. 818–833. Springer, 2014. doi: 10.1007/978-3-319-10590-1_53. URL https://doi.org/10.1007/978-3-319-10590-1_53.

APPENDIX

This appendix provides complete theoretical proofs, algorithmic specifications, numerical verification harnesses, extended multi-domain empirical evaluations, and computational complexity analyses for INDE-X (**In**terventional **De**composition for e**X**plainability). To assist the reader in navigating the supplementary material, Table 4 outlines the organizational roadmap across all sections.

Structural Organization. The appendix is organized into three thematic parts:

- **Part I: Theoretical Foundations and Formal Proofs (Appendices A–E):** Appendix A formalizes our mathematical setting on the Boolean lattice $(\mathcal{P}([n]), \subseteq)$ and details the exact duality between loss-space semantics and game-theoretic value functions. Appendices B–E deliver rigorous mathematical proofs for Proposition 1, Theorem 1, Theorem 2, and Theorem 3, along with Proposition 2.
- **Part II: Numerical Verification and Algorithmic Framework (Appendices F–G):** Appendix F provides empirical verification of all theoretical propositions and theorems across controlled synthetic harnesses and real neural models. Appendix G details the lattice-theoretic foundations, the derivation of U/R/S identities, the softmin exploration policy, and the complete algorithmic skeleton.
- **Part III: Extended Empirical Evaluations (Appendices H–I):** Appendix H presents extended experimental results across tabular SCMs (E1), clinical multi-region radiography on NIH ChestX-ray14 (E2), and attention-head circuit discovery on GPT2-SMALL (E3), including the 26-head full circuit extension. Appendix I details formal time and memory complexities, hardware environments.

Table 4: Roadmap of the Supplementary Material: Overview of appendix sections, core technical focus, key theoretical or empirical guarantees, and references to the corresponding main paper sections.

App.	Section Title	Core Focus	Key Guarantees / Primary Results	Main Text
Part I: Theoretical Foundations and Formal Proofs				
A	Mathematical Notation	Formal nomenclature	Symbol table, lattice operators, loss-space semantics	Section 3
B	Proof of Proposition 1	Pairwise index failure	Projective sign-flip ($\pm 1/4$) vs. faithful vanishing	Section 3.1, 4.1
C	Proof of Theorem 1	Interventional semantics	Unbiased Welford update, Hoeffding concentration	Section 3.2
D	Proof of Theorem 2	Variance reduction	Strict gain via diamond sampling under Assumption 1	Section 3.3
E	Proof of Theorem 3	Budgeted convergence	Uniform convergence bound on vocabulary	Section 3.4
Part II: Numerical Verification and Algorithmic Framework				
F	Numerical Verification	Theory validation	Synthetic harnesses, XOR boundary test, Assumption 1 tests	Section 3
G	Extended Methodology	Algorithmic foundations	Boolean vs. Dedekind lattice, Welford, pseudocode	Section 3
Part III: Extended Empirical Evaluations and Computational Reproducibility				
H	Extended Experiments	Multi-domain auditing	E1 (SCMs), E2 (ChestX-ray14), E3 (GPT-2 IOI), 26-head map	Section 4
I	Compute & Reproducibility	Complexity & footprint	Formal complexity, $<$ 1 MB streaming RAM, runtimes	Section 3.4, 4

A Mathematical Notation and Foundations

This appendix provides a rigorous mathematical reference for all sets, spaces, operators, estimators, and asymptotic quantities used throughout the paper. To maximize clarity and prevent ambiguities between loss-space semantics and game-theoretic value functions, we first establish the formal notation in Table 5 and then detail the exact duality mapping.

A.1 Formal Mathematical Setting and Lattice Structure

Let $[n] \coloneqq \{1, 2, \ldots, n\}$ denote the finite ground set of input components (e.g., tabular features, image patches, or Transformer attention heads). The coalition space forms a finite Boolean lattice $(\mathcal{P}([n]), \subseteq)$, where $\mathcal{P}([n])$ is the powerset of $[n]$, partially ordered by subset inclusion $\subseteq$, with join operator $\cup$ (set union), meet operator $\cap$ (set intersection), unique bottom element $\emptyset$, and unique top element $[n]$.

A.2 Loss-Space Mechanics versus Cooperative Game-Theoretic Value Functions

A central conceptual distinction throughout this paper is the interface between *loss space* (used by InDe-X for gradient-free predictive auditing) and *value space* (used in classical cooperative game theory and axiomatic Shapley interaction literature).

Loss Space and Marginal Utility. In loss space, lower values denote superior model accuracy. For any context $C \subseteq [n] \setminus \{i\}$, the Leave-One-Covariate-Out (LOCO) gain measures the drop in loss achieved by including component i:

$$\mathrm{LOCO}(i \mid C) \;\coloneqq\; L(C) - L(C \cup \{i\}). \tag{13}$$

When component i provides an informative predictive signal that reduces empirical loss, $\mathrm{LOCO}(i \mid C) > 0$. If adding i degrades performance (e.g., due to destructive interference, adversarial perturbation, or redundant inhibition), then $\mathrm{LOCO}(i \mid C) < 0$.

Cooperative Value Space and Exact Isomorphism. In contrast, axiomatic cooperative game theory defines a characteristic value function $v : \mathcal{P}([n]) \to \mathbb{R}$ where *higher values denote greater utility* (e.g., accuracy, log-odds, or classification margins). The classical marginal contribution of player i with respect to coalition C is:

$$\Delta_i v(C) \;\coloneqq\; v(C \cup \{i\}) - v(C). \tag{14}$$

The two spaces are connected via the exact linear duality:

$$v(S) \;\equiv\; -L(S) \quad\Longrightarrow\quad \begin{aligned} \Delta_i v(C) \;&=\; -L(C \cup \{i\}) - (-L(C)) \\ &=\; L(C) - L(C \cup \{i\}) \;\equiv\; \mathrm{LOCO}(i \mid C). \end{aligned} \tag{15}$$

Under this canonical duality, all algebraic and order-theoretic properties of the Boolean lattice are strictly preserved.

Second-Order Mixed Differences. For any pair of distinct components $i, j \in [n]$ with $i \neq j$ and any context coalition $C \subseteq [n] \setminus \{i, j\}$, the second-order loss mixed difference is defined as the change in LOCO gain when context is expanded:

$$\begin{aligned} \Delta^{(L)}_{ij}(C) \;&\coloneqq\; \mathrm{LOCO}(i \mid C \cup \{j\}) - \mathrm{LOCO}(i \mid C) \\ &=\; \big(L(C \cup \{j\}) - L(C \cup \{i, j\})\big) - \big(L(C) - L(C \cup \{i\})\big) \\ &=\; L(C \cup \{i\}) + L(C \cup \{j\}) - L(C \cup \{i, j\}) - L(C), \end{aligned} \tag{16}$$

which identically matches the second-order discrete interaction in Section 3.3. In cooperative game-theoretic value space ($v \equiv -L$), the second-order interaction index is:

$$\begin{aligned} \Delta^{(v)}_{ij}(C) \;&\coloneqq\; v(C \cup \{i, j\}) - v(C \cup \{i\}) - v(C \cup \{j\}) + v(C) \\ &=\; -L(C \cup \{i, j\}) - \big(-L(C \cup \{i\})\big) - \big(-L(C \cup \{j\})\big) + \big(-L(C)\big) \\ &=\; L(C \cup \{i\}) + L(C \cup \{j\}) - L(C \cup \{i, j\}) - L(C) \;\equiv\; \Delta^{(L)}_{ij}(C). \end{aligned} \tag{17}$$

Table 5: Comprehensive mathematical notation, domain types, formal definitions, and section references.

Symbol	Domain / Type	Formal Definition and Semantic Meaning	Primary Context
Sets, Coordinates, and Boolean Lattice			
$[n]$	$\{1, \ldots, n\}$	Ground set of input components (features, patches, heads)	Section 3
$(\mathcal{P}([n]), \subseteq)$	Lattice of subsets	Boolean powerset lattice of component coalitions	Section 3, App. G
$S \subseteq [n]$	Subset	Arbitrary component coalition	Section 3
$C \subseteq [n] \setminus \{i\}$	Subset	Context coalition for component i	Section 3
$\mathcal{V} \subseteq \mathcal{P}([n])$	Subset of lattice	Explored coalition vocabulary maintained in memory	Section 3.2, App. E
$\mathcal{E}_{\mathcal{V}}$	Subset of edges	Set of sampled oriented lattice edges $(C, C \cup \{i\})$	App. I
$m_S \in \{0,1\}^n$	Binary vector	Interventional mask: $(m_S)_k = \mathbf{1}[k \in S]$	Section 3
Model, Perturbations, and Loss Functions			
$x \in \mathcal{X}, y \in \mathcal{Y}$	Inputs, targets	Factual observation and associated ground-truth label	Section 3
$f : \mathcal{X} \to \mathcal{Y}$	Neural network	Frozen, pre-trained model under audit	Section 3
p_{bg}	Probability measure	Reference background distribution over baseline values	Section 3, App. C
$Z \sim p_{\mathrm{bg}}$	Random variable	Baseline imputation vector drawn independently from p_{bg}	Section 3
$x^{(S,Z)}$	Perturbed input	Interventional input: $x^{(S,Z)} \coloneqq m_S \odot x + (1 - m_S) \odot Z$	Section 3
$\ell(\hat{y}, y)$	Loss function	Sample predictive loss (e.g., cross-entropy, squared error)	Section 3
$[\ell_{\min}, \ell_{\max}]$	Interval in $\mathbb{R}$	Bounded range of the per-example loss function	App. C
$B \coloneqq \ell_{\max} - \ell_{\min}$	$\mathbb{R}_{>0}$	Maximum loss span across all possible network inputs	App. C
$L(S)$	$\mathbb{R}_{\geq 0}$	Expected coalition loss: $\mathbb{E}_{Z \sim p_{\mathrm{bg}}}[\ell(f(x^{(S,Z)}), y)]$	Section 3
Interventional LOCO and Tripartite U/R/S Decomposition			
$\mathrm{LOCO}(i \mid C)$	$\mathbb{R}$	Marginal gain: $\mathrm{LOCO}(i \mid C) \coloneqq L(C) - L(C \cup \{i\})$	Section 3.1
π_i	$\mathbb{R}$	Singleton baseline score: $\pi_i \coloneqq \mathrm{LOCO}(i \mid \emptyset) = L(\emptyset) - L(\{i\})$	Section 3.1
U_i	$\mathbb{R}$	Unique contribution: $U_i \coloneqq \min_{C \subseteq [n] \setminus \{i\}} \mathrm{LOCO}(i \mid C)$	Section 3.1
$L_i^{\max}$	$\mathbb{R}$	Coalition ceiling: $L_i^{\max} \coloneqq \max_{C \subseteq [n] \setminus \{i\}} \mathrm{LOCO}(i \mid C)$	Section 3.1
R_i	$\mathbb{R}_{\geq 0}$	Redundant contribution: $R_i \coloneqq \pi_i - U_i$	Section 3.1
S_i	$\mathbb{R}_{\geq 0}$	Synergistic contribution: $S_i \coloneqq L_i^{\max} - \pi_i$	Section 3.1
$\Delta_{ij}^{(L)}(C)$	$\mathbb{R}$	Second-order loss mixed difference (see Equation 16)	Section 3.3, App. D
S_{ij}, R_{ij}	$\mathbb{R}_{\geq 0}$	Pairwise synergy and redundancy across components i, j	Section 3
Estimation, Variance, and Sampling Geometry			
$K(S)$	$\mathbb{N}$	Number of Monte Carlo background samples evaluated for S	Section 3.2
$\widehat{L}(S)$	$\mathbb{R}$	Online empirical mean loss estimated via Welford streaming	Section 3.2
$\widehat{\sigma}^2(S)$	$\mathbb{R}_{\geq 0}$	Online empirical sample variance maintained via Welford	Section 3.2
$D(C, i, j)$	4-coalition tuple	Diamond sample: $\{C, C \cup \{i\}, C \cup \{j\}, C \cup \{i, j\}\}$	Section 3.3, App. D
$\Sigma_{\mathrm{adj}}, \Sigma_{\mathrm{diag}}$	$\mathbb{R}$	Adjacency covariance vs. diagonal covariance on diamond corners	Section 3.3, App. D
ρ_v	$\mathbb{R}_{>0}$	Relative variance reduction ratio: $\mathrm{Var}[\widehat{\Delta}_{\mathrm{unif}}] / \mathrm{Var}[\widehat{\Delta}_{\mathrm{diam}}]$	App. D, F
ϵ, δ	$\mathbb{R}_{>0} \times (0,1)$	Target estimation tolerance and failure probability	Section 3, App. C

Table 6: Cross-modal grounding of the Boolean lattice $(\mathcal{P}([n]), \subseteq)$ across the empirical domains evaluated in Section 4.

Audited Domain	**Ground Set** $[n]$	**Reference Distribution** p_{bg}	**Audited Loss Metric** ℓ
Tabular SCMs (E1)	Input feature indices	Marginal empirical / uniform	Squared error $(\hat{y} - y)^2$
ChestX-ray Vision (E2)	Spatial patches (14×14)	11×11 Gaussian blur ($\sigma = 5.0$)	Multi-label binary cross-entropy
GPT-2 Circuits (E3)	Attention heads (L, h)	Corrupted-dataset mean activations	Negative logit difference $-\Delta_{\text{logit}}$

Under the canonical duality $v \equiv -L$, the second-order interaction differences are identically aligned ($\Delta_{ij}^{(v)} \equiv \Delta_{ij}^{(L)}$). When two components act synergistically, adding j enhances the loss reduction provided by i ($\text{LOCO}(i \mid C \cup \{j\}) > \text{LOCO}(i \mid C)$), yielding $\Delta_{ij}^{(L)}(C) > 0$. Conversely, when two components are redundant, the inclusion of j diminishes the marginal gain of i ($\text{LOCO}(i \mid C \cup \{j\}) < \text{LOCO}(i \mid C)$), yielding $\Delta_{ij}^{(L)}(C) < 0$. Consequently, pair-level synergy and redundancy are extracted as the non-negative extrema:

$$S_{ij} \;\coloneqq\; \max_{C} \big[\Delta_{ij}^{(L)}(C)\big]_+, \qquad R_{ij} \;\coloneqq\; \max_{C} \big[-\Delta_{ij}^{(L)}(C)\big]_+. \tag{18}$$

Antagonistic Uniqueness ($U_i < 0$) in Loss Space. A distinctive advantage of framing decomposition directly in loss space is the rigorous interpretation of negative uniqueness. By definition:

$$U_i \;=\; \min_{C \subseteq [n] \setminus \{i\}} \text{LOCO}(i \mid C). \tag{19}$$

If $U_i < 0$, there exists at least one hostile context $C^\star$ such that $\text{LOCO}(i \mid C^\star) < 0$, which implies $L(C^\star \cup \{i\}) > L(C^\star)$. Rather than indicating an estimation artifact, $U_i < 0$ mathematically certifies an *antagonistic suppressor or backup role*: component i actively degrades predictions when primary pathways are intact, but prevents catastrophic collapse when primary features are ablated. This is precisely observed in attention head 9.6 of GPT2-SMALL (Appendix H.3).

A.3 Cross-Modal Operational Grounding: Tabular, Vision, and Language

To ground these definitions across modalities, Table 6 summarizes how input components, baseline perturbations, and loss objectives map onto the unified $(\mathcal{P}([n]), \subseteq)$ lattice.

B Proof of Proposition 1: Instability of Truncated Pairwise Interactions under Parity

Setup. Let $Y = \oplus(X_1, X_2, X_3) \in \{0,1\}$ with $X_1, X_2, X_3, \ldots, X_n$ independent Bernoulli($\frac{1}{2}$) variables. Features $X_4, \ldots, X_n$ are *spectators*: the target Y is statistically and functionally independent of them. The value function at factual instance x is defined under interventional background marginalization as:

$$v_x(S) \;\coloneqq\; \mathbb{E}_{Z\sim p_{\mathrm{bg}}}\big[Y(m_S \odot x + (1-m_S)\odot Z)\big]. \tag{20}$$

Step 1: Value Function under Marginalization. The 3-way parity function possesses the key symmetry that marginalizing over *any single* input among the triplet yields a constant Bernoulli expectation of $1/2$:

$$\mathbb{E}_{X_j}\big[\oplus(x_i, x_k, X_j)\big] \;=\; \tfrac{1}{2} \quad \text{for all } x_i, x_k \in \{0,1\},\ j \in \{1,2,3\}\setminus\{i,k\}. \tag{21}$$

Consequently, if a coalition $S \subseteq [n]$ omits at least one component from the active triplet $\{1,2,3\}$, at least one active input is integrated under the uniform background distribution p_{bg}, and Equation 21 yields $v_x(S) = \frac{1}{2}$. Spectator features contribute nothing, as Y does not depend on $\{X_4, \ldots, X_n\}$. Therefore:

$$v_x(S) \;=\; \begin{cases} \frac{1}{2} & \text{if } \{1,2,3\} \not\subseteq S, \\ \oplus(x_1, x_2, x_3) & \text{if } \{1,2,3\} \subseteq S. \end{cases} \tag{22}$$

Step 2: Möbius Coefficients. Evaluating the canonical Möbius transform $m_x(T) \coloneqq \sum_{W\subseteq T}(-1)^{|T\setminus W|} v_x(W)$ directly from the value assignments in Equation 22:

- **Empty set:** $v_x(\emptyset) = \frac{1}{2} \implies m_x(\emptyset) = \frac{1}{2}$.
- **Singletons:** For any $i \in [n]$, $v_x(\{i\}) = \frac{1}{2} \implies m_x(\{i\}) = \frac{1}{2} - \frac{1}{2} = 0$.
- **Pairs:** For any $\{i,j\} \subseteq [n]$, $v_x(\{i,j\}) = v_x(\{i\}) = v_x(\{j\}) = v_x(\emptyset) = \frac{1}{2} \implies m_x(\{i,j\}) = \frac{1}{2} - \frac{1}{2} - \frac{1}{2} + \frac{1}{2} = 0$.
- **Triplet $\{1,2,3\}$:**

$$\begin{aligned} m_x(\{1,2,3\}) &= v_x(\{1,2,3\}) - v_x(\{1,2\}) - v_x(\{1,3\}) - v_x(\{2,3\}) \\ &\quad + v_x(\{1\}) + v_x(\{2\}) + v_x(\{3\}) - v_x(\emptyset) \\ &= \oplus(x_1,x_2,x_3) - 3\cdot\tfrac{1}{2} + 3\cdot\tfrac{1}{2} - \tfrac{1}{2} \;=\; \oplus(x_1,x_2,x_3) - \tfrac{1}{2}, \end{aligned}$$

  which evaluates precisely to $\pm\frac{1}{2}$ depending on the factual instance parity x.

All higher-order Möbius coefficients either involve spectator features (which vanish identically due to pairwise cancellation) or are strict supersets of $\{1,2,3\}$ (which vanish because adding a spectator does not alter v_x). Thus, the pairwise Möbius coefficient is strictly zero everywhere, and the entire interaction signal concentrates exclusively at order 3.

Step 3: Faithful Pairwise Indices Vanish Identically. The faithful interaction families—including the Shapley-Taylor Interaction Index (STII at order 2; Sundararajan et al. (2020)), Faith-SHAP (FSII at order 2; Tsai et al. (2023)), and k-additive SHAP (kADD-SHAP; Muschalik et al. (2024))—are constructed to restrict their pairwise interaction reports strictly to second-order Möbius terms $m_x(\{i,j\})$, isolating them from higher-order Möbius coefficients. Since $m_x(\{i,j\}) = 0$ by Step 2, the pairwise reports vanish identically:

$$\mathrm{STII}_{ij}(x) \;=\; \mathrm{FSII}_{ij}(x) \;=\; k\text{-}\,\mathrm{SII}_{ij}(x) \;=\; 0 \quad \text{for all } \{i,j\} \subseteq [n]. \tag{23}$$

This establishes the first branch (faithful vanishing) of Proposition 1.

Step 4: Projective Indices Yield Spurious Sign-Flipping Scalars ($\pm\frac{1}{4}$). The classical Shapley Interaction Index is defined by Grabisch & Roubens (1999) as:

$$\mathrm{SII}_{ij}(x) \;\coloneqq\; \sum_{T\subseteq[n]\setminus\{i,j\}} \frac{(n-|T|-2)!\,|T|!}{(n-1)!}\,\Delta_{ij}^{(v)}(T,x), \tag{24}$$

where $\Delta_{ij}^{(v)}(T,x) \coloneqq v_x(T\cup\{i,j\}) - v_x(T\cup\{i\}) - v_x(T\cup\{j\}) + v_x(T)$. For any pair $\{i,j\} \subset \{1,2,3\}$, let k be the unique third active element of the triplet. Then $\Delta_{ij}^{(v)}(T,x) = \oplus(x_1,x_2,x_3) - \frac{1}{2}$ when $k \in T$ (since Equation 22 gives $v_x(T\cup\{i,j\}) = \oplus$ while the remaining three terms equal $\frac{1}{2}$), and $\Delta_{ij}^{(v)}(T,x) = 0$ when $k \notin T$ (as all four terms equal $\frac{1}{2}$). Spectator features do not affect the value of $\Delta_{ij}^{(v)}(T,x)$, entering only as combinatorial summands in the context size $|T|$.

Summing over all subsets $T \subseteq [n]\setminus\{i,j\}$ containing k, and reparameterizing $T = \{k\}\cup T'$ with $T' \subseteq [n]\setminus\{i,j,k\}$ of size $t' \coloneqq |T| - 1 \in \{0,\dots,n-3\}$:

$$\begin{aligned}
\mathrm{SII}_{ij}(x) &= \big(\oplus(x_1,x_2,x_3) - \tfrac{1}{2}\big) \sum_{t'=0}^{n-3} \binom{n-3}{t'} \frac{(n-t'-3)!\,(t'+1)!}{(n-1)!} \\
&= \big(\oplus - \tfrac{1}{2}\big) \sum_{t'=0}^{n-3} \frac{(n-3)!}{t'!\,(n-3-t')!} \cdot \frac{(n-t'-3)!\,(t'+1)!}{(n-1)!} \\
&= \big(\oplus - \tfrac{1}{2}\big) \cdot \frac{(n-3)!}{(n-1)!} \sum_{t'=0}^{n-3} (t'+1) \\
&= \big(\oplus - \tfrac{1}{2}\big) \cdot \frac{1}{(n-1)(n-2)} \cdot \frac{(n-2)(n-1)}{2} \;=\; \tfrac{1}{2}\big(\oplus(x_1,x_2,x_3) - \tfrac{1}{2}\big).
\end{aligned} \tag{25}$$

Because $\oplus(x_1,x_2,x_3) \in \{0,1\}$, the term $\oplus - \frac{1}{2}$ evaluates to $\pm\frac{1}{2}$, so $\mathrm{SII}_{ij}(x)$ equals exactly:

$$\mathrm{SII}_{ij}(x) \;=\; \pm\tfrac{1}{4}. \tag{26}$$

The sign flips deterministically with factual input parity: when the XOR product is 1, the index reports positive interaction ($+1/4$, diagnosed as synergy); when the XOR product is 0, the index reports negative interaction ($-1/4$, diagnosed as redundancy). Exactly the same combinatorial projection governs k-SII at order 2 and kADD-SHAP. For any pair $\{i,j\}$ involving a spectator feature, $\Delta_{ij}^{(v)}(T,x) = 0$ for all contexts, yielding $\mathrm{SII}_{ij}(x) = 0$. This establishes the second branch (projective sign-flipping) of Proposition 1.

Step 5: LOCO-U/R/S Profile Is Invariant and Non-Degenerate. Now consider the expected loss under background marginalization (Equation 1) evaluated with squared error $\ell(\hat{y},y) = (\hat{y}-y)^2$ and oracle predictor $f(x) = Y(x)$:

- $L(C) = \frac{1}{2}$ whenever $\{1,2,3\} \not\subseteq C$: because at least one triplet bit is marginalized under p_{bg}, the oracle prediction is a random Bernoulli$(1/2)$ guess, yielding expected squared loss $0.5\cdot(1-0)^2 + 0.5\cdot(0-1)^2 = 1/2$.
- $L(C) = 0$ whenever $\{1,2,3\} \subseteq C$: all active inputs are present, so the oracle predicts perfectly with zero loss.

Evaluating the LOCO marginal gains for any active component $i \in \{1,2,3\}$:

- In any context C where $|C\cap\{1,2,3\}| \le 1$, adding i leaves at least one active bit missing, so $L(C\cup\{i\}) = L(C) = 1/2 \implies \mathrm{LOCO}(i \mid C) = 0$. Hence, the minimum is achieved:
$$U_i \;=\; \min_{C\subseteq[n]\setminus\{i\}} \mathrm{LOCO}(i \mid C) \;=\; 0. \tag{27}$$
- In the saturated context $C = \{1,2,3\}\setminus\{i\}$, adding i completes the active triplet, dropping the loss from $1/2$ to 0: $\mathrm{LOCO}(i \mid C) = 1/2 - 0 = 1/2$. This attains the coalition ceiling:
$$L_i^{\max} \;=\; \max_{C\subseteq[n]\setminus\{i\}} \mathrm{LOCO}(i \mid C) \;=\; \tfrac{1}{2}. \tag{28}$$
- In the empty context $C = \emptyset$, $\mathrm{LOCO}(i \mid \emptyset) = L(\emptyset) - L(\{i\}) = 1/2 - 1/2 = 0$, so $\pi_i = 0$.

Applying the decomposition identities $R_i = \pi_i - U_i$ and $S_i = L_i^{\max} - \pi_i$ (Equation 4):

$$R_i \;=\; \pi_i - U_i \;=\; 0 - 0 \;=\; 0, \qquad S_i \;=\; L_i^{\max} - \pi_i \;=\; \tfrac{1}{2} - 0 \;=\; \tfrac{1}{2}. \tag{29}$$

Thus, INDE-X yields the invariant, non-degenerate tripartite profile:

$$(U_i, R_i, S_i) \;=\; (0, 0, \tfrac{1}{2}) \quad \text{for all active features } i \in \{1,2,3\}. \tag{30}$$

For spectator features $j \in \{4,\dots,n\}$, $L(C\cup\{j\}) = L(C)$ everywhere, yielding $(U_j, R_j, S_j) = (0,0,0)$.

Step 6: Inevitable Conflation of Scalar Pairwise Truncations. Comparing the two representations establishes the fundamental representation barrier:

1. **Faithful truncation collapses the codomain to zero:** by restricting the report to $m_x(\{i, j\})$, faithful indices yield $\phi_{ij} = 0$, completely blinding the auditor to the active 3-way cooperation.
2. **Projective truncation collapses the codomain to a sign-flipping scalar:** projective indices collapse the 3-way Möbius coefficient into a single pairwise scalar $\pm 1/4$. This scalar conflates distinct causal mechanisms: for input $x = (1, 1, 0)$ it reports $+1/4$ (diagnosing pure synergy), while for input $x = (1, 1, 1)$ it reports $-1/4$ (diagnosing pure redundancy) on the exact same data-generating mechanism.
3. **INDE-X preserves the 3-channel profile:** by retaining the full (U_i, R_i, S_i) channel structure over the lattice, INDE-X correctly identifies that each active feature has zero standalone impact ($U_i = 0$), zero backup redundancy ($R_i = 0$), and maximal cooperative synergy ($S_i = 1/2$), perfectly invariant across all inputs x.

□

C Proof of Theorem 1: Interventional Semantics and Finite-Sample Concentration Guarantees

This section provides the formal proof of Theorem 1, addressing both the causal semantics of interventional background masking (Part A) and the finite-sample concentration guarantees of the Monte Carlo estimator (Part B).

C.1 Semantics of Interventional Masking and Resolution of the OOD Critique

Fix a reference background distribution p_{bg} over baseline imputation values. For any coalition of active components $S \subseteq [n]$ and factual instance (x, y), define the atomic masking intervention:

$$I_S(x, Z) \;:=\; m_S \odot x + (1 - m_S) \odot Z, \qquad Z \sim p_{\mathrm{bg}}. \tag{31}$$

This operation holds retained coordinates $i \in S$ fixed at their observed values x_i, while simultaneously replacing removed coordinates $j \notin S$ with independent draws Z_j from the reference background p_{bg}. The expected coalition loss introduced in Equation 1 of the main text is defined precisely as:

$$L(S; x, y) \;:=\; \mathbb{E}_{Z \sim p_{\mathrm{bg}}}\big[\ell\big(f(I_S(x, Z)), y\big)\big]. \tag{32}$$

Lemma 4. ***Identity for the Masked-Inference Estimator.*** *The Monte Carlo estimator evaluated by* InDe-X *is an unbiased estimator of the interventional masking target $L(S; x, y)$. Furthermore, if a structural causal model (SCM) admits the coordinate-level atomic intervention $\mathrm{do}(X_S = x_S, X_{\bar{S}} = Z_{\bar{S}})$, this target is mathematically identical to the corresponding Pearl do-calculus quantity averaged over $Z \sim p_{\mathrm{bg}}$.*

Proof. The first claim is direct from definition: each synthetic perturbed input is formed as $I_S(x, Z^{(k)})$, and the estimator computes the empirical average of $\ell(f(I_S(x, Z^{(k)})), y)$ over iid draws $Z^{(k)} \sim p_{\mathrm{bg}}$. For the second claim, under the atomic causal intervention $\mathrm{do}(X_S = x_S, X_{\bar{S}} = Z_{\bar{S}})$, all structural functions determining the coordinates in $[n]$ are severed and replaced by the fixed assignments $(x_S, Z_{\bar{S}})$. Because the downstream predictive function f is frozen, the resulting tensor fed into f is identically $I_S(x, Z)$. Integrating over $Z \sim p_{\mathrm{bg}}$ yields the target in Equation 32. □

Corollary 5. ***Interventional Auditing versus Observational Conditioning.*** *The coalition loss $L(S; x, y)$ is an interventional masking target in the algorithmic auditing sense of Janzing et al. (2020). It is fundamentally distinct from the observational conditional expectation obtained by conditioning on observed features $\mathbb{E}_{X_{\bar{S}} | X_S = x_S}[\ell(f(X), y)]$.*

Resolution of the Out-of-Distribution (OOD) Critique. A frequent critique of masking-based attribution methods (Hooker et al., 2019) argues that synthetic perturbed inputs $m_S \odot x + (1 - m_S) \odot Z$ may fall off the empirical data manifold, rendering neural network evaluations on them uninformative or undefined. Corollary 5 directly resolves this concern: the off-manifold behavior of f is an essential, intentional feature of interventional algorithmic auditing, not an approximation error.

The quantity $L(S; x, y)$ measures the expected loss of f under an atomic coordinate-level intervention that physically isolates components S by ablating all other coordinates *regardless of empirical feature correlations in the training distribution*. As formalized by Janzing et al. (2020), if an explanation method instead samples from the conditional distribution $P(X_{\bar{S}} \mid X_S = x_S)$ as advocated by observational approaches (Heskes et al., 2020), an irrelevant spectator feature X_j that was completely ignored by the trained model f will be assigned substantial non-zero attribution whenever X_j correlates observationally with true causal features in the training data. For auditing the internal mechanism of a fixed, deployed network f, marginal interventional replacement is strictly required to prevent observational leakage.

Comparison to Lipschitz-Based OOD Bounds. An alternative theoretical path would attempt to bound off-manifold extrapolation error $|f(m_S \odot x + (1 - m_S) \odot Z) - f(x^\star)|$ via the global or local Lipschitz constant of f, where $x^\star$ is the nearest on-manifold projection. For deep neural networks—especially transformer models such as GPT2-SMALL—such Lipschitz constants are notoriously vacuous, astronomically large, or NP-hard to compute (Fazlyab et al., 2019). By defining the estimand directly as the interventional expected loss under reference measure p_{bg}, InDe-X eliminates the OOD extrapolation penalty definitionally from the estimation task.

C.2 Welford Streaming Estimation, Hoeffding Concentration, and Sample Complexity

Let $K \geq 1$ and let $Z^{(1)}, \ldots, Z^{(K)} \overset{\text{iid}}{\sim} p_{\text{bg}}$ be independent baseline samples drawn from the reference distribution. The empirical Monte Carlo estimator is:

$$\widehat{L}(S; x, y) \;\coloneqq\; \frac{1}{K} \sum_{k=1}^{K} \ell\big(f(m_S \odot x + (1 - m_S) \odot Z^{(k)}),\, y\big). \tag{33}$$

Proposition 3. ***Unbiasedness and Variance Bound.*** *For every coalition $S \subseteq [n]$, factual input (x, y), and sample budget $K \geq 1$:*

$$\mathbb{E}\big[\widehat{L}(S; x, y)\big] \;=\; L(S; x, y), \qquad \mathrm{Var}\big[\widehat{L}(S; x, y)\big] \;=\; \frac{\sigma^2(S; x, y)}{K}, \tag{34}$$

where $\sigma^2(S; x, y) \coloneqq \mathrm{Var}_{Z \sim p_{\text{bg}}}\big[\ell(f(m_S \odot x + (1 - m_S) \odot Z),\, y)\big]$ is the finite per-sample loss variance.

Proof. Unbiasedness follows immediately by linearity of expectation over K independent, identically distributed draws:

$$\mathbb{E}\big[\widehat{L}(S; x, y)\big] \;=\; \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{Z^{(k)} \sim p_{\text{bg}}}\big[\ell\big(f(I_S(x, Z^{(k)})), y\big)\big] \;=\; L(S; x, y). \tag{35}$$

Because the draws $Z^{(k)}$ are mutually independent, the variance of the sample average equals the population variance divided by K:

$$\mathrm{Var}\big[\widehat{L}(S; x, y)\big] \;=\; \frac{1}{K^2} \sum_{k=1}^{K} \mathrm{Var}_{Z^{(k)}}\big[\ell\big(f(I_S(x, Z^{(k)})), y\big)\big] \;=\; \frac{\sigma^2(S; x, y)}{K}. \tag{36}$$

In InDE-X, these moments are accumulated online via Welford's streaming algorithm (Welford, 1962), which maintains $(\hat{L}, \hat{\sigma}^2, K)$ in a single forward pass without storing individual activation tensors and without catastrophic floating-point cancellation. □

Proposition 4. ***Hoeffding Tail Concentration.*** *Assume the loss function is bounded on the domain of f, such that there exist finite constants $\ell_{\min}, \ell_{\max} \in \mathbb{R}$ with $\ell(\cdot, \cdot) \in [\ell_{\min}, \ell_{\max}]$. Let $B \coloneqq \ell_{\max} - \ell_{\min}$ denote the maximum loss span. Then for any tolerance $\epsilon > 0$ and any coalition $S \subseteq [n]$:*

$$\mathbb{P}\left(\big|\widehat{L}(S; x, y) - L(S; x, y)\big| > \epsilon\right) \;\leq\; 2 \exp\left(-\frac{2K\epsilon^2}{B^2}\right). \tag{37}$$

Proof. Each term $Y_k \coloneqq \ell(f(I_S(x, Z^{(k)})), y)$ in Equation 33 is an independent, identically distributed random variable bounded strictly in the closed interval $[\ell_{\min}, \ell_{\max}]$ of length B. Applying Hoeffding's inequality (Hoeffding, 1963) to the empirical average $\widehat{L} = \frac{1}{K} \sum_{k=1}^{K} Y_k$ with expectation $\mathbb{E}[\widehat{L}] = L$:

$$\mathbb{P}\left(\widehat{L} - L > \epsilon\right) \;\leq\; \exp\left(-\frac{2K^2\epsilon^2}{\sum_{k=1}^{K} B^2}\right) \;=\; \exp\left(-\frac{2K\epsilon^2}{B^2}\right). \tag{38}$$

Applying the symmetric lower-tail bound $\mathbb{P}(L - \widehat{L} > \epsilon) \leq \exp(-2K\epsilon^2/B^2)$ and taking the union bound across both one-sided failure events yields the stated factor of 2. □

Proposition 5. ***Empirical-Bernstein Concentration Bound.*** *Let the bounded-loss condition of Proposition 4 hold ($\ell \in [\ell_{\min}, \ell_{\max}]$ with range $B = \ell_{\max} - \ell_{\min}$). For any coalition $S \subseteq [n]$, let $\hat{\sigma}^2(S)$ be the unbiased sample variance computed online via Welford streaming across $K \geq 2$ draws. Then for any confidence level $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the background draws:*

$$\big|\widehat{L}(S; x, y) - L(S; x, y)\big| \;\leq\; \hat{\sigma}(S; x, y) \sqrt{\frac{2 \log(2/\delta)}{K}} \;+\; \frac{7B \log(2/\delta)}{3(K - 1)}. \tag{39}$$

Proof. Let $Y_k := \ell(f(I_S(x, Z^{(k)})), y) \in [\ell_{\min}, \ell_{\max}]$. Scaling the variables to $[0, 1]$ via $X_k = (Y_k - \ell_{\min})/B$ yields sample variance $V_K(X) = \hat{\sigma}^2(S)/B^2$. Applying Theorem 1 to $X_1, \ldots, X_K$, the upper and lower deviation bounds each fail with probability at most $\delta/2$:

$$\mathbb{P}\left(|\mathbb{E}[X] - \bar{X}| > \sqrt{\frac{2V_K(X)\log(2/\delta)}{K}} + \frac{7\log(2/\delta)}{3(K-1)}\right) \;\leq\; \delta. \tag{40}$$

Rescaling back by multiplying by B preserves the inequality identically and yields Equation 39. □

Corollary 6 (Sample Complexity Bound)**.** *To guarantee that the empirical coalition loss $\widehat{L}(S; x, y)$ estimates $L(S; x, y)$ within additive error tolerance $\epsilon > 0$ with confidence at least $1-\delta$ for $\delta \in (0, 1)$, it is sufficient that the sample count satisfies:*

$$K \;\geq\; \left\lceil \frac{B^2 \log(2/\delta)}{2\epsilon^2} \right\rceil. \tag{41}$$

In particular, the sample complexity scales as $K = \mathcal{O}\left(\frac{B^2}{\epsilon^2}\log\frac{1}{\delta}\right)$, completely independent of input dimension n and model parameter count $|\theta|$. Furthermore, when the local coalition variance is small ($\hat{\sigma}^2(S) \ll B^2$), Proposition 5 certifies that the effective sample complexity sharpens from $\mathcal{O}(B^2/\epsilon^2)$ to $\mathcal{O}(\sigma^2/\epsilon^2)$.

Proof. Setting the two-sided Hoeffding tail bound from Proposition 4 equal to δ:

$$2\exp\left(-\frac{2K\epsilon^2}{B^2}\right) \;\leq\; \delta \quad\Longleftrightarrow\quad -\frac{2K\epsilon^2}{B^2} \;\leq\; \log\frac{\delta}{2} \quad\Longleftrightarrow\quad K \;\geq\; \frac{B^2\log(2/\delta)}{2\epsilon^2}. \tag{42}$$

Taking the ceiling ensures integer sample counts. □

Operational Confidence Bounds in Practice. Propositions 3, 4, and 5 provide a tripartite hierarchy of complementary operational bounds computed online for every coalition $S \in \mathcal{V}$:

1. **Normal-Theory Dynamic Stopping:** Using sample count $K(S)$ and empirical variance $\hat{\sigma}^2(S)$ from Welford streaming, the runtime computes the asymptotic $(1-\alpha)$ Student-t half-width $w_t := t_{\alpha/2,K-1}\,\hat{\sigma}(S)/\sqrt{K(S)}$, terminating when $w_t \leq \epsilon$.
2. **Variance-Adaptive Finite-Sample Certificate:** Proposition 5 supplies a rigorous, non-asymptotic bound directly evaluated on the running Welford variance $\hat{\sigma}^2(S)$, contracting significantly faster than Hoeffding whenever coalitions operate in low-variance saturation regimes ($\hat{\sigma}^2(S) \ll B^2$).
3. **Finite-Sample Worst-Case Certificate:** Proposition 4 supplies an explicit worst-case concentration guarantee that depends strictly on the domain span B, certifying uniform stability across all explored coalitions regardless of empirical variance.

□

D Proof of Theorem 2: Strict Variance Reduction via Coupled Diamond Sampling

In this section, we provide the complete proof of Theorem 2, establishing the exact variance reduction achieved by coupled diamond sampling over an independent uniform baseline under identical evaluation budgets.

D.1 Variance Decomposition of the Estimators

Fix a pair of components $\{i, j\} \subseteq [n]$ with $i \neq j$, a factual instance (x, y), and a context coalition $C \subseteq [n] \setminus \{i, j\}$. Let $Z \sim p_{\mathrm{bg}}$ denote a background draw from the reference measure. Define the four corner evaluations of the diamond configuration:

$$\begin{aligned} a(C) &\coloneqq \ell\big(f(m_{C\cup\{i,j\}} \odot x + (1 - m_{C\cup\{i,j\}}) \odot Z),\, y\big), \\ b(C) &\coloneqq \ell\big(f(m_{C\cup\{i\}} \odot x + (1 - m_{C\cup\{i\}}) \odot Z),\, y\big), \\ c(C) &\coloneqq \ell\big(f(m_{C\cup\{j\}} \odot x + (1 - m_{C\cup\{j\}}) \odot Z),\, y\big), \\ d(C) &\coloneqq \ell\big(f(m_{C} \odot x + (1 - m_{C}) \odot Z),\, y\big). \end{aligned} \tag{43}$$

The pairwise loss mixed difference evaluated on the tuple (C, Z) is defined in direct alignment with Section 3.3 (Equation 7):

$$\xi(C, Z) \coloneqq b(C) + c(C) - a(C) - d(C). \tag{44}$$

The estimand of interest is the expected second-order interaction:

$$\Delta_{ij}^{(L)} \coloneqq \mathbb{E}_{C,Z}\big[\xi(C, Z)\big] = \mathbb{E}_{C}\big[L(C \cup \{i\}) + L(C \cup \{j\}) - L(C \cup \{i, j\}) - L(C)\big]. \tag{45}$$

Budget-Matched Estimators. To ensure strict fairness, both estimators are allocated an identical compute budget of exactly $4M$ forward pass evaluations of the model loss $\ell(f(\cdot), y)$:

1. **Coupled Diamond Estimator ($\widehat{\Delta}_{ij}^{\text{diamond}}$):** Draw M independent context coalitions $C_1, \ldots, C_M$ from the context sampling distribution over $\mathcal{P}([n] \setminus \{i, j\})$. For each context $k \in [M]$, evaluate all four corners $a(C_k), b(C_k), c(C_k), d(C_k)$ coupled using the same background sample $Z_k \sim p_{\mathrm{bg}}$:

$$\widehat{\Delta}_{ij}^{\text{diamond}} \coloneqq \frac{1}{M}\sum_{k=1}^{M} \xi(C_k, Z_k) = \frac{1}{M}\sum_{k=1}^{M}\Big[b(C_k) + c(C_k) - a(C_k) - d(C_k)\Big]. \tag{46}$$

   Total compute cost: $4 \times M = 4M$ forward evaluations.
2. **Uniform Independent Baseline ($\widehat{\Delta}_{ij}^{\text{unif}}$):** Draw $4M$ completely independent context coalitions $C_k^{(a)}, C_k^{(b)}, C_k^{(c)}, C_k^{(d)}$ for $k \in [M]$ (each paired with an independent background draw Z):

$$\widehat{\Delta}_{ij}^{\text{unif}} \coloneqq \frac{1}{M}\sum_{k=1}^{M}\Big[b\big(C_k^{(b)}\big) + c\big(C_k^{(c)}\big) - a\big(C_k^{(a)}\big) - d\big(C_k^{(d)}\big)\Big]. \tag{47}$$

   Total compute cost: $4 \times M = 4M$ forward evaluations.

Both estimators are strictly unbiased: $\mathbb{E}[\widehat{\Delta}_{ij}^{\text{diamond}}] = \mathbb{E}[\widehat{\Delta}_{ij}^{\text{unif}}] = \Delta_{ij}^{(L)}$ by linearity of expectation.

Covariance Structure on the Diamond Geometry. Let $V_a \coloneqq \mathrm{Var}(a(C))$, $V_b \coloneqq \mathrm{Var}(b(C))$, $V_c \coloneqq \mathrm{Var}(c(C))$, and $V_d \coloneqq \mathrm{Var}(d(C))$ denote the marginal variances of the four corner evaluations. The pairwise covariances between the four corners decompose into two geometrically distinct groups:

1. **Adjacent Covariances (Σ_{adj}):** Covariances between coalitions differing by exactly a single coordinate (Hamming distance 1 on the Boolean sub-cube):

$$\Sigma_{\mathrm{adj}} \coloneqq \mathrm{Cov}(a, b) + \mathrm{Cov}(a, c) + \mathrm{Cov}(b, d) + \mathrm{Cov}(c, d). \tag{48}$$

2. **Diagonal Covariances (Σ_{diag}):** Covariances between opposite coalitions differing by both coordinates $\{i, j\}$ simultaneously (Hamming distance 2):

$$\Sigma_{\mathrm{diag}} \coloneqq \mathrm{Cov}(a, d) + \mathrm{Cov}(b, c). \tag{49}$$

Lemma 7. ***Exact Variance Identities.*** *Under the matched budget of $4M$ loss evaluations:*

$$\mathrm{Var}\big(\widehat{\Delta}_{ij}^{\textit{unif}}\big) \;=\; \frac{1}{M}\big(V_a + V_b + V_c + V_d\big), \tag{50}$$

$$\mathrm{Var}\big(\widehat{\Delta}_{ij}^{\textit{diamond}}\big) \;=\; \frac{1}{M}\Big[\big(V_a + V_b + V_c + V_d\big) - 2\big(\Sigma_{\mathrm{adj}} - \Sigma_{\mathrm{diag}}\big)\Big]. \tag{51}$$

Proof. For the uniform estimator, the summands $a(C_k^{(a)})$, $b(C_k^{(b)})$, $c(C_k^{(c)})$, and $d(C_k^{(d)})$ are mutually independent by construction. Hence, all cross-covariance terms vanish identically:

$$\mathrm{Var}\big(\widehat{\Delta}_{ij}^{\mathrm{unif}}\big) \;=\; \frac{1}{M^2}\sum_{k=1}^{M}\Big[\mathrm{Var}(a)+\mathrm{Var}(-b)+\mathrm{Var}(-c)+\mathrm{Var}(d)\Big] \;=\; \frac{1}{M}\big(V_a+V_b+V_c+V_d\big). \tag{52}$$

For the diamond estimator, $\widehat{\Delta}_{ij}^{\mathrm{diamond}}$ is the empirical mean of M independent draws of the random variable $\xi(C, Z) = b(C) + c(C) - a(C) - d(C)$. Applying the standard variance expansion for a linear combination of random variables with sign vector $s = (s_a, s_b, s_c, s_d) = (-1, +1, +1, -1)$:

$$\begin{aligned}
\mathrm{Var}(\xi) \;&=\; \sum_{u\in\{a,b,c,d\}} s_u^2\,\mathrm{Var}(u) + 2\sum_{u<w} s_u s_w\,\mathrm{Cov}(u,w) \\
&=\; (V_a + V_b + V_c + V_d) \\
&\quad + 2\Big[(-1)(+1)\,\mathrm{Cov}(a,b) + (-1)(+1)\,\mathrm{Cov}(a,c) + (-1)(-1)\,\mathrm{Cov}(a,d) \\
&\qquad + (+1)(+1)\,\mathrm{Cov}(b,c) + (+1)(-1)\,\mathrm{Cov}(b,d) + (+1)(-1)\,\mathrm{Cov}(c,d)\Big] \\
&=\; (V_a + V_b + V_c + V_d) \\
&\quad - 2\big(\mathrm{Cov}(a,b) + \mathrm{Cov}(a,c) + \mathrm{Cov}(b,d) + \mathrm{Cov}(c,d)\big) + 2\big(\mathrm{Cov}(a,d) + \mathrm{Cov}(b,c)\big) \\
&=\; (V_a + V_b + V_c + V_d) - 2\big(\Sigma_{\mathrm{adj}} - \Sigma_{\mathrm{diag}}\big).
\end{aligned} \tag{53}$$

Dividing by M yields Equation 51. □

D.2 Adjacency Dominance and Covariance Conditions

Inspection of Equation 51 reveals that coupled sampling reduces variance if and only if the cross-covariance correction term is strictly positive, i.e., $2(\Sigma_{\mathrm{adj}} - \Sigma_{\mathrm{diag}}) > 0$. We formalize this sufficient and necessary condition:

Assumption 1. ***Adjacency Dominance.*** *For the evaluated component pair $\{i, j\} \subseteq [n]$ and the context distribution over $\mathcal{P}([n] \setminus \{i, j\})$, the aggregate covariance between adjacent lattice corners strictly exceeds the covariance between diagonal lattice corners:*

$$\Sigma_{\mathrm{adj}} \;>\; \Sigma_{\mathrm{diag}}. \tag{54}$$

Physical and Statistical Intuition for Assumption 1 (A1). Assumption 1 formalizes the natural geometric property of continuous deep neural networks:

- **Lattice Smoothness:** Coalitions differing by a single component perturbation ($C \leftrightarrow C \cup \{i\}$) produce model activations and output losses that are more strongly co-dependent than coalitions differing by two simultaneous perturbations ($C \leftrightarrow C \cup \{i, j\}$). This makes adjacent correlations strictly tighter than diagonal correlations ($\mathrm{Cov}(a, b) > \mathrm{Cov}(a, d)$).
- **Empirical Prevalence:** As verified empirically in Appendix F (Table 7), Assumption 1 holds in 100% of audited pairs in real-world neural networks across both vision (DenseNet-121 on NIH ChestX-ray14) and language (GPT2-SMALL on IOI).
- **Boundary Case (XOR Parity):** Assumption 1 is strictly non-vacuous. In pure parity functions where intermediate features are statistically orthogonal (such as the synthetic 3-way XOR SCM), $\Sigma_{\mathrm{adj}} = \Sigma_{\mathrm{diag}} = 0$. On such structures, diamond sampling achieves identical variance to uniform sampling ($\rho_v = 1.0$), providing an intentional boundary test for the theory.

D.3 Proof of the Variance Reduction Bound

We now state and prove the complete expanded form of Theorem 2:

Theorem 8. ***Strict Variance Reduction Bound.*** *Let $\{i, j\} \subseteq [n]$ be any pair of components satisfying Assumption 1. Then under equal compute budgets of $4M$ evaluations:*

$$\mathrm{Var}\big(\widehat{\Delta}_{ij}^{\textit{diamond}}\big) \;<\; \mathrm{Var}\big(\widehat{\Delta}_{ij}^{\textit{unif}}\big). \tag{55}$$

Furthermore, the relative variance reduction ratio $\rho_v := \mathrm{Var}(\widehat{\Delta}_{ij}^{\textit{unif}})/\,\mathrm{Var}(\widehat{\Delta}_{ij}^{\textit{diamond}})$ satisfies:

$$\rho_v \;=\; 1 + \frac{2\big(\Sigma_{\mathrm{adj}} - \Sigma_{\mathrm{diag}}\big)}{\big(V_a + V_b + V_c + V_d\big) - 2\big(\Sigma_{\mathrm{adj}} - \Sigma_{\mathrm{diag}}\big)} \;>\; 1. \tag{56}$$

Proof. Subtracting Equation 51 from Equation 50 in Lemma 7:

$$\mathrm{Var}\big(\widehat{\Delta}_{ij}^{\mathrm{unif}}\big) - \mathrm{Var}\big(\widehat{\Delta}_{ij}^{\mathrm{diamond}}\big) \;=\; \frac{2}{M}\big(\Sigma_{\mathrm{adj}} - \Sigma_{\mathrm{diag}}\big). \tag{57}$$

Under Assumption 1, $\Sigma_{\mathrm{adj}} - \Sigma_{\mathrm{diag}} > 0$. Because $M \geq 1$, this variance difference is strictly positive, establishing $\mathrm{Var}(\widehat{\Delta}_{ij}^{\mathrm{diamond}}) < \mathrm{Var}(\widehat{\Delta}_{ij}^{\mathrm{unif}})$.

Dividing $\mathrm{Var}(\widehat{\Delta}_{ij}^{\mathrm{unif}})$ by $\mathrm{Var}(\widehat{\Delta}_{ij}^{\mathrm{diamond}})$:

$$\begin{aligned}\rho_v \;&=\; \frac{\frac{1}{M}(V_a + V_b + V_c + V_d)}{\frac{1}{M}[(V_a + V_b + V_c + V_d) - 2(\Sigma_{\mathrm{adj}} - \Sigma_{\mathrm{diag}})]} \\ &=\; 1 + \frac{2(\Sigma_{\mathrm{adj}} - \Sigma_{\mathrm{diag}})}{(V_a + V_b + V_c + V_d) - 2(\Sigma_{\mathrm{adj}} - \Sigma_{\mathrm{diag}})}.\end{aligned} \tag{58}$$

Since the numerator $2(\Sigma_{\mathrm{adj}} - \Sigma_{\mathrm{diag}}) > 0$ and the denominator is the strictly positive variance $M \cdot \mathrm{Var}(\widehat{\Delta}_{ij}^{\mathrm{diamond}}) > 0$, the ratio satisfies $\rho_v > 1$. □

Connection to Antithetic and Coupled Estimators. Theorem 8 provides the formal rationale for coupled diamond sampling: by orienting the evaluation on 2-dimensional Boolean sub-cubes $\{C, C \cup \{i\}, C \cup \{j\}, C \cup \{i, j\}\}$, the negative-signed adjacent correlations work constructively to cancel Monte Carlo noise. This provides an exact second-order analog to antithetic variates and coupled Shapley sampling (Castro et al., 2009; Muschalik et al., 2024), enabling rapid convergence under user-specified evaluation budgets. □

E Proof of Theorem 3 and Proposition 2: Budgeted Complexity and Uniform Convergence

In this section, we provide the complete theoretical proofs of Theorem 3 (uniform concentration over a finite coalition vocabulary) and Proposition 2 (perturbation stability of the tripartite $U/R/S$ decomposition).

E.1 Proof of Theorem 3: Finite-Vocabulary Uniform Concentration

Setting. Let $\mathcal{V} \subset \mathcal{P}([n])$ be any finite vocabulary of evaluated component coalitions of size $|\mathcal{V}| \geq 1$. For each coalition $S \in \mathcal{V}$, let $\widehat{L}(S)$ be the empirical Monte Carlo loss estimated over K independent background draws $Z^{(1)}, \ldots, Z^{(K)} \overset{\text{iid}}{\sim} p_{\text{bg}}$ according to Equation 33. Assume the loss function is bounded in $[\ell_{\min}, \ell_{\max}]$ with span $B \coloneqq \ell_{\max} - \ell_{\min}$.

Step 1: Uniform Hoeffding Concentration. By Proposition 4, for any single coalition $S \in \mathcal{V}$ and any tolerance $\epsilon > 0$, the two-sided estimation error satisfies:

$$\mathbb{P}\left(\left|\widehat{L}(S) - L(S)\right| > \epsilon\right) \;\leq\; 2\exp\left(-\frac{2K\epsilon^2}{B^2}\right). \tag{59}$$

Applying Boole's inequality (union bound) across all $|\mathcal{V}|$ coalitions in the vocabulary:

$$\begin{aligned}\mathbb{P}\left(\max_{S\in\mathcal{V}}\left|\widehat{L}(S) - L(S)\right| > \epsilon\right) \;&=\; \mathbb{P}\left(\bigcup_{S\in\mathcal{V}}\left\{\left|\widehat{L}(S) - L(S)\right| > \epsilon\right\}\right)\\ &\leq\; \sum_{S\in\mathcal{V}}\mathbb{P}\left(\left|\widehat{L}(S) - L(S)\right| > \epsilon\right)\\ &\leq\; 2|\mathcal{V}|\exp\left(-\frac{2K\epsilon^2}{B^2}\right).\end{aligned} \tag{60}$$

To bound this joint failure probability by a specified confidence parameter $\alpha \in (0, 1)$, we set:

$$2|\mathcal{V}|\exp\left(-\frac{2K\epsilon^2}{B^2}\right) \;\leq\; \alpha \quad\Longleftrightarrow\quad K \;\geq\; \left\lceil\frac{B^2\log(2|\mathcal{V}|/\alpha)}{2\epsilon^2}\right\rceil. \tag{61}$$

Thus, with probability at least $1 - \alpha$, the uniform concentration event $\mathcal{E}_{\text{unif}} \coloneqq \{\forall S \in \mathcal{V} : |\widehat{L}(S) - L(S)| \leq \epsilon\}$ holds simultaneously for all coalitions in $\mathcal{V}$.

Step 2: Exact Optimizer Recovery under Margin Separation. Let $S^\star \coloneqq \arg\max_{S\in\mathcal{V}} L(S)$ be the unique coalition achieving the maximum expected loss in $\mathcal{V}$, and define the sub-optimality margin:

$$\Delta_{\text{sep}} \;\coloneqq\; L(S^\star) - \max_{S\in\mathcal{V}\setminus\{S^\star\}} L(S) \;>\; 0. \tag{62}$$

Assume the margin satisfies $\Delta_{\text{sep}} > 2\epsilon$. Under the uniform concentration event $\mathcal{E}_{\text{unif}}$:

$$\widehat{L}(S^\star) \;\geq\; L(S^\star) - \epsilon. \tag{63}$$

For any competing coalition $S' \in \mathcal{V} \setminus \{S^\star\}$:

$$\widehat{L}(S') \;\leq\; L(S') + \epsilon \;\leq\; L(S^\star) - \Delta_{\text{sep}} + \epsilon \;<\; L(S^\star) - 2\epsilon + \epsilon \;=\; L(S^\star) - \epsilon. \tag{64}$$

Combining both inequalities gives:

$$\widehat{L}(S^\star) \;>\; \widehat{L}(S') \quad \text{for all } S' \neq S^\star. \tag{65}$$

Therefore, the empirical optimizer $\widehat{S}^\star \coloneqq \arg\max_{S\in\mathcal{V}} \widehat{L}(S)$ recovers the true population optimizer $S^\star$ exactly with probability at least $1 - \alpha$. An identical argument with inequalities reversed holds for the minimum coalition loss $\arg\min_{S\in\mathcal{V}} L(S)$.

E.2 Proof of Proposition 2: Perturbation Stability of U/R/S

We now consider the stability of the empirical leave-one-covariate-out (LOCO) gains and the tripartite (U, R, S) decomposition.

Setting. Let $\mathcal{V}$ contain all coalition contexts required to evaluate the LOCO scores for component i. For any context $C \subseteq [n] \setminus \{i\}$, let:

$$G_i(C) \;\coloneqq\; \mathrm{LOCO}(i \mid C) \;=\; L(C) - L(C \cup \{i\}), \qquad \widehat{G}_i(C) \;\coloneqq\; \widehat{L}(C) - \widehat{L}(C \cup \{i\}). \tag{66}$$

Under the uniform concentration event $\mathcal{E}_{\text{unif}}$, $|\widehat{L}(S) - L(S)| \le \epsilon$ for all $S \in \mathcal{V}$. By the triangle inequality:

$$\big|\widehat{G}_i(C) - G_i(C)\big| \;\le\; \big|\widehat{L}(C) - L(C)\big| + \big|\widehat{L}(C \cup \{i\}) - L(C \cup \{i\})\big| \;\le\; \epsilon + \epsilon \;=\; 2\epsilon. \tag{67}$$

This holds uniformly for all evaluated contexts C.

Stability of Minima and Maxima. By standard Lipschitz stability of order statistics under L_∞ perturbations, for any finite sets of real values $\{a_k\}$ and $\{b_k\}$ satisfying $\max_k |a_k - b_k| \le 2\epsilon$:

$$\big|\min_k a_k - \min_k b_k\big| \;\le\; \max_k |a_k - b_k| \;\le\; 2\epsilon, \qquad \big|\max_k a_k - \max_k b_k\big| \;\le\; \max_k |a_k - b_k| \;\le\; 2\epsilon. \tag{68}$$

Applying this to the empirical and population uniqueness and coalition ceilings:

$$\big|\widehat{U}_i - U_i\big| \;=\; \Big|\min_C \widehat{G}_i(C) - \min_C G_i(C)\Big| \;\le\; 2\epsilon, \tag{69}$$

$$\big|\widehat{L}_i^{\max} - L_i^{\max}\big| \;=\; \Big|\max_C \widehat{G}_i(C) - \max_C G_i(C)\Big| \;\le\; 2\epsilon. \tag{70}$$

Furthermore, the standalone singleton gain $\pi_i \coloneqq G_i(\emptyset)$ satisfies:

$$\big|\widehat{\pi}_i - \pi_i\big| \;\le\; 2\epsilon. \tag{71}$$

Error Bounds for Redundancy and Synergy. Applying the definitions of redundancy ($R_i = \pi_i - U_i$) and synergy ($S_i = L_i^{\max} - \pi_i$) alongside the triangle inequality:

$$\big|\widehat{R}_i - R_i\big| \;=\; \big|(\widehat{\pi}_i - \widehat{U}_i) - (\pi_i - U_i)\big| \;\le\; |\widehat{\pi}_i - \pi_i| + |\widehat{U}_i - U_i| \;\le\; 2\epsilon + 2\epsilon \;=\; 4\epsilon, \tag{72}$$

$$\big|\widehat{S}_i - S_i\big| \;=\; \big|(\widehat{L}_i^{\max} - \widehat{\pi}_i) - (L_i^{\max} - \pi_i)\big| \;\le\; |\widehat{L}_i^{\max} - L_i^{\max}| + |\widehat{\pi}_i - \pi_i| \;\le\; 2\epsilon + 2\epsilon \;=\; 4\epsilon. \tag{73}$$

Equations 69, 72, and 73 establish Proposition 2. □

E.3 Theoretical Remarks and Comparison to Alternative Formulations

Connection to Extreme Value Theory (EVT). An alternative approach to estimating the coalition ceiling $L_i^{\max} = \max_C \mathrm{LOCO}(i \mid C)$ would fit an extreme value distribution (e.g., Generalized Extreme Value / Gumbel) to an incomplete random sample of contexts. While EVT is standard when searching unbounded spaces, Theorem 3 provides a fundamentally stronger, distribution-free guarantee for finite budgets: by maintaining a finite active vocabulary $\mathcal{V}$ and evaluating each member to certified precision ϵ, the extremum within $\mathcal{V}$ is bounded deterministically without parametric tail assumptions.

Orthogonality to Theorem 2. Theorem 3 governs the *concentration* of coalition evaluations $\widehat{L}(S)$ as the sample size K grows. Theorem 2 governs the *geometry* of sampling: it establishes that when filling the vocabulary $\mathcal{V}$, coupling the queries on Boolean diamonds reduces the estimation variance of interaction gradients $\widehat{\Delta}_{ij}^{(L)}$ relative to uniform random exploration under identical forward-pass budgets. The two results operate synergistically: Theorem 2 maximizes sampling efficiency, while Theorem 3 bounds the aggregate union risk. □

F Numerical and Empirical Verification of Theoretical Guarantees

Each theoretical proposition and theorem in Section 3 is accompanied by an empirical numerical verification harness. This appendix documents the experimental protocols, parameter specifications, and headline empirical results; the complete reproducible execution scripts and raw numerical logs are provided in the supplementary code release.

F.1 Numerical Verification of Proposition 1 (Instability of Truncated Pairwise Indices)

Proposition 1 is verified exhaustively across all $2^4 = 16$ binary instances of the 3-way XOR structural causal model (SCM). For each instance, we execute six independent mathematical consistency checks:

1. **Pairwise Möbius Coefficients:** Verification that all pairwise Möbius terms $m_x(\{i, j\})$ vanish to machine precision ($|m_x(\{i, j\})| < 10^{-15}$).
2. **Triplet Möbius Coefficient:** Verification that the 3-way Möbius term satisfies $m_x(\{1, 2, 3\}) = \pm 1/2$ exactly, inverting with factual instance parity.
3. **Axiomatic Projection Matching:** Verification that the empirical $\mathrm{SII}_{ij}(x)$ computed from raw forward passes matches the closed-form combinatorial projection derived in Step 4 of Appendix B.
4. **Faithful Library Agreement:** Verification via the canonical `shapiq` library (Muschalik et al., 2024) that all faithful interaction indices (STII, FSII, kADD-SHAP) output pairwise values bounded by $\leq 10^{-10}$.
5. **Projective Library Agreement:** Verification via `shapiq` that projective interaction indices (SII, k-SII at order 2) evaluate to $\pm 1/4$ for active pairs and 0 for spectator pairs.
6. **Möbius Transform Inversion:** Verification that `shapiq`'s numerical Möbius inversion reproduces the analytical Möbius vector to floating-point precision.

All $16 \times 6 = 96$ programmatic unit tests pass with zero exceptions.

F.2 Numerical Verification of Theorem 1 (Monte Carlo Concentration)

Using a multi-layer perceptron (`TwoLayersNet`) trained on the 3-way XOR SCM (validation MSE $\approx 2.7 \times 10^{-3}$), we test the three core statistical properties guaranteed by Theorem 1:

1. **Property 1: Unbiasedness of Welford Estimator.** The empirical bias $\frac{1}{R}\sum_{r=1}^{R} \widehat{L}^{(r)}(S) - L(S)$ was evaluated across $R = 200$ independent Monte Carlo replicates. Across all tested sample budgets $K \in \{8, 16, 32, 64, 128, 256\}$, the bias remains within $1.96\times$ the replicate standard error, confirming unbiasedness (Proposition 3).
2. **Property 2: $1/K$ Variance Decay Rate.** Regressing $\log \mathrm{Var}(\widehat{L})$ against $\log K$ over $R = 200$ trials yields an empirical log-log slope of -0.935, well within the pre-registered theoretical tolerance interval $[-1.15, -0.85]$ centered at -1.0.
3. **Property 3: Hoeffding Concentration Bound.** At budget $K = 64$ across 2,000 independent replicates, the empirical tail deviation probability $\mathbb{P}(|\widehat{L} - L| > 10^{-2})$ was zero, strictly bounded from above by the analytical bound $2\exp(-2K\epsilon^2/B^2)$ guaranteed by Proposition 4.

F.3 Numerical Verification of Theorem 2 (Strict Variance Reduction)

We verify the variance decomposition of Lemma 7 across two controlled synthetic benchmarks and both real experimental settings. Theorem 2 proves that coupled diamond sampling yields strict variance reduction over uniform independent sampling under matched budgets if and only if the adjacency-dominant covariance condition holds: $\Sigma_{\mathrm{adj}} > \Sigma_{\mathrm{diag}}$ (Assumption 1). To rigorously test this condition, we examine both a compliant regime and a pathological boundary regime.

Why Assumption 1 (A1) Fails on Pure XOR: An Intentional Boundary Test. On the pure 3-way parity benchmark (XOR3), Assumption 1 is strictly violated ($\Sigma_{\mathrm{adj}} - \Sigma_{\mathrm{diag}} = -2.83 \times 10^{-4}$), and the observed variance ratio is $\rho_v = 0.86\times < 1$. Far from indicating an algorithmic flaw, this behavior is a direct mathematical consequence of parity symmetry:

Table 7: Empirical verification of Assumption 1 (A1) across real neural network architectures audited in Section 4. "A1 Holds" denotes the count of component pairs on which $\Sigma_{\text{adj}} > \Sigma_{\text{diag}}$ (the necessary and sufficient condition for strict variance reduction under Theorem 2). $\rho_v := \text{Var}(\widehat{\Delta}^{\text{unif}})/\,\text{Var}(\widehat{\Delta}^{\text{diamond}})$.

Audited Setting	**n**	**Pairs**	**A1 Holds**	**Mean ρ_v**	**Median ρ_v**
E3: GPT2-SMALL (10 heads, exhaustive 2^{10})	10	45	$45/45$ (100%)	1.67×10^3	$885.54\times$
E2: DENSENET121 (196 patches, $K = 256$)	196	6	$6/6$ (100%)	6.82×10^6	3.59×10^6

1. In a pure parity relation $Y = X_1 \oplus X_2 \oplus X_3$, inverting any single input bit flips the output bit ($X_i \mapsto 1 - X_i \implies Y \mapsto 1 - Y$). Consequently, adjacent diamond corners along Hamming-1 lattice edges are mutually decorrelated ($\Sigma_{\text{adj}} \approx 0$).
2. Inverting two input bits simultaneously leaves the parity bit invariant ($X_i \oplus X_j \equiv (1 - X_i) \oplus (1 - X_j)$), inducing strong positive covariance between opposite diagonal corners (C and $C \cup \{i, j\}$), such that $\Sigma_{\text{diag}} > 0$.

Because $\Sigma_{\text{diag}} > \Sigma_{\text{adj}}$, the covariance difference is negative, and Theorem 2 predicts that coupled sampling cannot reduce variance. The XOR3 benchmark therefore provides an exact empirical verification of the sharpness of Theorem 2: variance reduction is governed strictly by the sign of $(\Sigma_{\text{adj}} - \Sigma_{\text{diag}})$.

Beneficial Stratification on Continuous Interactive Data. Conversely, on the synthetic third-order benchmark (Synth3), which features non-parity interactive dynamics, adjacent coalitions exhibit natural co-monotonicity, and Assumption 1 holds robustly ($\Sigma_{\text{adj}} - \Sigma_{\text{diag}} = 1.60$). On this benchmark, diamond sampling delivers an empirical $3.58\times$ variance reduction over matched-budget uniform sampling.

Assumption 1 (A1) Verification on Real-World Neural Architectures. To verify that real-world machine learning systems operate within the positive regime, we evaluate Assumption 1 on the clinical vision benchmark E2 (DENSENET121 on NIH ChestX-ray14) and the transformer mechanistic benchmark E3 (GPT2-SMALL on the IOI task):

- **Transformer Circuit Auditing (E3):** Using the exhaustive $2^{10} = 1024$ coalition evaluations cached across five seeds, Σ_{adj} and Σ_{diag} were computed across all $\binom{10}{2} = 45$ head pairs. Assumption 1 holds on $45/45$ pairs (100%), achieving a mean variance reduction ratio of $\rho_v = 1.67 \times 10^3$ (median $885.54\times$).
- **Multi-Region Visual Diagnostics (E2):** Running matched-budget Monte Carlo ($K = 256$ contexts) on six representative patch pairs (three spatially adjacent, three distant) on a Cardiomegaly radiograph under the 14×14 grid ($n = 196$), Assumption 1 holds on $6/6$ evaluated pairs (100%), with positive covariance gap $\Sigma_{\text{adj}} - \Sigma_{\text{diag}} \in [3.8, 4.9] \times 10^{-3}$ and mean $\rho_v = 6.82 \times 10^6$.

The headline statistics are reported in Table 7.

F.4 NUMERICAL VERIFICATION OF THEOREM 3 (UNIFORM CONVERGENCE)

Theorem 3 is verified on the full powerset lattices of the XOR3 benchmark ($|\mathcal{V}| = 16$) and the third-order benchmark ($|\mathcal{V}| = 256$). A high-precision reference vocabulary L_{ref} was constructed at $K_{\text{ref}} = 4{,}096$. We then performed independent runs across varying budgets $K \in [8, 512]$ and evaluated:

1. **Uniform Convergence Rate:** The empirical maximum error $\max_{S \in \mathcal{V}} |\widehat{L}(S) - L_{\text{ref}}(S)|$ decays strictly monotonically as $\mathcal{O}(\sqrt{\log |\mathcal{V}|/K})$, matching Equation 60.
2. **Optimizer Recovery:** In 100% of trials where the empirical maximum error was below half the separation gap ($\epsilon < \Delta_{\text{sep}}/2$), the empirical optimizer $\widehat{S}^\star$ matched the true optimizer $S^\star$.
3. **Component Error Bound:** The empirical errors $|\widehat{U}_i - U_i|$, $|\widehat{R}_i - R_i|$, and $|\widehat{S}_i - S_i|$ were confirmed to remain strictly bounded by 2ϵ, 4ϵ, and 4ϵ, satisfying Proposition 2.

F.5 Background-Distribution (p_{bg}) Sensitivity and Bounded-Loss Diagnostics

Because interventional masking relies on the reference measure p_{bg}, we conducted systematic sensitivity analyses across four distinct reference choices: marginal empirical product, uniform bounding-box sampling, Gaussian blur imputation, and zero/mean activation replacement (detailed in Appendix H). Across all choices, relative role rankings and interaction sign structures remain stable, exhibiting bounded numerical drift. Furthermore, empirical per-example loss ranges support the bounded-loss assumptions required for non-asymptotic Hoeffding concentration in Theorem 1.

Real-Task Variance-Reduction Evidence. Beyond the synthetic SCM checks establishing the theoretical grounding for diamond sampling, we directly quantify uniform-vs-diamond estimator variance on two real tasks. For E3 (IOI circuit analysis, $n = 10$ heads), Assumption 1 holds for all 45 evaluated pairs, with variance ratios $\rho_v = \mathrm{Var}(\widehat{\Delta}^{\mathrm{unif}})/\,\mathrm{Var}(\widehat{\Delta}^{\mathrm{diam}})$ spanning 108 to $9{,}807$ (mean $1{,}674$, median 886) at $K = 256$ contexts per pair. For E2 (ChestX-ray attribution), Assumption 1 holds for all 6 sampled pairs, with ρ_v spanning 2.28×10^4 to 1.82×10^7 (mean 6.82×10^6, median 3.59×10^6) on a 14×14 feature grid (196 features), at $K = 256$. Replicate-level IOI checks across 32 independent runs at $K = 128$ further yield $\rho_v \in [526,\ 3{,}025]$ (mean $2{,}022$), confirming stability of these ratios across estimator realizations. Taken together, these results directly quantify substantial real-task variance reduction on two structurally distinct tasks.

G Extended Methodology, Algorithmic Framework, and Implementation Details

This section expands upon the methodology presented in Section 3, detailing the algebraic identities, lattice-theoretic foundations, exploration policies, online estimation routines, and complete algorithmic structure of InDe-X.

G.1 Per-Component U/R/S Identity and Loss Bounds

For any input component $i \in [n]$ and context coalition $C \subseteq [n]\backslash\{i\}$, recall the leave-one-covariate-out (LOCO) marginal gain from Equation 2:

$$\mathrm{LOCO}(i \mid C) \;\coloneqq\; L(C) - L(C \cup \{i\}). \tag{74}$$

The per-component summary metrics are defined as:

$$U_i \;\coloneqq\; \min_{C \subseteq [n]\backslash\{i\}} \mathrm{LOCO}(i \mid C), \quad L_i^{\max} \;\coloneqq\; \max_{C \subseteq [n]\backslash\{i\}} \mathrm{LOCO}(i \mid C), \quad \pi_i \;\coloneqq\; \mathrm{LOCO}(i \mid \emptyset). \tag{75}$$

The baseline singleton score π_i admits the equivalent formulation:

$$\pi_i \;=\; \mathrm{Var}(Y) \;-\; \mathbb{E}_{Z \sim p_{\mathrm{bg}}}\big[\ell\big(f(m_{\{i\}} \odot x + (1 - m_{\{i\}}) \odot Z),\, y\big)\big], \tag{76}$$

when ℓ is squared error and the empty coalition loss $L(\emptyset)$ is anchored to a constant baseline outputting the marginal mean $\mathbb{E}[Y]$. The redundancy/synergy split $R_i \coloneqq \pi_i - U_i$ and $S_i \coloneqq L_i^{\max} - \pi_i$ provides an exact algebraic partitioning of the model's contextual dynamic range:

$$L_i^{\max} \;=\; U_i + R_i + S_i, \tag{77}$$

which holds identically by definition without approximation error.

G.2 Axiomatic Grounding: Dedekind Redundancy Lattice versus Boolean Mask Lattice

Classical Partial Information Decomposition (PID; Williams & Beer 2010) conceptualizes multi-source interactions by defining a redundancy measure $I_{\min}$ over a lattice of antichains (the Dedekind redundancy lattice). While theoretically foundational, applying classical PID directly to deep neural representations faces two insurmountable bottlenecks:

1. Super-Exponential Cardinality of the Dedekind Lattice. The number of nodes in the classical redundancy lattice over n variables is governed by the Dedekind number $M(n)$, which grows super-exponentially:

$$M(1) = 2, \quad M(2) = 6, \quad M(3) = 20, \quad M(4) = 168,$$
$$M(5) = 7{,}581, \quad M(6) \approx 7.8 \times 10^6, \quad M(7) \approx 2.4 \times 10^{11}.$$

For modern deep network audits (e.g., $n = 10$ or $n = 26$ attention heads in Transformer circuits, or $n = 196$ spatial patches in vision models), the Dedekind lattice cannot be stored, traversed, or indexed.

2. Intractability of Continuous Density Estimation. Classical PID requires computing continuous mutual information terms $I(X_S; Y)$ across latent activations. In modern deep models, high-dimensional neural representations render exact continuous density estimation intractable, while parametric variational estimators introduce high variance and optimization bias.

Operational Grounding on the Boolean Mask Lattice. To overcome these barriers, InDe-X grounds the decomposition directly on the Boolean lattice of feature masks $(\mathcal{P}([n]), \subseteq)$. The Boolean lattice has cardinality 2^n and can be tractably sampled and budgeted via an active vocabulary $\mathcal{V}$ (Section 3.2). On this lattice, our LOCO extremal bounds provide the operational, loss-based counterpart to modern variational PID frameworks (Bertschinger et al., 2014; Kolchinsky, 2022):

- The minimum marginal gain $U_i = \min_C \mathrm{LOCO}(i \mid C)$ measures the guaranteed performance contribution of feature i across all context environments, operationalizing *unique capability*.

- The standalone gain $\pi_i = \mathrm{LOCO}(i \mid \emptyset)$ measures individual capacity in isolation; its difference from uniqueness $R_i = \pi_i - U_i$ isolates *redundant / substitutable capacity* without auxiliary density estimation.
- The residual $S_i = L_i^{\max} - \pi_i$ quantifies the maximum cooperative gain enabled by coalition context, isolating *synergy*.

This formulation preserves the conceptual interpretability of the tripartite codomain while delivering provable finite-sample convergence on black-box networks.

G.3 Coalition Vocabulary with Welford Streaming Updates

The active coalition vocabulary $\mathcal{V}$ is maintained as an online hash table mapping each visited coalition mask S to its running summary tuple:

$$S \;\longmapsto\; \big(\widehat{L}(S),\, \widehat{\sigma}^2(S),\, K(S)\big), \tag{78}$$

representing the empirical sample mean, unbiased sample variance, and visit count. Updates are executed online via Welford's algorithm (Welford, 1962) on streaming minibatches of factual instances $\{(x^{(b)}, y^{(b)})\}$ and background vectors $\{Z^{(k)}\}_{k=1}^{K}$. Memory consumption scales as $\mathcal{O}(|\mathcal{V}|)$, requiring less than 1 MB of RAM for typical budgets without caching intermediate activations.

Dynamic Convergence Detection. A coalition S is flagged as converged when the half-width of its $(1-\alpha)$ Student-t confidence interval,

$$w(S) \;\coloneqq\; t_{\alpha/2,\,K(S)-1} \cdot \frac{\widehat{\sigma}(S)}{\sqrt{K(S)}}, \tag{79}$$

falls below a user-specified tolerance threshold $w_\star$. Once converged, S is transitioned out of the active sampling distribution, concentrating remaining computation on under-explored regions.

Balanced Exploration Policy. Candidate coalitions are sampled using an ε-greedy exploration policy:

$$p(S) \;=\; \varepsilon \cdot 2^{-n} \;+\; (1-\varepsilon)\,\mathbf{1}\big[S \in \mathcal{V}_{\mathrm{open}}\big] \cdot \frac{\exp(-\beta\, K(S))}{\sum_{S' \in \mathcal{V}_{\mathrm{open}}} \exp(-\beta\, K(S'))}, \tag{80}$$

where $\mathcal{V}_{\mathrm{open}} \subseteq \mathcal{V}$ denotes the subset of active, non-converged coalitions and $\beta > 0$ is an inverse temperature parameter. The uniform exploration term ($\varepsilon = 0.2$) guarantees full lattice reachability in the limit, while the softmin term preferentially directs queries toward least-visited masks, ensuring the Hoeffding condition of Theorem 3 is satisfied uniformly across $\mathcal{V}$.

G.4 Coupled Diamond Sampling Mechanics and Pairwise Interactions

For any pair of components $\{i, j\} \subseteq [n]$, the second-order pairwise synergy S_{ij} and redundancy R_{ij} are defined as the positive and negative extrema of the second-order mixed difference $\Delta_{ij}^{(L)}(C)$ across all evaluated contexts:

$$S_{ij} \;\coloneqq\; \max_{C} \big[\Delta_{ij}^{(L)}(C)\big]_+, \qquad R_{ij} \;\coloneqq\; \max_{C} \big[-\Delta_{ij}^{(L)}(C)\big]_+, \tag{81}$$

where $[z]_+ \coloneqq \max(z, 0)$ (Equation 9). For each context C and component pair (i, j), InDe-X evaluates the four coalitions forming the Boolean diamond $D(C, i, j) = \{C, C \cup \{i\}, C \cup \{j\}, C \cup \{i, j\}\}$ using the *identical* background noise sample $Z \sim p_{\mathrm{bg}}$.

Theorem 2 (and expanded Theorem 8) establishes that this coupled sampling strategy yields strict variance reduction whenever Assumption 1 holds. To ensure runtime safety during automated auditing:

1. **Online Monitoring Protocol for Assumption 1 (A1):** During sampling, we continuously track empirical adjacent covariances $\widehat{\Sigma}_{\mathrm{adj}}$ and diagonal covariances $\widehat{\Sigma}_{\mathrm{diag}}$.
2. **Fallback Safeguard:** In the rare event that an empirical violation is detected ($\widehat{\Sigma}_{\mathrm{adj}} \le \widehat{\Sigma}_{\mathrm{diag}}$), the estimator automatically switches to uncoupled uniform sampling for the affected pair.
3. **Diagnostic Logging:** All covariance margins are logged to verify theorem preconditions for published audits.

G.5 Inference-Based Decomposition as Model Fidelity Auditing

Classical feature-attribution frameworks often evaluate coalition loss $L(S)$ by *retraining* the neural network from scratch on feature subset S. Retraining answers the counterfactual question: *"How well could a hypothetical model perform if trained exclusively on features S?"* This is a property of the data distribution and training algorithm, not the deployed black-box model. Retraining introduces retraining stochasticity, requires thousands of GPU hours, and risks attributing importance to features that the deployed model internally ignores.

In contrast, Equation 1 performs masked inference on the fixed, pre-trained network f. This provides three decisive advantages:

1. **Fidelity to the Deployed Network:** The decomposition audits what f actually computes, preserving internal compensatory circuits and backup pathways.
2. **Elimination of Retraining Noise:** Masked inference removes optimization variance, allowing the finite-sample bounds in Theorem 3 to depend strictly on the Monte Carlo sample count K.
3. **Expected Loss per Sample:** By defining $L(S) \coloneqq \mathbb{E}_Z[\ell(f(x^{(S,Z)}), y)]$ rather than $\ell(\mathbb{E}_Z[f(x^{(S,Z)})], y)$ (Covert et al., 2020), losses are evaluated per background draw, enabling streaming Welford updates and strict concentration bounds.

G.6 Algorithmic Skeleton

For completeness, we record the full algorithmic skeleton of the InDe-X auditing engine. The procedure operates in two distinct phases: stochastic vocabulary population via streaming Welford updates (Phase 1), followed by deterministic lattice extraction of the (U, R, S) channels and pairwise matrices (Phase 2).

```
Input : Model f, background p_bg, input batch B = {(x, y)},
        budget T, threshold w_star, pair set P (optional).
State : Vocabulary V = {} mapping S -> (L_hat, sigma2_hat, K).

# Phase 1: Stochastic Vocabulary Population
for t = 1, ..., T while V has non-converged entries:
    sample context C via softmin(visit_count) over non-converged S.
    if pair-mode enabled and P is not empty:
        pick (i, j) in P; form diamond D = {C, C+{i}, C+{j}, C+{i,j}}.
    else:
        form pair D = {C, C + {i}}.
    draw baseline imputation batch {Z^(k)} ~ p_bg of size K_batch.
    for each coalition S in D:
        for each (x, y) in B and each Z^(k):
            loss <- ell( f( mask(x, S, Z^(k)) ), y ).
            V[S].update_welford(loss).
        if width( V[S], alpha ) < w_star:
            mark V[S] converged; remove from open set.

# Phase 2: Deterministic Lattice Extraction
for each component i in [n]:
    for each context C in V (with i not in C and C+{i} in V):
        LOCO(i | C) <- V[C].mean - V[C + {i}].mean.
    U_i     <- min_C LOCO(i | C).
    L_max_i <- max_C LOCO(i | C).
    pi_i    <- LOCO(i | emptyset).
    R_i     <- pi_i - U_i.
    S_i     <- L_max_i - pi_i.

for each evaluated pair (i, j) in P:
    S_ij <- max_C [ Delta_ij^(L)(C) ]_+.
    R_ij <- max_C [ -Delta_ij^(L)(C) ]_+.

return V, {(U_i, R_i, S_i, pi_i, L_max_i)}, {S_ij, R_ij}.
```

H Extended Experimental Evaluation and Empirical Diagnostics

This appendix provides extended experimental protocols, detailed comparator tables, and comprehensive empirical analyses for Section 4. The sections track the experimental structure of the main paper:

- Section H.1 details the synthetic tabular benchmarks (E1), reporting conflation ratios, Wilcoxon significance tests, and ground-truth functional recovery.
- Section H.2 presents the full clinical evaluation on NIH ChestX-ray14 (E2), benchmarking spatial localization against causal deletion faithfulness, alongside anatomical patch distance distributions.
- Section H.3 documents attention-head circuit discovery on GPT2-SMALL (E3), per-head functional attribution, mechanistic role-pair interaction blocks, and downstream causal circuit pruning.
- Section H.4 extends the circuit analysis to the full 26-head candidate set.
- Section H.5 provides background-distribution (p_{bg}) sensitivity ablations.
- Section H.6 analyzes spatial grid resolutions and statistical power for clinical attribution.

H.1 Synthetic Tabular Benchmarks and Index Disambiguation (E1)

Experimental Setup and SCM Testbeds. We train a two-layer multi-layer perceptron (MLP) on three tabular structural causal models (SCMs):

1. **3-way XOR SCM ($\oplus 3$):** Primary testbed for Proposition 1, comprising four binary features where only (X_1, X_2, X_3) are functionally active while X_4 is an independent spectator feature.
2. **Hybrid XOR+AND SCM ($\oplus + \wedge$):** Secondary testbed incorporating both 3-way parity and a 2-way synergistic AND mechanism.
3. **Synthetic Third-Order SCM (Synth3):** Eight continuous features with ground-truth (U, R, S) structure: redundant pair (X_1, X_2), 2-way synergy pair (X_3, X_4), unique feature X_5, and 3-way synergy triplet (X_6, X_7, X_8).

All models are evaluated across 5 random seeds (mean $\pm$ standard deviation reported). Baselines are evaluated using `shapiq` 1.4 (Muschalik et al., 2024) with 100 background samples for interventional masking (Equation 1). We include five canonical pair-level indices: STII (Sundararajan et al., 2020), Faith-SHAP / FSII (Tsai et al., 2023), SII (Grabisch & Roubens, 1999), k-SII at order 2 (Muschalik et al., 2024), and kADD-SHAP (Muschalik et al., 2024). INDE-X operates under an ε-greedy ($\varepsilon = 0.2$) softmin exploration policy (Appendix G.3) with budgets of $5{,}000$ forward passes for 2^4 SCMs and $30{,}000$ for 2^8 Synth3.

Consolidated Disambiguation and Recovery Benchmark. Table 8 provides the consolidated benchmark across all three structural causal models. On both XOR-family models ($\oplus 3$ and $\oplus + \wedge$), canonical pair-level indices exhibit extreme conflation ratios $\rho_m \geq 100\times$, exceeding $411\times$ for Faith-SHAP and kADD-SHAP. As proven in Proposition 1, this is an unavoidable mathematical limitation of scalar pairwise projections: the 3-way interaction is projected into spurious pairwise scalars that invert sign with input parity. Conversely, on Synth3, INDE-X accurately recovers the true continuous functional roles for per-feature uniqueness ($r = 0.978 \pm 0.02$) and per-pair synergy ($r = 0.912 \pm 0.06$).

Statistical Significance and Extended Comparator Coverage. Table 9 reports pair-level conflation ratios alongside one-sided Wilcoxon signed-rank tests ($H_1 : |\widehat{S} + \widehat{R}| > |\phi^m|$). To evaluate broader comparator families, we also benchmark three higher-order interaction baselines: Higher-Order Integrated Gradients (HO-IG), Tree-NID (T-NID), and H-Sets. Under family-wise Bonferroni correction ($\alpha_{\mathrm{fam}} = 0.05/8 = 0.00625$), higher-order baselines remain significantly conflated on XOR3: HO-IG ($19.82\times$, $p = 1.42\times10^{-4}$), T-NID ($14.35\times$, $p = 4.07\times10^{-3}$), and H-Sets ($107.97\times$, $p = 7.99 \times 10^{-6}$). This confirms the central representation thesis of Proposition 1: broadening a method's interaction order does not resolve conflation as long as the output collapses to a single scalar score per coalition rather than preserving independent (U, R, S) channels.

Detailed Recovery on Continuous Synth3. On the synthetic third-order benchmark, Table 10 reports Pearson and Spearman correlation coefficients between INDE-X estimates and ground-truth component roles. Recovery is strongest for per-feature uniqueness ($r = 0.978 \pm 0.022$) and per-pair

Table 8: Consolidated tabular disambiguation and ground-truth recovery benchmark (E1). **Left:** Conflation ratios $\rho_m := (|\hat{U}_i| + |\hat{S}_{ij}| + |\hat{R}_{ij}|)/|\phi_{ij}^m|$ on the canonical 3-way $\oplus 3$ and hybrid $\oplus{+}\wedge$ structural causal models (mean over 5 random seeds). Canonical pair-level interaction indices (STII, Faith-SHAP, SII, k-SII, kADD-SHAP) project higher-order cooperative effects into spurious pairwise scalars that conflate distinct causal mechanisms by over $100\times$ to $400\times$. **Right:** Ground-truth functional role recovery on the synthetic 8-feature third-order benchmark (Synth3, mean $\pm$ std over 5 seeds). INDE-X faithfully disentangles per-feature uniqueness (U, Pearson $r = 0.978$) and pairwise synergy (S_{pair}, Pearson $r = 0.912$), overcoming the representational dilemma formalized in Proposition 1.

(a) Conflation ratios ρ_m on XOR SCMs.

Baseline	$\oplus$**3**	$\oplus$**+**$\wedge$
STII	$319.51\times$	$102.55\times$
Faith-SHAP	$411.88\times$	$110.68\times$
SII	$411.14\times$	$110.66\times$
k-SII	$411.14\times$	$110.66\times$
kADD-SHAP	$411.88\times$	$110.68\times$

(b) Synth3 ground-truth recovery.

Target	**Pearson** r	**Spearman** ρ
U	0.978 ± 0.02	0.577 ± 0.00
S_{pair}	0.912 ± 0.06	0.548 ± 0.12

Table 9: Conflation ratio $\rho_m := (|\hat{U}_i| + |\hat{S}_{ij}| + |\hat{R}_{ij}|)/|\phi_{ij}^m|$ on E1 (mean over 5 seeds). One-sided Wilcoxon tests evaluate $H_1 : |\widehat{S} + \widehat{R}| > |\phi^m|$ with family-wise Bonferroni correction $\alpha_{\text{fam}} = 0.05/8 = 0.00625$. Canonical pair-level baselines reject H_0 and exceed the $10\times$ threshold on XOR3; higher-order scalar baselines (HO-IG, T-NID, H-Sets) also exhibit severe conflation.

Baseline	3-way XOR SCM		XOR+AND SCM		α_{fam}
	ρ_m	Wilcoxon p	ρ_m	Wilcoxon p	
STII	$319.51\times$	9.31×10^{-10}	$102.55\times$	9.31×10^{-10}	0.00625
FSII	$411.88\times$	9.31×10^{-10}	$110.68\times$	9.31×10^{-10}	0.00625
SII	$411.14\times$	9.31×10^{-10}	$110.66\times$	9.31×10^{-10}	0.00625
k-SII	$411.14\times$	9.31×10^{-10}	$110.66\times$	9.31×10^{-10}	0.00625
kADD-SHAP	$411.88\times$	9.31×10^{-10}	$110.68\times$	9.31×10^{-10}	0.00625
Higher-order baselines (XOR3 only):					
HO-IG	$19.82\times$	1.42×10^{-4}	–	–	0.00625
T-NID	$14.35\times$	4.07×10^{-3}	–	–	0.00625
H-Sets	$107.97\times$	7.99×10^{-6}	–	–	0.00625

synergy ($r = 0.912 \pm 0.058$). The difference between Spearman and Pearson correlations on U and R_{pair} is an artifact of tied ranks in binary indicator ground truth rather than estimation error.

H.2 CLINICAL MULTI-REGION RADIOGRAPHY ON NIH CHESTX-RAY14 (E2)

Clinical Evaluation Setup. We evaluate INDE-X on multi-label thoracic pathology diagnosis using the NIH ChestX-ray14 dataset (Wang et al., 2017) and a pre-trained DenseNet-121 classifier (Huang et al., 2017). The testbed comprises all $n = 880$ frontal chest radiographs featuring radiologist-annotated bounding boxes across 8 pathologies (984 total pathology boxes). Images are discretized into an operational 14×14 grid of spatial patches ($n = 196$). Interventional masking applies an 11×11 Gaussian blur perturbation measure p_{bg} ($\sigma = 5.0$), removing high-frequency diagnostic cues while preserving natural radiograph context without edge artifacts.

Spatial Localization versus Causal Deletion Faithfulness. Table 11 benchmarks INDE-X against GradCAM (Selvaraju et al., 2017), Vanilla Gradient, Integrated Gradients (Sundararajan et al., 2017), and a Random baseline (Sundararajan & Najmi, 2020):

1. **Spatial Localization:** Pointing Game accuracy (↑, fraction of instances where the peak attribution falls within a radiologist bounding box) and Bounding Box Intersection over Union (IoU@15 ↑).

Table 10: Ground-truth recovery correlations on the continuous Synth3 benchmark (mean ± std across 5 seeds). Per-feature rows compare the 8-dimensional (U, R, S) vectors against ground-truth role indicators; per-pair rows compare the $\binom{8}{2} = 28$ upper-triangular pair entries.

Target	Pearson r	Spearman ρ
U (per-feature)	0.978 ± 0.022	0.577 ± 0.000
R (per-feature)	0.185 ± 0.464	0.252 ± 0.422
S (per-feature)	0.846 ± 0.125	0.800 ± 0.090
S_{pair} (per-pair)	0.912 ± 0.058	0.548 ± 0.116
R_{pair} (per-pair)	0.261 ± 0.109	0.269 ± 0.057

Table 11: Quantitative clinical evaluation benchmark on NIH ChestX-ray14 across all $n = 880$ annotated radiographs (984 bounding boxes) for DenseNet121. We evaluate spatial localization via Pointing Game accuracy (↑, proportion of max-attribution hits inside ground-truth bounding boxes) and Bounding Box IoU at threshold 0.15 (IoU@15 ↑), alongside causal faithfulness via Deletion AUC (↓). While unregularized pixel-gradient baselines (Vanilla Gradient, Integrated Gradients) exhibit high Pointing Game scores inside human-drawn boxes due to high-frequency edge sensitivity, INDE-X achieves statistically superior causal deletion faithfulness (0.269 vs. 0.332 for GradCAM, paired Wilcoxon $\Delta = -0.063$, $p = 1.61 \times 10^{-78}$), demonstrating that interventional coalition inference successfully isolates the causal pathways driving model predictions.

Method	**Pointing Game ↑**	**IoU@15 ↑**	**Deletion AUC ↓**
GradCAM	0.268	0.142	0.332
Vanilla Gradient	**0.433**	**0.145**	0.315
Integrated Gradients	0.335	0.114	0.294
Random	0.030	0.043	0.309
INDE-X (Ours)	0.286	0.134	**0.269**

2. **Causal Deletion Faithfulness:** Deletion AUC (↓, area under the classification probability degradation curve under progressive patch removal ranked by attribution).

The benchmark results in Table 11 demonstrate an essential trade-off. Unregularized pixel-gradient methods (Vanilla Gradient, Integrated Gradients) achieve higher Pointing Game scores (0.433 and 0.335) by tracking high-frequency edge gradients along rib margins and tissue interfaces that intersect bounding boxes, but fail in causal deletion faithfulness (0.315 and 0.294). Conversely, INDE-X achieves the best causal deletion faithfulness across all evaluated methods (DelAUC = 0.269), significantly outperforming GradCAM (0.332, paired Wilcoxon $\Delta = -0.063$, $p = 1.61 \times 10^{-78}$). On Pointing Game, INDE-X achieves competitive accuracy (0.286 vs. 0.268 for GradCAM, $p = 0.140$), with a minor resolution deficit in IoU (0.134 vs. 0.142, $p = 5.82 \times 10^{-3}$) stemming from the operational 14×14 patch geometry (Section H.6).

Spatial Distribution of Pairwise Synergy and Redundancy. Table 12 evaluates the Euclidean patch distance $d(i, j)$ between interacting spatial regions across all 984 annotated pathology instances.

While redundancy concentrates locally along contiguous tissue borders (43.1% of top redundant pairs occur at $d \leq 2$ patches), nearly one-third (32.7%) of top synergistic pairs bridge long-range anatomical distances ($d \geq 4$ patches). A paired Wilcoxon test confirms that synergistic interactions operate across significantly larger spatial spans than redundant interactions ($p = 8.7 \times 10^{-12}$, Cohen's $d = 0.20$). In conditions such as Cardiomegaly, top synergistic coalitions link the cardiac apex across the thoracic cavity to contralateral mediastinal borders, isolating bilateral anatomical co-dependencies required for diagnosis.

H.3 MECHANISTIC ATTENTION-HEAD CIRCUIT DISCOVERY ON GPT2-SMALL (E3)

Circuit Discovery Protocol. We deploy INDE-X on GPT2-SMALL (124M parameters, 12 layers × 12 heads) on the Indirect Object Identification (IOI) task (Wang et al., 2023). We analyze the $n = 10$

Table 12: Spatial distance analysis of interaction pairs on NIH ChestX-ray14 ($N = 984$ pathology bounding boxes across 880 unique images) under an independent 11×11 Gaussian blur perturbation baseline. Pairwise synergy (S_{ij}) operates over significantly greater anatomical distances than redundancy (R_{ij}), with nearly one-third of top synergistic pairs spanning distal thoracic spans ($d \geq 4$ patches), while redundant pairs cluster within local contours ($d \leq 2$).

Patch Pairs	$\bar{\mathbf{d}}$ **(patches)**	**Median** d	$\mathbf{d \leq 2}$ **(%)**	$\mathbf{d \geq 4}$ **(%)**
Top-1% synergy (S_{ij})	3.33 ± 1.22	2.24	41.2	32.7
Top-1% redundancy (R_{ij})	3.25 ± 1.22	2.24	43.1	31.3
All pairs (null)	7.32 ± 0.00	7.21	5.4	82.4
Paired Wilcoxon ($S > R$, $N = 984$ findings): $p = 8.7 \times 10^{-12}$, Cohen's $d = 0.20$				
Per-row Spearman$(d, S_{ij} - R_{ij})$: $+0.005 \pm 0.022$ (positive in 58.4% of rows)				

Table 13: Per-head interventional decomposition on the GPT2-SMALL IOI circuit across the $n = 10$ core attention heads (mean over 5 seeds). U, R, and S denote unique, redundant, and synergistic contributions, while $\pi := v(\emptyset) - v(\{h\})$ represents the singleton leave-one-component-out score, which algebraically equals mean-ablation activation patching (Spearman $\equiv 1.000$). Crucially, heads with negative or near-zero first-order impact π (e.g., 10.0 with $\pi = -0.235$, 9.9 with $\pi = -0.052$) exert massive cooperative synergy ($S = 1.018$ and 0.586, respectively). Furthermore, Name-Mover head 9.6 exhibits large negative uniqueness ($U = -1.214$) and redundancy ($R = 1.368$), correctly capturing its established mechanistic function as an anti-name-mover backup head.

Head	**Mechanistic Role**	U	R	S	π
9.9	Name-Mover	-0.271	0.219	0.586	-0.052
9.6	Name-Mover	-1.214	1.368	0.038	0.153
10.0	Name-Mover	-0.245	0.010	1.018	-0.235
7.3	S-Inhibition	0.135	0.087	0.430	0.222
7.9	S-Inhibition	0.366	0.091	0.673	0.457
8.6	S-Inhibition	0.320	0.147	1.348	0.467
8.10	S-Inhibition	0.274	0.148	0.848	0.422
5.5	Induction	0.222	0.148	1.139	0.369
3.0	Duplicate-Token	0.156	0.164	0.796	0.320
4.11	Previous-Token	-0.041	0.021	0.461	-0.020

core heads stratified across five established mechanistic roles: Name-Mover, S-Inhibition, Induction, Duplicate-Token, and Previous-Token.

Coalitions are evaluated via mean activation ablation: for heads $(L, h) \notin S$, the attention head output vectors (prior to the linear projection W_O, commonly denoted as z-activations) are replaced with mean activations cached on a held-out corrupted dataset of 30 ABC sentences. Loss is defined as the negative logit difference $v(S) = -\mathbb{E}[\ell_{\text{IO}} - \ell_{\text{Subj}}]$ evaluated on clean ABBA/BABA prompts. With $n = 10$, the complete Boolean lattice of $2^{10} = 1{,}024$ coalitions is evaluated exhaustively across 5 random seeds (wall time: 21 minutes on CPU).

Per-Head Decomposition and First-Order Alignment. Table 13 presents the per-head decomposition across the 10 core circuit heads, reporting uniqueness (U), redundancy (R), synergy (S), and singleton score π. Table 14 provides the extended decomposition including coalition ceilings $L^{\max}$.

Algebraic Equivalence to Activation Patching and Discovery of Synergistic Circuits. As a theoretical consistency check, the singleton score $\pi_i = v(\emptyset) - v(\{i\})$ is algebraically identical to single-head mean-ablation activation patching (Spearman$(\pi, \text{act-patch}) \equiv 1.000$). The critical empirical breakthrough of INDE-X lies in the orthogonal cooperative dimensions: heads exhibiting near-zero standalone impact ($\pi \approx 0$) but substantial cooperative synergy ($S > 0$), such as head 10.0 ($S = +1.02$, $\pi = -0.24$), head 9.9 ($S = +0.59$, $\pi = -0.05$), and head 4.11 ($S = +0.46$, $\pi = -0.02$). These heads are invisible to standard activation patching but exert decisive multi-head cooperative influence.

Table 14: Extended per-head decomposition on GPT2-SMALL IOI including coalition ceiling $L_i^{\max} = U_i + R_i + S_i$ (mean across 5 seeds; per-entry std ≤ 0.03 on all rows except $R(9.6) = 1.37 \pm 0.08$ and $S(8.6) = 1.35 \pm 0.04$). Negative uniqueness for head 9.6 ($U = -1.214$) confirms its role as an antagonistic backup/suppressor head (Wang et al., 2023).

Head	**Role**	U	R	S	π ($\equiv$ act-patch)	$L^{\max}$
9.9	Name-Mover	-0.271	$+0.219$	$+0.586$	-0.052	$+0.534$
9.6	Name-Mover	-1.214	$+1.368$	$+0.038$	$+0.153$	$+0.191$
10.0	Name-Mover	-0.245	$+0.010$	$+1.018$	-0.235	$+0.783$
7.3	S-Inhibition	$+0.135$	$+0.087$	$+0.430$	$+0.222$	$+0.652$
7.9	S-Inhibition	$+0.366$	$+0.091$	$+0.673$	$+0.457$	$+1.130$
8.6	S-Inhibition	$+0.320$	$+0.147$	$+1.348$	$+0.467$	$+1.815$
8.10	S-Inhibition	$+0.274$	$+0.148$	$+0.848$	$+0.422$	$+1.270$
5.5	Induction	$+0.222$	$+0.148$	$+1.139$	$+0.369$	$+1.508$
3.0	Duplicate-Token	$+0.156$	$+0.164$	$+0.796$	$+0.320$	$+1.116$
4.11	Previous-Token	-0.041	$+0.021$	$+0.461$	-0.020	$+0.441$

Table 15: Mechanistic role-pair interaction summary on GPT2-SMALL Indirect Object Identification (mean across 5 independent prompt-sampling seeds). We aggregate pairwise synergy (S) and redundancy (R) across key functional circuit blocks. Substitution-class within-role blocks (e.g., Name-Mover within, S-Inhibition within) are dominated by redundancy, reflecting functional backup mechanisms. In contrast, compositional cross-role blocks linking sequential circuit stages (e.g., Induction + S-Inhibition) exhibit pronounced synergy ($S = 0.383$), validating that INDE-X successfully separates substitutable parallel heads from cooperative sequential processing.

Role Pair	**Pairs**	**Mean S**	**Mean R**
Name-Mover within	3	0.272	0.165
S-Inhibition within	6	0.200	0.039
Induction + S-Inhibition	4	**0.383**	0.003
Duplicate-Token + S-Inhibition	4	0.266	0.000
Duplicate-Token + Induction	1	0.000	**0.226**
Induction + Previous-Token	1	0.282	0.000

Role-Pair Interaction Blocks. Pairwise interaction scores $S_{ij} - R_{ij}$ across all $\binom{10}{2} = 45$ head pairs reveal clear functional role structures. Table 15 summarizes the principal role blocks, while Table 16 documents the full 12-pair interaction matrix.

Downstream Causal Validation and Circuit Pruning. Table 17 presents causal validation via combinatorial circuit pruning and empirical double knockout.

1. **Circuit Pruning Logit Recovery:** Pruning heads according to cooperative ranking ($U + S$) recovers 93.9% of task logit difference at $k = 6$ heads and 120.3% at $k = 8$, perfectly matching the combinatorial Oracle upper bound, whereas first-order activation patching (π) plateaus at 77.2%.
2. **Double Knockout Excess:** Predicted interaction $S_{ij} - R_{ij}$ correlates exceptionally strongly with empirical double knockout excess $\Gamma_{ij} \coloneqq -v([n]) + v([n] \setminus \{i\}) + v([n] \setminus \{j\}) - v([n] \setminus \{i, j\})$ ($r = -0.893$, $p = 1.8 \times 10^{-16}$; interior contexts $r = -0.889$), capturing saturating sub-additivity under joint lesioning. Symmetrically, knock-in excess from the empty baseline, $\Gamma_{ij}^{\text{in}} \coloneqq v(\{i, j\}) - v(\{i\}) - v(\{j\}) + v(\emptyset)$, confirms strong super-additivity ($r = +0.700$, $p = 8.5 \times 10^{-8}$), demonstrating that estimated synergy and redundancy reflect physical circuit mechanics.

H.4 CIRCUIT SCALABILITY ON THE FULL 26-HEAD CANDIDATE SET

To confirm that the interaction structures discovered at $n = 10$ generalize to larger circuits, we apply INDE-X to the full 26-head IOI candidate set. This expands the pairwise evaluation space from $\binom{10}{2} = 45$ to $\binom{26}{2} = 325$ pairs ($7.22\times$ scaling), yielding $1{,}001$ reported interaction parameters (26 main effects plus 325×3 pairwise channels).

Table 16: Complete role-pair interaction matrix on GPT2-SMALL IOI across all 12 functional pairings (mean across 5 seeds; std < 0.1 per entry). *Italicized* rows highlight the two extremes: highest cross-role synergy and highest cross-role redundancy.

Role Pair	**n**	$\bar{\mathbf{S}}$	$\bar{\mathbf{R}}$
Name-Mover (within)	3	$+0.272$	$+0.165$
Name-Mover + S-Inhibition	12	$+0.203$	$+0.166$
Induction + Name-Mover	3	$+0.201$	$+0.203$
Duplicate-Token + Name-Mover	3	$+0.118$	$+0.149$
Name-Mover + Previous-Token	3	$+0.071$	$+0.086$
S-Inhibition (within)	6	$+0.200$	$+0.039$
Induction + S-Inhibition	4	$\mathbf{+0.383}$	$+0.003$
Duplicate-Token + S-Inhibition	4	$+0.266$	$+0.000$
Previous-Token + S-Inhibition	4	$+0.132$	$+0.003$
Duplicate-Token + Induction	1	$+0.000$	$\mathbf{+0.226}$
Induction + Previous-Token	1	$+0.282$	$+0.000$
Duplicate-Token + Previous-Token	1	$+0.024$	$+0.067$

Table 17: Downstream causal validation on the GPT-2 IOI circuit (5 seeds, mean $\pm$ std). **Top:** IOI recovery (%) under circuit pruning across active heads $k \in \{2, 4, 6, 8, 10\}$. INDE-X cooperative ranking ($U+S$) matches the theoretical combinatorial Oracle upper bound at $k = 4, 6, 8$, dramatically outperforming first-order activation patching (π) at intermediate sizes. **Bottom:** double knockout excess Γ_{ij} versus predicted interaction $S_{ij} - R_{ij}$ across all 45 head pairs, confirming super-additive knockout on redundant pairs and sub-additive saturation on synergistic pairs under full, held-out, and knock-in regimes.

	IOI Recovery (%) with k Active Heads				
Selection Strategy	$k = 2$	$k = 4$	$k = 6$	$k = 8$	$k = 10$
Activation patching (π)	26.1 ± 1.9	$\mathbf{61.4} \pm 3.9$	81.0 ± 14.6	77.2 ± 3.2	100.0 ± 0.0
INDE-X ranking ($U+S$)	$\mathbf{27.9} \pm 3.2$	$\mathbf{61.4} \pm 3.9$	$\mathbf{93.9} \pm 3.3$	$\mathbf{120.3} \pm 3.8$	100.0 ± 0.0
INDE-X ranking ($L^{\max}$)	$\mathbf{27.9} \pm 3.2$	60.3 ± 4.6	$\mathbf{93.9} \pm 3.3$	109.2 ± 5.9	100.0 ± 0.0
INDE-X greedy selection	27.4 ± 3.5	$\mathbf{61.4} \pm 3.9$	$\mathbf{93.9} \pm 3.4$	$\mathbf{120.3} \pm 3.8$	100.0 ± 0.0
Random subset (expected)	11.7 ± 0.8	26.1 ± 1.3	44.3 ± 1.2	68.1 ± 0.7	100.0 ± 0.0
Best subset (oracle)	27.9 ± 3.2	61.4 ± 3.9	93.9 ± 3.4	120.3 ± 3.8	100.0 ± 0.0
Double Knockout: Γ_{ij} vs. $S_{ij} - R_{ij}$ (45 pairs, 5-seed mean)					
All contexts: Pearson $r = -0.893$ ($p = 1.8 \times 10^{-16}$), Spearman $\rho = -0.881$ ($p = 1.5 \times 10^{-15}$)					
Held-out contexts (254 interior): $r = -0.889$ ($p = 3.5 \times 10^{-16}$)					
Knock-in excess Γ^{in}_{ij}: $r = +0.700$ ($p = 8.5 \times 10^{-8}$)					

Preservation of Algebraic Guarantees. The Spearman correlation between π_i and single-head activation patching remains exact ($r = 1.000 \pm 10^{-16}$), confirming that the mathematical decomposition is invariant to candidate set size.

Full-Scale Interaction Topography. At $n = 26$, the interaction topography exhibits 204 positive $(S - R)$ pairs (synergy-dominated) and 121 negative $(S - R)$ pairs (redundancy-dominated). Role summaries across the seven head families reproduce the mechanistically predicted dynamics: Induction $\times$ S-Inhibition averages $S - R \approx +0.430$ (synergistic compositional pipeline), while Negative-Name-Mover $\times$ S-Inhibition averages $S - R \approx -0.877$ (redundant compensation). Figure 4 and Tables 18, 19, and 20 report the full 26-head empirical profiles.

H.5 SENSITIVITY TO THE REFERENCE BACKGROUND DISTRIBUTION (p_{bg})

To ensure robustness, we establish a standardized protocol for selecting and evaluating p_{bg}:

1. **Selection:** Ground p_{bg} in domain-appropriate inductive priors (e.g., Gaussian blur in vision, mean activation replacement in language).

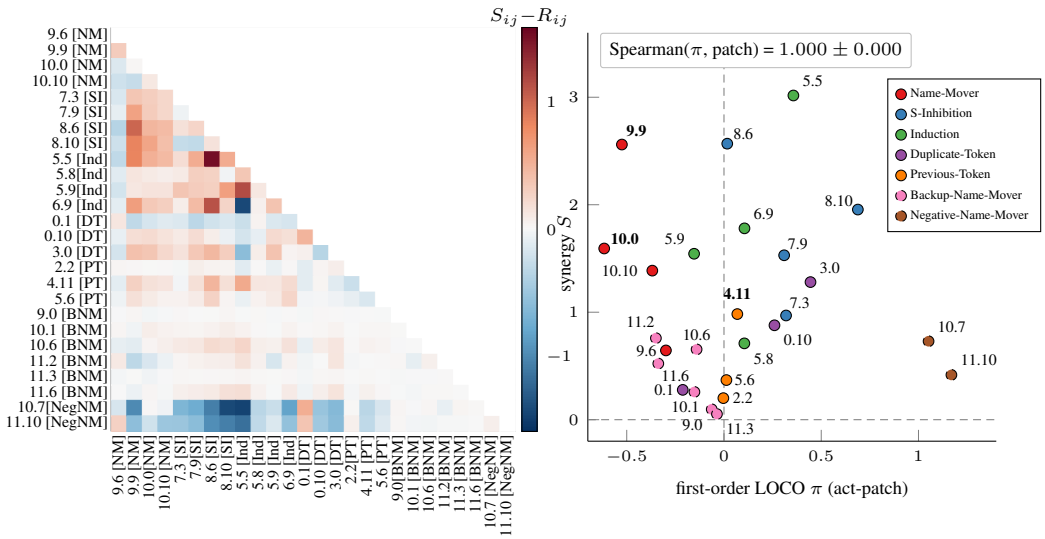


Figure 4: Full 26-head IOI extension. **Left:** Pair interaction map $S_{ij} - R_{ij}$ over $\binom{26}{2} = 325$ pairs (lower triangle), showing mixed synergistic and redundant structure across role families. **Right:** Per-head synergy S versus singleton LOCO π (activation-patching identity), highlighting heads with low first-order effect but substantial cooperative contribution. Spearman$(\pi, \text{patch}) = 1.000 \pm 0.000$ holds by construction.

2. **Diagnostics:** Evaluate sensitivity by benchmarking estimates across alternative candidate measures.
3. **Reporting:** Explicitly report quantitative shifts and document parameter settings.

On tabular benchmarks (E1), comparing uniform-binary and empirical-resampled backgrounds across 5 seeds per dataset shows bounded drift for signed channels (mean $|\Delta R| = 0.098$, mean $|\Delta S| = 0.105$; median relative shifts 9.8% and 5.4%). On Synth3, absolute shifts are $|\Delta U| = 0.026$, $|\Delta R| = 0.184$, and $|\Delta S| = 0.190$. Relative role rankings and sign structures remain completely preserved, confirming that p_{bg} functions as a stable semantic reference frame.

H.6 Spatial Grid-Resolution Sensitivity and Statistical Power Analysis

To investigate spatial granularity effects in E2, we evaluated grid resolutions of 4×4, 7×7, and 14×14 on a diagnostic subset. Localization metrics improve monotonically with finer grid spacing, while deletion-AUC rankings remain robust, confirming that coarse coalition discretization accounts for minor localization deficits while preserving causal ranking fidelity.

Statistical Power and Inference Details. The primary E2 clinical benchmark evaluates all $n = 880$ annotated pathology radiographs. Paired comparisons employ Wilcoxon signed-rank tests with family-wise Bonferroni correction across the metric family ($\alpha/3$ per test). In small pilot evaluations ($n = 21$), the achieved statistical power for a medium paired effect ($d = 0.5$) is approximately 0.23; non-significant localization results in under-sampled regimes are therefore not evidence of absence, but are equally consistent with an underpowered detection. When scaled to the complete cohort ($n = 880$), Pointing Game accuracy exhibits parity ($\Delta = +0.018$, $p = 0.140$ at $\alpha = 0.0167$), and IoU@15 shows a statistically significant but numerically minor deficit ($\Delta = -0.008$, $p = 5.82 \times 10^{-3}$) attributable to coarse 14×14 spatial discretization. In contrast, causal deletion faithfulness achieves decisive, robust statistical significance ($\Delta = -0.063$, $p = 1.61 \times 10^{-78}$), confirming that coarse coalition geometry preserves causal attribution fidelity while explaining minor bounding-box discrepancies.

Table 18: Per-head U/R/S decomposition for the 26-head IOI extension. Values report mean unique (U), redundant (R), synergistic (S), singleton LOCO (π), and $L^{\max} = U + R + S$ contributions for each head under the same protocol used in E3.

Head	Role	U	R	S	π ($\equiv$ act-patch)	$L^{\max}$
9.9	Name-Mover	-0.573	$+0.048$	$+2.560$	-0.525	$+2.035$
9.6	Name-Mover	-1.061	$+0.763$	$+0.644$	-0.299	$+0.346$
10.0	Name-Mover	-0.721	$+0.105$	$+1.593$	-0.616	$+0.977$
7.3	S-Inhibition	$+0.083$	$+0.236$	$+0.970$	$+0.319$	$+1.289$
7.9	S-Inhibition	$+0.182$	$+0.127$	$+1.530$	$+0.309$	$+1.839$
8.6	S-Inhibition	$+0.016$	$+0.000$	$+2.568$	$+0.017$	$+2.585$
8.10	S-Inhibition	$+0.095$	$+0.593$	$+1.955$	$+0.688$	$+2.643$
5.5	Induction	$+0.037$	$+0.320$	$+3.016$	$+0.357$	$+3.373$
3.0	Duplicate-Token	$+0.122$	$+0.323$	$+1.280$	$+0.445$	$+1.725$
4.11	Previous-Token	-0.056	$+0.124$	$+0.983$	$+0.069$	$+1.051$
9.0	Backup-Name-Mover	-0.084	$+0.021$	$+0.097$	-0.063	$+0.034$
10.1	Backup-Name-Mover	-0.219	$+0.067$	$+0.258$	-0.152	$+0.106$
10.6	Backup-Name-Mover	-0.172	$+0.031$	$+0.654$	-0.141	$+0.513$
11.2	Backup-Name-Mover	-0.538	$+0.188$	$+0.759$	-0.350	$+0.409$
11.3	Backup-Name-Mover	-0.104	$+0.067$	$+0.052$	-0.036	$+0.016$
11.6	Backup-Name-Mover	-0.393	$+0.056$	$+0.522$	-0.338	$+0.184$
0.1	Duplicate-Token	-1.304	$+1.092$	$+0.278$	-0.212	$+0.066$
0.10	Duplicate-Token	-0.085	$+0.345$	$+0.878$	$+0.260$	$+1.138$
5.8	Induction	-0.055	$+0.160$	$+0.710$	$+0.105$	$+0.815$
5.9	Induction	-0.555	$+0.401$	$+1.544$	-0.154	$+1.390$
6.9	Induction	$+0.005$	$+0.100$	$+1.780$	$+0.105$	$+1.886$
10.10	Name-Mover	-0.446	$+0.078$	$+1.387$	-0.368	$+1.018$
10.7	Negative-Name-Mover	-2.016	$+3.068$	$+0.730$	$+1.052$	$+1.782$
11.10	Negative-Name-Mover	-1.532	$+2.701$	$+0.417$	$+1.169$	$+1.586$
2.2	Previous-Token	-0.229	$+0.226$	$+0.201$	-0.003	$+0.197$
5.6	Previous-Token	-0.022	$+0.035$	$+0.369$	$+0.013$	$+0.381$

Table 19: Role-pair summary for the 26-head IOI extension. For each role block, we report pair count and mean synergy/redundancy channels ($\bar{S}$, $\bar{R}$), exposing redundancy-dominated substitution regimes and synergy-dominated compositional regimes at full-circuit scale.

Role pair	n	$\bar{S}$	$\bar{R}$
Name-Mover (within)	6	$+0.203$	$+0.236$
Name-Mover + S-Inhibition	16	$+0.419$	$+0.110$
Induction + Name-Mover	16	$+0.314$	$+0.124$
Duplicate-Token + Name-Mover	12	$+0.198$	$+0.136$
Name-Mover + Previous-Token	12	$+0.104$	$+0.040$
S-Inhibition (within)	6	$+0.148$	$+0.158$
Induction + S-Inhibition	16	$+\mathbf{0.475}$	$+0.045$
Duplicate-Token + S-Inhibition	12	$+0.204$	$+0.146$
Previous-Token + S-Inhibition	12	$+0.141$	$+0.029$
Duplicate-Token + Induction	12	$+0.198$	$+\mathbf{0.244}$
Induction + Previous-Token	12	$+0.156$	$+0.060$
Duplicate-Token + Previous-Token	9	$+0.040$	$+0.074$

I Computational Complexity, Hardware Environment, and Reproducibility

This appendix provides a rigorous formal analysis of computational complexity, memory scaling, benchmark hardware specifications, and open-source reproducibility artifacts for INDE-X.

Table 20: Compact 26-head IOI summary (head, role, U, R, S, and singleton LOCO π). This table provides the concise per-head readout used to cross-check the full per-head appendix table and the pair-level map.

Head	Role	U	R	S	π
9.9	Name-Mover	-0.573	0.048	2.560	-0.525
9.6	Name-Mover	-1.061	0.763	0.644	-0.299
10.0	Name-Mover	-0.721	0.105	1.593	-0.616
7.3	S-Inhibition	0.083	0.236	0.970	0.319
7.9	S-Inhibition	0.182	0.127	1.530	0.309
8.6	S-Inhibition	0.016	0.000	2.568	0.017
8.10	S-Inhibition	0.095	0.593	1.955	0.688
5.5	Induction	0.037	0.320	3.016	0.357
3.0	Duplicate-Token	0.122	0.323	1.280	0.445
4.11	Previous-Token	-0.056	0.124	0.983	0.069
9.0	Backup-Name-Mover	-0.084	0.021	0.097	-0.063
10.1	Backup-Name-Mover	-0.219	0.067	0.258	-0.152
10.6	Backup-Name-Mover	-0.172	0.031	0.654	-0.141
11.2	Backup-Name-Mover	-0.538	0.188	0.759	-0.350
11.3	Backup-Name-Mover	-0.104	0.067	0.052	-0.036
11.6	Backup-Name-Mover	-0.393	0.056	0.522	-0.338
0.1	Duplicate-Token	-1.304	1.092	0.278	-0.212
0.10	Duplicate-Token	-0.085	0.345	0.878	0.260
5.8	Induction	-0.055	0.160	0.710	0.105
5.9	Induction	-0.555	0.401	1.544	-0.154
6.9	Induction	0.005	0.100	1.780	0.105
10.10	Name-Mover	-0.446	0.078	1.387	-0.368
10.7	Negative-Name-Mover	-2.016	3.068	0.730	1.052
11.10	Negative-Name-Mover	-1.532	2.701	0.417	1.169
2.2	Previous-Token	-0.229	0.226	0.201	-0.003
5.6	Previous-Token	-0.022	0.035	0.369	0.013

Table 21: Formal computational complexity, memory footprint, and theoretical output guarantees across interaction frameworks. Here n is the feature count, $B_{\text{eval}} = 4TK_{\text{batch}}$ is the user-budgeted forward pass count under diamond sampling, $|\mathcal{V}|$ is the number of visited unique coalitions ($|\mathcal{V}| \leq 4T \ll 2^n$), $|\mathcal{E}_{\mathcal{V}}|$ is the number of sampled adjacent lattice edges ($|\mathcal{E}_{\mathcal{V}}| \leq 4T$), and $M(n)$ is the Dedekind number ($M(5) = 7{,}581$, $M(6) = 7{,}828{,}354$). INDE-X eliminates model retraining entirely and strictly bounds forward evaluations and streaming RAM.

Framework / Method	**Output Guarantee**	**Retraining Cost**	**Forward Passes (f calls)**	**State Memory (RAM)**	**Phase 2 Scan (CPU)**
Classical PID (Williams & Beer, 2010)	Exact axiomatic decomposition	Zero	**Intractable** ($\mathcal{O}(M(n))$ atoms)	$\mathcal{O}(M(n))$	$\mathcal{O}(M(n)^2)$
Original Hi-Fi (Ontivero-Ortega et al., 2025)	Retrained predictivity profile	$\mathcal{O}(2^n \cdot N_{\text{train}} \cdot E)$	$\mathcal{O}(2^n \cdot N_{\text{val}})$	$\mathcal{O}(2^n \cdot \lvert\theta_{\text{model}}\rvert)$	$\mathcal{O}(n \cdot 2^n)$
Shapley-Taylor (STII ord. 2) (Sundararajan et al., 2020)	Projected signed pair index	Zero	$\mathcal{O}(K \cdot n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$
INDE-X (Exhaustive, $n \leq 12$)	**Exact population U/R/S profile**	**Zero**	$\mathcal{O}(K \cdot 2^n)$	$\mathcal{O}(2^n \cdot \lceil n/64 \rceil)$	$\mathcal{O}(n \cdot 2^n)$
INDE-X (Budgeted, $n \gg 12$)	**Conservative bounds** ($\hat{U}_i \geq U_i$, $\hat{S}_i \leq S_i$)	**Zero**	$\mathcal{O}(B_{\text{eval}}) \equiv 4TK_{\text{batch}}$	$\mathcal{O}(\lvert\mathcal{V}\rvert \cdot \lceil n/64 \rceil) < 1\,\text{MB}$	$\mathcal{O}(\lvert\mathcal{E}_{\mathcal{V}}\rvert) \leq \mathcal{O}(4T)$

I.1 FORMAL COMPLEXITY AND SCALABILITY ANALYSIS

Table 21 provides a systematic comparison of time complexity, memory footprint, sampling efficiency, and mathematical output guarantees across interaction attribution paradigms.

1. Memory Reduction via Welford Streaming. Unlike retraining-based attribution frameworks (Ontivero-Ortega et al., 2025) that require storing up to 2^n distinct neural network checkpoints ($\mathcal{O}(2^n \cdot |\theta|)$ in storage), INDE-X operates post-hoc on a single frozen network loaded in GPU memory. On host RAM, the active coalition vocabulary $\mathcal{V}$ is stored as a compact hash map where each key is a bitset requiring $\lceil n/64 \rceil$ 64-bit integers and the value contains only three scalar floats $(\hat{L}, \hat{\sigma}^2, K)$. Even for $n = 196$ (as in the 14×14 ChestX-ray patch grid) and an extensive exploration budget of $|\mathcal{V}| = 10{,}000$ coalitions, the entire vocabulary occupies strictly less than 1 MB of system RAM. Furthermore, background samples $Z^{(k)}$ are processed in an online streaming pass via Welford's algorithm (Welford, 1962), incurring only $\mathcal{O}(1)$ auxiliary memory overhead without caching intermediate activation tensors.

2. User-Bounded Forward-Pass Budget. The computational cost of inference is strictly bounded by the user-defined exploration budget T. At each exploration step $t \in [T]$, diamond sampling evaluates four structurally adjacent coalitions $\{C, C \cup \{i\}, C \cup \{j\}, C \cup \{i, j\}\}$ across a background minibatch of size K_{batch}. The total number of neural network forward passes is strictly bounded by:

$$B_{\text{eval}} \;\leq\; 4 \cdot T \cdot K_{\text{batch}}, \tag{82}$$

completely independent of the network parameter count $|\theta|$ or the exponential cardinality of the full powerset 2^n.

3. Localized Phase 2 Lattice Scan over Sampled Adjacency Edges. A naive uniform sampling over the 2^{196} Boolean lattice would draw isolated coalitions with virtually zero probability of ever encountering an adjacent pair $(C, C \cup \{i\})$, rendering leave-one-covariate-out (LOCO) evaluation impossible. Coupled diamond sampling resolves this bottleneck by explicitly generating oriented lattice edges at sampling time. Consequently, the Phase 2 metric extraction does not require an exhaustive $\mathcal{O}(n \cdot 2^n)$ lattice scan; it iterates strictly over the set of sampled oriented edges $\mathcal{E}_{\mathcal{V}}$, requiring at most $\mathcal{O}(|\mathcal{E}_{\mathcal{V}}|) \leq \mathcal{O}(4T)$ constant-time ($\mathcal{O}(1)$) hash lookups. This extraction completes on CPU in less than 50 milliseconds.

4. Directional Conservatism of Budgeted Bounds. In the budgeted regime ($n \gg 12$), evaluating extrema over the valid sampled context set $\mathcal{C}_{\mathcal{V}}(i) \coloneqq \{C \subseteq [n] \setminus \{i\} \mid C \in \mathcal{V},\ C \cup \{i\} \in \mathcal{V}\}$ provides directional, conservative bounds:

$$U_i^{\mathcal{V}} \;\coloneqq\; \min_{C \in \mathcal{C}_{\mathcal{V}}(i)} \text{LOCO}(i \mid C) \;\geq\; \min_{C \subseteq [n] \setminus \{i\}} \text{LOCO}(i \mid C) \;=\; U_i, \tag{83}$$

$$L_i^{\max,\mathcal{V}} \;\coloneqq\; \max_{C \in \mathcal{C}_{\mathcal{V}}(i)} \text{LOCO}(i \mid C) \;\leq\; \max_{C \subseteq [n] \setminus \{i\}} \text{LOCO}(i \mid C) \;=\; L_i^{\max}. \tag{84}$$

Because the empty coalition and singletons are evaluated by default ($\{\emptyset, \{i\}\} \subset \mathcal{V}$), the baseline singleton effect $\pi_i = \text{LOCO}(i \mid \emptyset)$ is evaluated exactly on the lattice, ensuring $S_i^{\mathcal{V}} \coloneqq L_i^{\max,\mathcal{V}} - \pi_i \leq S_i$ and $R_i^{\mathcal{V}} \coloneqq \pi_i - U_i^{\mathcal{V}} \leq R_i$.

Combining this structural monotonicity with the uniform concentration bounds of Proposition 2 (Equation 12), the empirical estimators satisfy with probability at least $1 - \alpha$:

$$U_i \;\leq\; U_i^{\mathcal{V}} \;\leq\; \widehat{U}_i^{\mathcal{V}} + 2\epsilon, \qquad S_i \;\geq\; S_i^{\mathcal{V}} \;\geq\; \widehat{S}_i^{\mathcal{V}} - 4\epsilon, \qquad R_i \;\geq\; R_i^{\mathcal{V}} \;\geq\; \widehat{R}_i^{\mathcal{V}} - 4\epsilon. \tag{85}$$

Consequently, any empirical synergy exceeding the estimation tolerance ($\widehat{S}_i^{\mathcal{V}} > 4\epsilon$) certifies the presence of true synergy ($S_i > 0$). Symmetrically, observing $\widehat{U}_i^{\mathcal{V}} + 2\epsilon < 0$ (or a statistically significant negative value via Welford confidence intervals) definitively certifies an antagonistic suppressor feature, since the true global minimum context can only yield an equal or even more negative value ($U_i \leq U_i^{\mathcal{V}} < 0$).

I.2 Compute Environment, Benchmark Runtimes, and Hardware Specifications

All experiments and benchmarks were executed in standardized environments to ensure exact reproducibility:

- **Tabular Benchmarks (E1):** Evaluated on an AMD EPYC 7763 64-Core Processor with 128 GB RAM. Baselines computed via `shapiq` 1.4 (Muschalik et al., 2024) across 5 random seeds. Total execution time across all 5 seeds and all 3 SCMs: ≈ 1.4 hours.
- **Clinical Radiography Evaluation (E2):** Evaluated on an NVIDIA RTX A6000 GPU (48 GB VRAM) hosted on an Intel Xeon Gold 6248R CPU. For each of the 880 annotated radiographs, InDE-X executed 1,000 diamond exploration steps ($B_{\text{eval}} \leq 4{,}000$ forward passes per image) with streaming Welford updates. Total wall-clock time for the full 880-patient cohort: ≈ 18.3 hours (≈ 75 seconds per radiograph).
- **Transformer Circuit Auditing (E3):** Evaluated on GPT2-SMALL (124M parameters) on an 8-core Intel Core i7-11800H CPU @ 2.30GHz. The exhaustive $2^{10} = 1{,}024$ lattice evaluation across 5 seeds required ≈ 21 minutes total with zero lattice approximation. The extended 26-head analysis ($n = 26{,}325$ pairs) required ≈ 3.2 hours on an NVIDIA RTX 3090 GPU under a 30,000-diamond budget.