CRIQ: FEATURE INTERACTION QUANTIFICATION IN DEEP NEURAL NETWORKS WITH COORDINATE-RAY INTEGRALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Higher-order game-theoretic interactions reveal critical neural network behaviors. However, existing black-box approximations suffer from combinatorial subset sampling bottleneck in high dimensions. To address this limitation, we introduce Coordinate-Ray Interaction Quantification (CRIQ), a scalable framework for interaction estimation in deep neural networks that maps discrete derivatives to coordinate-ray integrals. To resolve non-differentiability in continuous piecewise-linear (CPWL) models and bypass high-order tensor differentiation in feature embeddings, we also introduce Iterated Dimensional Reduction (IDR), which decomposes discrete derivatives into an alternating sum of 1D first-order derivative integrals. We derive CRIQ's sample complexity, propose an estimation algorithm and measure its convergence stability empirically, showing that CRIQ makes feature-level interaction quantification in high-dimensional neural networks computationally feasible.

1 INTRODUCTION

With the call for transparency (Adadi & Berrada, 2018; Dewitte & Smuha, 2025) and related regulations (European Parliament, 2016-05-04), feature attribution methods based on cooperative game theory have become crucial in interpreting model predictions (Rozemberczki et al., 2022). However, recent research has shown that individual feature contributions are insufficient to capture the full model behavior (Covert et al., 2020; Muschalik et al., 2024b): feature *interactions* carry key information in real-world tasks like natural language processing (Chen et al., 2020; Singhvi et al., 2025; Kang et al., 2025; Butler et al., 2026). Interaction indices (IIs) (Fujimoto et al., 2006), which measure the synergies and redundancies between groups of features, can offer a much richer understanding of a model's decision-making process.

Despite their importance, computing higher-order game-theoretic interactions in deep neural networks is computationally expensive. Existing approaches primarily operate within a model-agnostic paradigm, extending classic perturbation-based attribution to interactions to improve quality (Sundararajan et al., 2020; Bordt & von Luxburg, 2023; Tsai et al., 2023) or computational efficiency (Fumagalli et al., 2023; Muschalik et al., 2024a; Kolpaczki et al., 2024). By treating the neural network as a black box, these methods calculate interactions from model queries across combinatorial subsets. This black-box setup leads to sample variance that scales with feature dimension (Figure 1). This scaling issue often forces practitioners to use heuristic pre-clustering (such as superpixel algorithm on images), which can introduce inductive biases.

To address this gap, we introduce **Coordinate-Ray Interaction Quantification** (**CRIQ**), a framework that evaluates atomic game interactions using mixed partial derivatives. CRIQ leverages the Multivariate Fundamental Theorem of Calculus (MFTC) to show that discrete derivatives are equal to integrals of mixed partial derivatives over hyper-rectangles. By parameterizing the intervals along orthogonal coordinate rays, CRIQ evaluates feature interactions without combinatorial subset evaluations, allowing scaling to high-dimensional models. To resolve vanishing higher order derivatives in continuous piecewise-linear architectures and avoid high-order tensor differentiation in multidimensional embeddings, we also propose **Iterated Dimensionality Reduction** (IDR), which reduces discrete derivatives into an alternating sum of 1D integrals over 1st order derivatives.

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.

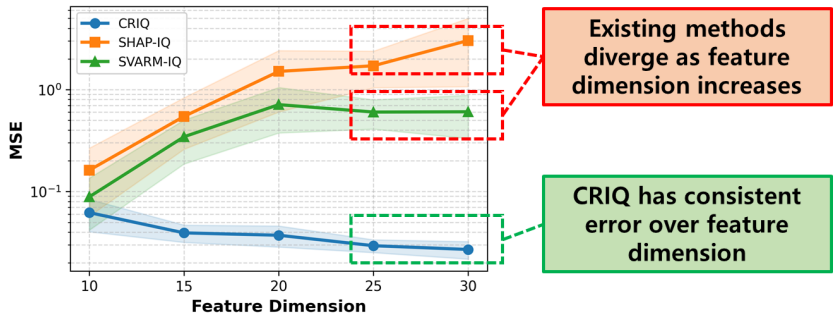

Figure 1: Illustration of CRIQ versus SHAP-IQ and SVARM-IQ on unanimity game neural network as number of features n increases while the query budget stays constant at $Q = 256$ for 2nd order Shapley Interaction Index. The mean squared error of SHAP-IQ increases with n, while that of CRIQ stays relatively constant.

The main contributions of this work are as follows:

- We establish the identity mapping model-induced discrete derivatives to continuous mixed partial derivative integrals over hyper-rectangles, resolving atomic game-theoretic interactions without subset-based evaluation.
- We address vanishing higher order derivatives and higher order tensor differentiation for interaction quantification by projecting higher order discrete derivatives onto integrals over 1st order derivatives.
- We derive CRIQ's sample complexity, develop an efficient estimation algorithm via reverse-mode amortization, and empirically demonstrate its performance over query budget and feature dimensions.

2 Related Works

Cooperative Game Theory. Feature attribution based on cooperative game theory represent the target model as a game and input features as players, interpreting the Shapley value as feature contribution to model prediction (Lundberg & Lee, 2017). The formulation has been generalized to higher-order Cardinal Interaction Indices (CIIs), including Shapley Interaction Index (Grabisch & Roubens, 1999), n-Shapley values (Bordt & von Luxburg, 2023), Shapley-Taylor Interaction Index (Sundararajan et al., 2020), and FaithShap (Tsai et al., 2023). Since these indices require evaluating combinatorial subsets of features, existing approaches such as SHAP-IQ (Fumagalli et al., 2023) and SVARM-IQ (Kolpaczki et al., 2024) treat the model as a black box and approximate the interactions with Monte Carlo sampling. This setup induces background variance that scales with feature dimension, making their calculations intractable in higher dimensions.

Sparse Fourier Expansions. Recent works reformulate game-theoretic interactions as Boolean Fourier transform (Gorji et al., 2025; Kang et al., 2025; Butler et al., 2026). Approaches such as SPEX (Kang et al., 2025) and ProxySPEX (Butler et al., 2026) leverage sparse recovery algorithms or train tree-based surrogates to identify the largest interaction terms. While effective, these approaches rely on sparsity assumptions that may not hold across deep architectures.

Gradient-based Feature Attribution. Integrated Gradients (IG) (Sundararajan et al., 2017) formulate feature attribution as nonatomic cooperative game, calculating Aumann-Shapley value by integrating derivatives along 1D diagonal trajectory from a reference baseline to the input. IG have been extended to different paths (Kapishnikov et al., 2021; Zhuo & Ge, 2024) or Hessians (Janizek et al., 2021). Despite the common characteristic of using derivatives, these methods address *nonatomic* games, which fundamentally differ from atomic games that our method is trying to resolve for neural networks.

Our method differs from existing methods by using mixed partial derivative representation of discrete derivatives, allowing for the estimation of *atomic* game-theoretic interactions (as opposed to *nonatomic* games) in high-dimensional models without requiring dimensionality-reducing preprocessing or interaction sparsity assumptions.

3 PRELIMINARIES

3.1 COOPERATIVE GAME THEORY

A cooperative game is defined as $G(N, v)$, where $N = \{1, 2, \ldots, n\}$ denotes n players and $v : 2^N \to \mathbb{R}$ is a payoff function such that $v(\emptyset) = 0$. For any coalition $S \subseteq N$, $v(S)$ represents the total payoff achieved by the members of S.

In cooperative game theory, player i's *value* is the weighted average of their *marginal contribution*, the change in v from i joining S or $\Delta_{\{i\}}v(S) = v(S \cup \{i\}) - v(S)$ (Shapley et al., 1953). More generally, the interaction between a subset of players $T \subseteq N$ is the weighted average of their *marginal interaction*, also known as *discrete derivatives* (Grabisch & Roubens, 1999).

Definition 1 *For a target coalition $T \subseteq N$ of cardinality $k = |T| \geq 1$, the discrete derivative $\Delta_T v(S)$ with respect to coalition $S \subseteq N \setminus T$ is:*

$$\Delta_T v(S) = \sum_{U \subseteq T} (-1)^{k-|U|} v(U \cup S) \tag{1}$$

A Cardinal Interaction Index (CII) is the weighted average of $\Delta_T v(S)$ across background cardinalities $s = |S|$ (Fujimoto et al., 2006):

$$I_{\text{CII}}(T) = \sum_{s=0}^{n-k} W(s) \sum_{\substack{S \subseteq N \setminus T \\ |S|=s}} \Delta_T v(S) \tag{2}$$

In particular, if $T = \{i\}$ and $W(s) = \frac{s!(n-k-s)!}{(n-k+1)!}$, we retrieve the Shapley value. Different $W(s)$ and T result in different CIIs, such as Banzhaf Interaction (Marichal & Mathonet, 2011), Shapley-Taylor Interaction (Sundararajan et al., 2020), FaithSHAP (Tsai et al., 2023), and n-Shapley (Bordt & von Luxburg, 2023).

3.2 MULTILINEAR EXTENSIONS

Cooperative games can be also analyzed through multilinear extension (MLE) (Owen, 1975):

Definition 2 *The multilinear extension $F : [0, 1]^n \to \mathbb{R}$ of game v is the continuous multilinear polynomial defined on the unit hypercube:*

$$F(u) = \sum_{S \subseteq N} \left(\prod_{j \in S} u_j \prod_{l \in N \setminus S} (1 - u_l) \right) v(S) = \mathbb{E}_{S \sim Ber(u)}[v(S)] \tag{3}$$

where $u = (u_1, \ldots, u_n) \in [0, 1]^n$ represents each player's coalition-joining probability.

Given $F(u)$, its partial derivatives with respect to u_T is equivalent to the expectation of discrete derivative $\Delta_T v(S)$.

Theorem 1 *For a coalition $T \subseteq N$ with $k = |T| \geq 1$ and probability vector $u \in [0, 1]^n$, the k-th order mixed partial derivative of $F(u)$ with respect to u_T equals the expected discrete derivative of game v over the background coalitions S:*

$$F_T(u) = \frac{\partial^k F(u)}{\prod_{j \in T} \partial u_j} = \sum_{S \subseteq N \setminus T} \left(\prod_{j \in S} u_j \prod_{l \in (N \setminus T) \setminus S} (1 - u_l) \right) \Delta_T v(S) \tag{4}$$

In particular, the right hand side can be interpreted as an expectation with respect to probabilities u_{-T}, i.e., $F_T(u) = \mathbb{E}_{S\sim Ber(u_{-T})}[\Delta_T v(S)]$.

A Probabilistic Interaction Index (PII) is calculated by integrating $F_T(u)$ along path $u_{-T}(t) \in [0,1]^{n-k}$ under a probability measure $d\mu(t)$ (Fujimoto et al., 2006):

$$I_{\text{PII}}(T) = \int_0^1 F_T(u(t))d\mu(t) = \int_0^1 d\mu(t)\, \mathbb{E}_{S\sim \text{Ber}(u_{-T})}[\Delta_T v(S)] \tag{5}$$

A PII becomes a CII if the probabilities are symmetric, i.e., $u_j(t) = g(t) \forall j \in N \setminus T$. In particular, setting $d\mu(t) = dt$ and $u_j(t) = t \forall j \in N \setminus i$ retrieves the Shapley value.

3.3 Cooperative Game Theory for Feature Attribution

In feature attribution, we represent a trained model as a game and its inputs as players. Specifically, let $f : \mathbb{R}^n \to \mathbb{R}$ be a k-differentiable model mapping input instance $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ to an output $y = f(x)$. We adopt the *interventional* (or baseline-based) framework, where feature absences are simulated by replacing feature values with reference values $R = (R_1, \ldots, R_n) \in \mathbb{R}^n$ drawn from a background distribution $\mathcal{D}$ (Lundberg et al., 2020; Sundararajan & Najmi, 2020). Setting feature indices to players $N = \{1, 2, \ldots, n\}$, each feature subset $S \subseteq N$, target instance x, and reference vector R determine a hybrid coalition vector $Z_S(x, R) \in \mathbb{R}^n$:

$$(Z_S(x,R))_i = \begin{cases} x_i & \text{if } i \in S \\ R_i & \text{if } i \notin S \end{cases} \tag{6}$$

When x and R are unambiguous, we write $Z_S \equiv Z_S(x, R)$ for brevity. Taking expectation over $\mathcal{D}$ induces the model-induced game $v : 2^N \to \mathbb{R}$:

$$v(S) = \mathbb{E}_{R\sim\mathcal{D}}\big[f(Z_S(x,R))\big] - \mathbb{E}_{R\sim\mathcal{D}}\big[f(R)\big] \tag{7}$$

Since expectation is a linear operator, it commutes with the discrete derivative operator. Therefore, $\Delta_T v(S)$ of the model-induced game can be written as:

$$\Delta_T v(S) = \mathbb{E}_{R\sim\mathcal{D}}\big[\Delta_T f(Z_S(x,R))\big] \tag{8}$$

where $\Delta_T f(Z_S(x,R)) = \sum_{U\subseteq T}(-1)^{k-|U|} f(Z_{U\cup S}(x,R))$. We refer to this term as model-induced discrete derivative.

3.4 Fundamental Theorem of Calculus

The Multivariate Fundamental Theorem of Calculus (MFTC) is defined as follows (Aksoy & Martelli, 2002; Owen, 2005; Bár, 2024):

Theorem 2 *Let $B = \prod_{m=1}^k [a_m, b_m] \subset \mathbb{R}^k$ be a k-dimensional axis-aligned hyper-rectangle, and let g be a k-times continuously differentiable function on an open domain $\Omega \supset B$. The integral of the k-th order mixed partial derivative of g over B is equal to the alternating sum of g evaluated at the 2^k vertices of B:*

$$\int_{a_1}^{b_1} \cdots \int_{a_k}^{b_k} \frac{\partial^k g(z_1, \ldots, z_k)}{\partial z_1 \ldots \partial z_k}\, dz_1 \ldots dz_k = \sum_{U\subseteq\{1,\ldots,k\}} (-1)^{k-|U|} g(w^U) \tag{9}$$

where each vertex vector $w^U \in \mathbb{R}^k$ is defined by:

$$w_m^U = \begin{cases} b_m & \text{if } m \in U, \\ a_m & \text{if } m \notin U \end{cases} \tag{10}$$

The alternating sum in Theorem 2 is equivalent to the discrete derivative $\Delta_T f(Z_S(x,R))$.

4 METHODOLOGY

We now introduce *Coordinate-Ray Interaction Quantification* (CRIQ). We use MFTC to show that the model-induced discrete derivative $\Delta_T f(Z_S(x,R))$ can be represented as the integral of mixed partial derivative of f. We then combine this form of $\Delta_T(S)$ with the definitions of CII and PII to estimate the interactions.

4.1 DISCRETE DERIVATIVES AS INTEGRAL OF MIXED PARTIAL DERIVATIVES

Based on MFTC, $\Delta_T f(Z_S(x,R))$ is equal to the integral of the mixed partial derivative for T by setting $g(z_T) = f(z_T, Z_S(x,R)_{-T})$ on $\prod_{m=1}^{k}[R_{i_m}, x_{i_m}]$.

Proposition 1 *Let f be k-continuously differentiable. By Theorem 2, we have the following identity between $\Delta_T f(Z_S(x,R))$ and the integral of mixed partial derivative with respect to feature $i \in T$ in the hyper-rectangle $\prod_{m=1}^{k}[R_{i_m}, x_{i_m}]$:*

$$\int_{R_{i_1}}^{x_{i_1}} \cdots \int_{R_{i_k}}^{x_{i_k}} \frac{\partial^k f(z_T, Z_S(x,R)_{-T})}{\prod_{j\in T} \partial z_j}\, dz_{i_1} \ldots dz_{i_k} = \Delta_T f(Z_S(x,R)) \tag{11}$$

The full proof is in Appendix C. We can rewrite the Equation 11 in terms of a *coordinate-ray integral*. Given basis vectors $\mathbf{e}_{i_m} \in \mathbb{R}^k$ for features $i_m \in T$, the coordinate ray is defined as:

$$r_T(\xi; Z_S) = Z_S + \sum_{m=1}^{k} \xi_m (x_{i_m} - R_{i_m}) e_{i_m} \tag{12}$$

Applying the transformation turns the integral into a product of the signed volume of the hyper-rectangle and the coordinate-ray integral.

$$\Delta_T f(Z_S) = \left(\prod_{j\in T}(x_j - R_j)\right) \int_{[0,1]^k} \frac{\partial^k f\big(\mathbf{r}_T(\boldsymbol{\xi}; Z_S)\big)}{\prod_{j\in T} \partial z_j}\, d\xi_1 \ldots d\xi_k \tag{13}$$

4.2 ITERATED DIMENSIONALITY REDUCTION (IDR)

While Proposition 1 shows the equivalence between discrete derivatives and mixed partial derivative integrals, it has two potential hurdles in deep learning architectures: (i) in Continuous Piecewise-Linear (CPWL) networks (e.g., ReLU activations), where higher-order derivatives vanish almost everywhere ($\nabla^k f(z) = 0$ a.e. for $k \geq 2$), and (ii) in models with multidimensional vector inputs (such as word token embeddings), where computing higher-order derivatives requires tensor evaluations. To address these challenges, we introduce Iterated Dimensionality Reduction (IDR), which applies 1D Fundamental Theorem of Calculus recursively along coordinate rays, projecting higher-order discrete derivatives onto 1st order derivatives.

Proposition 2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous. For coalition $T \subseteq N$ with $k = |T| \geq 1$, context $S \subseteq N \setminus T$, reference $R \sim D$, and a chosen pivot feature $p \in T$, the model-induced discrete derivative decomposes into an alternating sum of 2^{k-1} 1D integrals of first-order derivative:*

$$\Delta_T f(Z_S(x,R)) = (x_p - R_p) \sum_{V \subseteq T\setminus\{p\}} (-1)^{k-1-|V|} \int_0^1 \frac{\partial f}{\partial z_p}\Big(r_p\big(\xi; Z_{V\cup S}(x,R)\big)\Big)\, d\xi \tag{14}$$

where $r_p(\xi; Z_{V\cup S}) = Z_{V\cup S} + \xi(x_p - R_p)e_p$ is the 1D coordinate ray along pivot p for $\xi \in [0,1]$.

Crucially, IDR is invariant to the choice of p. The full proof is provided in Appendix D. See Appendix A for feature embedding extension.

4.3 Evaluation of Probabilistic and Cardinal Interaction Indices

We can substitute the integral formulation of $\Delta_T f(Z_S)$ into the definitions of CII and PII to find the model-induced CII and PII.

Cardinal Interaction Indices (CII). The substitution results in:

$$I_{\text{CII}}(T) = \sum_{s=0}^{n-k} W(s) \sum_{\substack{S \subseteq N \setminus T \\ |S|=s}} \Delta_T f(Z_S) = \sum_{s=0}^{n-k} W(s) \binom{n-k}{s} \mathbb{E}_{\substack{S \sim \mathcal{U}_s \\ R \sim \mathcal{D}}} \left[\Delta_T f(Z_S) \right] \tag{15}$$

where $\mathcal{U}_s$ is the uniform distribution over subsets of $N \setminus T$ with size s. The Monte Carlo estimate $\hat{I}_{\text{CII}}(T)$ is calculated by generating samples (s, S, R) following $s \sim P(s)$, $S \sim \mathcal{U}_S$, and $R \sim D$, evaluating the mixed partial derivative, and performing weighted average with weight $K_{\text{weight}} = \frac{W(s)\binom{n-k}{s}}{P(s)}$. In particular, if $P(s) = W(s)\binom{n-k}{s}$, $K_{\text{weight}} = 1$.

Probabilistic Interaction Indices (PII). The substitution results in:

$$I_{\text{PII}}(T) = \int_0^1 d\mu(t)\, \mathbb{E}_{\substack{S \sim \text{Ber}(u_{-T}(t)) \\ R \sim \mathcal{D}}} \left[\Delta_T f\big(Z(S,R)\big) \right] \tag{16}$$

We can get the Monte Carlo estimate $\hat{I}_{\text{PII}}(T)$ by sampling (t, S, R) following $t \sim \mu(t)$, $S \sim Ber(u_{-T}(t))$, and $R \sim D$, evaluating the mixed partial derivative, and averaging the results.

The pseudo-code for both methods is shown in Appendix B.

4.4 Convergence Guarantees

We establish theoretical guarantees on the sample complexity of CRIQ as follows.

Proposition 3 *For coalition $T \subseteq N$ with $k = |T| \geq 1$, let the maximum feature displacement $D_{\max} = \sup_{R \in \mathcal{D},\, i \in N} |x_i - R_i|$, maximum mixed partial derivative $G_{\max}^{(k)} = \sup_{z \in B} \left| \frac{\partial^k f(z)}{\prod_{j \in T} \partial z_j} \right| < \infty$, and the maximum sample weight multiplier $K_{\max} = \sup_s |K_{weight}(s)|$. For $\epsilon > 0$ and $\delta \in (0,1)$, the estimator $\widehat{I}_M(T)$ of interaction index $I(T)$ with Monte Carlo sample budget M converges at $\mathbb{P}\big(|\widehat{I}_M(T) - I(T)| \geq \epsilon\big) \leq \delta$ provided that M satisfies:*

$$M \geq \frac{2C^2 \ln(2/\delta)}{\epsilon^2} \tag{17}$$

where $C = K_{\max}(D_{\max})^k G_{\max}^{(k)}$ using mixed partial derivatives and $K_{\max} D_{\max} 2^{k-1} G_{\max}^{(1)}$ using IDR.

The full proof is presented in Appendix E. A key aspect of Proposition 3 is that for a single coalition T, the sample bound is independent of n, especially if $K_{\max} = 1$ (for PII and CII where $S \sim P(s)$).

4.5 Practical Global Scaling with Reverse-Mode Amortization.

Practical pipelines require quantifying all $\binom{n}{k}$ interactions simultaneously. A naive Monte Carlo implementation using M from Proposition 3 would require $\mathcal{O}(M\binom{n}{k})$ evaluations, which becomes intractable at large n. In practice, we use the reverse mode automatic differentiation to amortize evaluations into shared backward passes. Specifically, we implement a 1-point estimator by anchoring the derivatives at the base state ($\xi = 0$) to share identical evaluation points over all $n-1$ candidate features, enabling the simultaneous update of a set of interactions in two backward passes per pivot configuration while preserving underlying game distributions (see Appendix B.1). Under this estimation, bounding $\mathbb{P}(\max_{T \in \mathcal{T}} |\hat{I}(T) - \mathbb{E}[\hat{I}(T)]| \geq \epsilon) \leq \delta$ over $\mathcal{T} = \{T | T \subseteq N, |T| = k\}$ requires a global query budget Q:

$$Q \geq \frac{2^k \binom{n}{k-1}}{k\epsilon^2} C^2 \left(\ln(2/\delta) + k \ln n\right) \tag{18}$$

The total estimation error satisfies $|\hat{I}(T) - I(T)| \leq \epsilon + \mathcal{E}_{\text{quad}}$, where $\mathcal{E}_{\text{quad}}$ is a quadrature discretization error. We emphasize that the underlying CRIQ formulation (Propositions 1 and 2) is an exact identity ($\mathcal{E}_{\text{quad}} = 0$). Our current implementation deliberately trades the asymptotic zero-bias at exhaustive budgets ($Q \sim 2^n$) for variance reduction in low-budget regime ($Q \ll 2^n$). For settings where asymptotic convergence to zero error is strictly required, $\mathcal{E}_{\text{quad}}$ can be eliminated by replacing 1-point estimator with standard M_ξ-point Gauss-Legendre quadrature estimator along rays at the cost of $2M_\xi$ backward passes (Appendix B.1).

5 Experiments

In this section, we empirically validate the theoretical properties of the proposed method and compare against several baseline methods. For brevity, we focus on Shapley interaction index of order $k = 2$ in the main text. We report the results for Shapley value ($k = 1$) and Banzhaf Interaction Indices ($k = 1, 2$) in Appendix F.

5.1 Metrics and Models

Metrics. We investigate asymptotic convergence (with increasing Q) and dimension divergence (with increasing n) using average MSE and ground truth rank correlation, where MSE over $\mathcal{T} = \{T | T \subseteq N, |T| = k\}$ is:

$$\text{MSE} = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \left(\widehat{I}(T) - I(T)\right)^2 \tag{19}$$

and ground truth rank correlation is:

$$\hat{\rho} = \rho(\widehat{I}(T), I(T)) \tag{20}$$

where ρ is the Spearman rank correlation.

Models and Tasks. We use the following models for the experiments:

- Synthetic CPWL Network: A 3-layer fully connected Multi-Layer Perceptron (MLP) with ReLU activations.
- DistilBERT: (Sanh et al., 2019). A smaller counterpart of BERT model that takes a sentence as the input.
- Unanimity Network: A ReLU network explicitly designed to mimic a unanimity game.

We perform synthetic classification for CPWL network, sentence sentiment analysis for DistilBERT, and value prediction for unanimity network. We evaluate 2nd order SII for logit of each model. In particular, for CPWL and DistilBERT, we have to calculate the ground truth values combinatorially with 2^n coalitions, so we limit $n \leq 12$ to compute exactly. For unanimity game, the exact SII can be calculated for higher n analytically, so we set $n = 30$ for convergence bound and $n = 10, 15, 20, 25, 30$ for dimension-free bound to test for scalability. All quantitative evaluations are performed over 5 instances averaged over 10 independent random seed initializations per instance for a total of 50 evaluations.

Explanation Methods. We evaluate CRIQ using both CII and PII setup with our 1-point estimator, and benchmark against several methods in the `shapiq` package (Muschalik et al., 2024a): SHAP-IQ (Fumagalli et al., 2023), SVARM-IQ (Kolpaczki et al., 2024), and SPEX (Kang et al., 2025). We do not include Integrated Gradients (Sundararajan et al., 2017) or Hessian (Janizek et al., 2021)

as they evaluate *nonatomic* games, which axiomatically diverge from the atomic cooperative games evaluated by our proposed approach.

The complete experimental details are presented in Appendix A.

5.2 Quantitative Evaluation

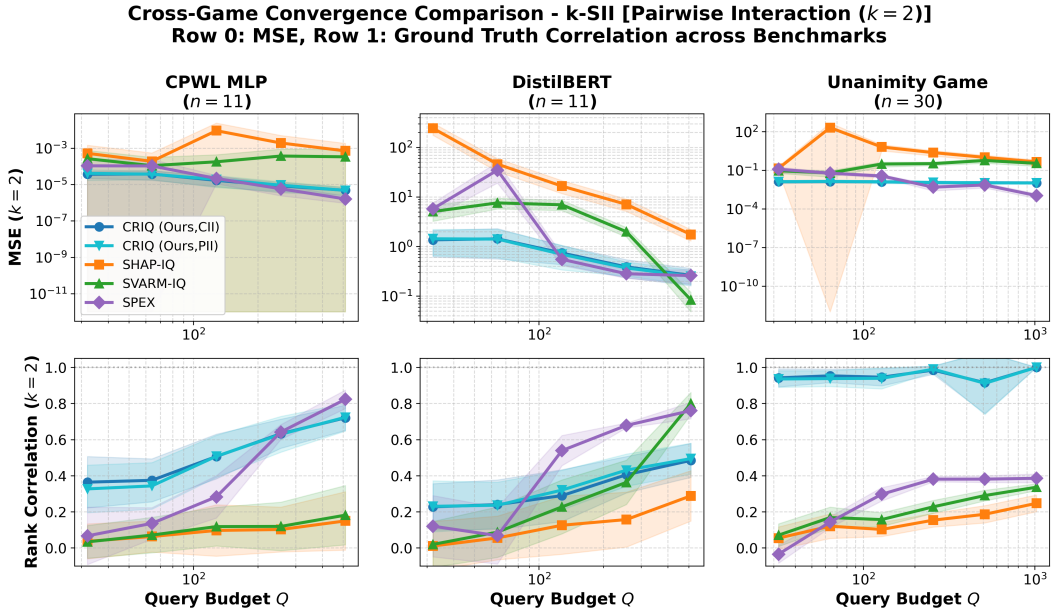

Figure 2: MSE and rank correlation of 2nd order Shapley Interaction Index estimated by CRIQ and baseline methods over query budget Q for unanimity network ($n = 30$), MLP network ($n = 11$) and DistilBERT ($n = 11$). The shaded bound indicates 1 standard deviation over all evaluations. CRIQ has lower MSE and higher correlation at low Q before converging asymptotically at high Q, while others converge to zero by design as Q approaches 2^n.

Convergence over query budget. Figure 2 presents 2nd order SII results over Q with ± 1 standard deviation bounds. Across all architectures, CRIQ tends to achieve better rank correlation and MSE in low-budget regime ($Q \ll 2^n$). At high budgets ($Q \sim 2^n$), black-box baselines converge toward zero MSE and correlation of 1 since they enumerate unique subsets, whereas CRIQ exhibits a bounded error floor due to discretization bias (Appendix B.1).

Robustness to feature dimension. In Figure 3, we report the results for fixed Q and increasing n across target models. We find that with fixed Q, the baseline MSE increase and rank correlations decrease with n. This degradation can be attributed to combinatorial dispersion: as the power set grows, the fixed budget covers an increasingly sparse fraction of the coalition space, inflating variance. In contrast, the estimation error of CRIQ remains stable across n. This stability occurs because, while Equation 18 establishes an $\mathcal{O}(n \ln n)$ global budget to bound the worst-case error across all T simultaneously, average MSE is driven by variance, which is bound independent of n from Proposition 3. CRIQ also naturally benefits from structural gradient decoupling in deep architectures without imposing any prior sparsity constraints. Under IDR, when two features (i, j) do not functionally interact within a model, toggling feature i produces no shift in the gradient along j ($\Delta g_j \equiv 0$), evaluating the interaction to zero with zero sample variance. Consequently, uncoupled features contribute zero variance to the global interaction estimate, preventing error accumulation as n increases.

Comparison against SPEX. We also observe distinct behavior against SPEX governed by interaction sparsity and model curvature. While both methods tend to outperform SHAP-IQ and

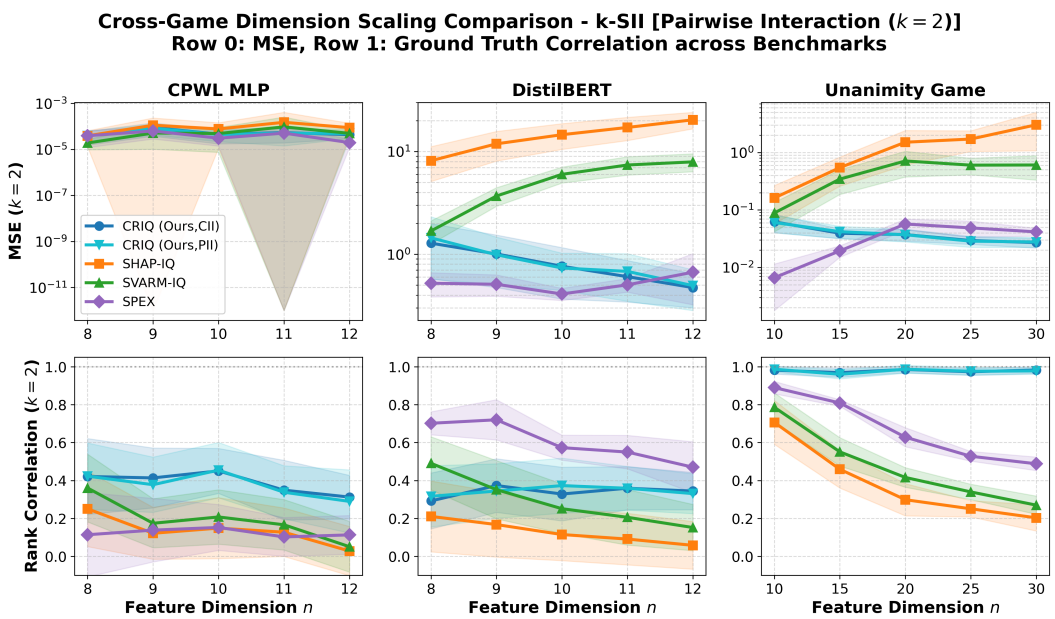

Figure 3: MSE and rank correlation of 2nd order Shapley Interaction Index estimated by CRIQ and baseline methods over feature dimension n for unanimity network, MLP network and DistilBERT for fixed Q. The shaded bound indicates 1 standard deviation over all evaluations. CRIQ has stable MSE and correlation, while others diverge as n increases.

SVARM-IQ, SPEX achieves higher rank correlation in DistilBERT at high Q and low n, whereas CRIQ outperforms SPEX on the unanimity game network. This difference reflects each method's estimation strategy. SPEX enforces an explicit sparsity prior that aligns well with localized interactions in language models, but becomes misspecified when true interactions are dense like unanimity games, decreasing SPEX's accuracy as n increases (Figure 3). In contrast, CRIQ requires no sparsity assumptions, guaranteeing dimension-independent variance bounds while passively capturing architectural decoupling. The performance gap in ground truth correlation in DistilBERT may also be attributed to the discretization bias under our 1-point estimator. For a mild discretization floor, the estimator enables CRIQ to scale to larger n (as seen by near-constant correlation in Figure 3 over n in DistilBERT).

6 Conclusion

In this work, we introduce Coordinate-Ray Interaction Quantification (CRIQ), a framework for estimating atomic game-theoretic interactions in high-dimensional models using coordinate-ray integrals. By applying the Multivariate Fundamental Theorem of Calculus, CRIQ maps discrete derivatives to integrals of continuous mixed partial derivatives, enabling efficient evaluation with automatic differentiation. For continuous piecewise-linear models and high-dimensional embeddings, we also introduce Iterated Dimensional Reduction (IDR), decomposing discrete derivatives into a sum of 1D integrals of 1st order derivatives. By providing a scalable, mathematically grounded foundation for quantifying higher-order feature interactions in deep neural networks, CRIQ opens the door for the analysis of complex models in real-world applications in terms of atomic cooperative game theory. Future research directions include optimizing the estimation algorithm and leveraging CRIQ interactions for mechanistic interpretability and model editing in neural networks.

AI Use Statement

In this work, we used generative AI tools for implementing the code for the experiments and for improving the readability of text. We have not used generative AI tools for other tasks with required

or recommended disclosure. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the full experimental settings in Appendix A, the pseudocode of the estimation algorithm in Appendix B, and all proofs in Appendices C through E.

REFERENCES

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Asuman Aksoy and Mario Martelli. Mixed partial derivatives and fubini's theorem. *The College Mathematics Journal*, 33(2):126–130, 2002.

Filip Bár. The fundamental theorem of calculus in higher dimensions. *arXiv preprint arXiv:2402.14564*, 2024.

Sebastian Bordt and Ulrike von Luxburg. From shapley values to generalized additive models and back. In *AISTATS*, 2023.

Landon Butler, Abhineet Agarwal, Justin Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. Proxyspex: Inference-efficient interpretability via sparse feature interactions in llms. *Advances in Neural Information Processing Systems*, 38:72306–72340, 2026.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5578–5593, 2020.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in neural information processing systems*, 33:17212–17223, 2020.

Pierre Dewitte and Nathalie Smuha. Ai meets the gdpr. navigating the impact of data protection on ai systems. 2025.

European Parliament. Regulation 2016/679 of the european parliament and of the council of april 27, 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *OJ*, L 119:1–88, 2016-05-04.

Katsushige Fujimoto, Ivan Kojadinovic, and Jean-Luc Marichal. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55 (1):72–99, 2006.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. Shap-iq: Unified approximation of any-order shapley interactions. *Advances in Neural Information Processing Systems*, 36:11515–11551, 2023.

Ali Gorji, Andisheh Amrollahi, and Andreas Krause. Shap values via sparse fourier representation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 18827–18860. Curran Associates, Inc., 2025. doi: 10.52202/085713-0634. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1b331c20064e37e204a5bcd12481bfac-Paper-Conference.pdf.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4):547–565, 1999.

Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54, 2021.

Justin Singh Kang, Landon Butler, Abhineet Agarwal, Yigit Efe Erginbas, Ramtin Pedarsani, Bin Yu, and Kannan Ramchandran. SPEX: Scaling feature interaction explanations for LLMs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=pRlKbAwczl.

Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 5048–5056. IEEE, 2021.

Patrick Kolpaczki, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier. SVARM-IQ: Efficient approximation of any-order Shapley interactions through stratification. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3520–3528. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/kolpaczki24a.html.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

Jean-Luc Marichal and Pierre Mathonet. Weighted banzhaf power and interaction indexes through weighted approximations of games. *European journal of operational research*, 211(2):352–358, 2011.

Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. shapiq: Shapley interactions for machine learning. *Advances in Neural Information Processing Systems*, 37:130324–130357, 2024a.

Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier. Beyond treeshap: Efficient computation of any-order shapley interactions for tree ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14388–14396, 2024b.

Art B Owen. Multidimensional variation for quasi-monte carlo. In *International Conference on Statistics in honour of Professor Kai-Tai Fang's 65th birthday*, pp. 49–74. World Scientific, 2005.

Guillermo Owen. Multilinear extensions and the banzhaf value. *Naval research logistics quarterly*, 22(4):741–750, 1975.

Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. The shapley value in machine learning. In *The 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence*, pp. 5572–5579. International Joint Conferences on Artificial Intelligence Organization, 2022.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Lloyd S Shapley et al. A value for n-person games. 1953.

Divyansh Singhvi, Diganta Misra, Andrej Erkelens, Raghav Jain, Isabel Papadimitriou, and Naomi Saphra. Using shapley interactions to understand how models use structure. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20727–20737, 2025.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9269–9278. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sundararajan20b.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International conference on machine learning*, pp. 9259–9268. PMLR, 2020.

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

Yue Zhuo and Zhiqiang Ge. Ig 2: Integrated gradient on iterative gradient path for feature attribution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7173–7190, 2024.

A EXPERIMENTAL SETTINGS

In this section, we provide the full specification of all empirical benchmarks.

A.1 HARDWARE, SOFTWARE, AND HYPERPARAMETERS

Hardware and Software. All benchmarks are executed on a machine equipped with an 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30GHz processor, 32GB RAM, and NVIDIA GeForce RTX 3080 GPU, running Windows 11 Education. For core library dependencies, we use `numpy` (v2.4.6), `scipy` (v1.18.0), `torch` (v2.13.0), `transformers` (v5.15.0), and `shapiq` (v1.7.0).

Hyperparameters. Across all benchmarks, each configuration is evaluated on 5 test instances over 10 independent trials for a total of 50 evaluations, with hyperparameters in Table 1. To maintain consistent terminology with black-box literature, we define the query budget Q as the total count of model calls. For black-box baselines, one call corresponds to a single forward pass. For CRIQ, one call corresponds to a single backward pass.

Table 1: Hyperparameter settings. All settings are evaluated across 5 independent instances with 10 random trials each. For the CPWL MLP, with $N_{\text{bg}} = 4$ reference samples, the effective coalition budget is $Q//4$.

MODEL	EVALUATION	DIMENSION (n)	QUERY BUDGET (Q)
DistilBERT	Convergence	11	32, 64, 128, 256, 512
	Dimension Scaling	8, 9, 10, 11, 12	128
CPWL MLP	Convergence	11	32, 64, 128, 256, 512
	Dimension Scaling	8, 9, 10, 11, 12	64
Unanimity Game	Convergence	30	32, 64, 128, 256, 512, 1024
	Dimension Scaling	10, 15, 20, 25, 30	256

Seeds. Each benchmark follows the following seed schedule based on master seed $S_0 = 42$:

- Model initialization: All neural networks are initialized with seed $S_{\text{weights}} = S_0 + n$, where n is feature dimension.
- Convergence benchmark: For instance index i, trial index t, and budget Q:

$$S_{\text{trial}}(i, t, Q) = S_0 + 1\,000 \cdot i + 50 \cdot t + Q.$$

- Dimension scaling benchmark: For dimension index d (i.e, the position of the dimension being tested), instance index i, and trial index t:

$$S_{\text{trial}}(d, i, t) = S_0 + 10\,000 \cdot d + 1\,000 \cdot i + 50 \cdot t.$$

Across all benchmarks, attribution accuracy is quantitatively scored against ground truth using mean squared error and rank correlation.

A.2 NATURAL LANGUAGE PROCESSING BENCHMARK.

Target Model. For natural language processing model, we use the pretrained DistilBERT (`distilbert-base-uncased-finetuned-sst-2-english` checkpoint). Our target function $f(x)$ is:

$$f(x) = \text{logit}_{\text{positive}}(x) - \text{logit}_{\text{negative}}(x).$$

The reference baseline token embedding is uniform padding token: $x_{\text{bg}} = [\text{[PAD]}]^n$. The integral is performed continuously between the [PAD] token embedding and original token's embedding.

Continuous Integration over Token Embeddings. In Transformer architectures like DistilBERT, each token is represented with a multidimensional embedding relative to [PAD] embedding. The 1D coordinate ray parameterization generalizes to vector spaces with the chain rule. Let embeddings of token i and [PAD] be $\mathbf{x}_i, \mathbf{R}_i \in \mathbb{R}^d$. Then, the linear path for token i is parameterized by $\mathbf{z}_i(\xi) = \mathbf{R}_i + \xi(\mathbf{x}_i - \mathbf{R}_i)$ for $\xi \in [0, 1]$. Differentiating model f with respect to ξ yields:

$$\frac{d}{d\xi} f(\mathbf{z}_i(\xi), Z_{-i}) = \langle \nabla_{\mathbf{z}_i} f(\mathbf{z}_i(\xi), Z_{-i}),\, \mathbf{x}_i - \mathbf{R}_i \rangle = \sum_{c=1}^{d} (x_{i,c} - R_{i,c}) \frac{\partial f}{\partial z_{i,c}} \tag{21}$$

By Iterated Dimensional Reduction (Proposition 2), the interaction between tokens i and j evaluates to the inner product between the token displacement vector and the gradient difference:

$$\Delta_{\{i,j\}} f(Z_S) = \int_0^1 \langle \mathbf{x}_j - \mathbf{R}_j,\, \nabla_{\mathbf{z}_j} f(r_j(\xi; Z_{S\cup\{i\}})) - \nabla_{\mathbf{z}_j} f(r_j(\xi; Z_S)) \rangle \, d\xi \tag{22}$$

Equation 22 ensures that token-level interactions satisfy exact discrete derivative equivalence regardless of embedding dimensionality d.

Input Sentences. We use the following sentences for DistilBERT experiments. For the dimension scaling experiments, each sentence is truncated to the first n words:

1. *"this charming lovely film has truly wonderful great emotional narrative depth and remarkable brilliance"*
2. *"a masterfully crafted movie with brilliant directing delivering hilarious comedy and truly witty dialogue"*
3. *"an incredibly touching story featuring remarkable beautiful scenes provide gripping cinematic masterpiece performance throughout"*
4. *"the fantastic screenplay delivers witty dialogue combined with superb captivating drama performance and intensity"*
5. *"spectacular visuals combined with magnificent exquisite acting provide classic captivating entertainment loved show throughout always"*

A.3 Continuous Piecewise Linear (CPWL) Deep ReLU Networks

Architecture and Initialization. All CPWL benchmarks utilize multilayer perceptrons with ReLU activations:

- Fixed dimension ($n = 11$): Used for query budget convergence experiment. A 3-hidden-layer network structured as:

$$\mathbb{R}^{11} \xrightarrow{\text{Linear}} \mathbb{R}^{48} \xrightarrow{\text{ReLU}} \mathbb{R}^{48} \xrightarrow{\text{ReLU}} \mathbb{R}^{48} \xrightarrow{\text{ReLU}} \mathbb{R}^{1} \quad \text{(bias-free output layer)}.$$

- Dynamic dimension: Used for dimension scaling experiment. The hidden layer dimension scales with input dimension n as $h(n) = \max(32, 4n)$:

$$\mathbb{R}^{n} \xrightarrow{\text{Linear}} \mathbb{R}^{h(n)} \xrightarrow{\text{ReLU}} \mathbb{R}^{h(n)} \xrightarrow{\text{ReLU}} \mathbb{R}^{h(n)} \xrightarrow{\text{ReLU}} \mathbb{R}^{1}.$$

Input Instances and Background Samples Target inputs $x \in \mathbb{R}^n$ are drawn uniformly over $[-1.25, 1.25]^n$, with clipping to prevent degenerate activations ($|x_i| \geq 0.2$). For each instance, we sample $N_{\text{bg}} = 4$ baselines by adding Gaussian white noise with standard deviation $\sigma = 0.35$ to the instance.

To ensure budget parity, the baseline methods are evaluated per background with a coalition budget of $Q_S = Q//N_{\text{bg}}$ queries, such that the total number of model calls equals Q for all methods.

A.4 UNANIMITY GAME RELU NETWORK

To evaluate scalability to larger dimensions where full combinatorial evaluation is intractable, we construct a specialized ReLU neural network to represent a unanimity game.

Let $S \subseteq N = \{1, \ldots, n\}$ denote a coalition. A unanimity game is defined by the logical conjunction:

$$u_S(z) = \prod_{i \in S} z_i = \begin{cases} 1, & \text{if } S \subseteq z, \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

Between a target $x = \mathbf{1}_n$ and baseline $x_{\text{bg}} = \mathbf{0}_n$, any binary coalition mask $z \in \{0, 1\}^n$ can evaluate this conjunction with a single-neuron ReLU threshold:

$$u_S(z) = \text{ReLU}\left(\sum_{i \in S} z_i - (|S| - 1)\right) \tag{24}$$

When all features in S are active ($\sum_{i \in S} z_i = |S|$), the argument equals 1 and the ReLU outputs 1. If at least a single feature is missing, the output is zero. The full synthetic game is formed as a linear combination of H carrier conjunctions:

$$f(z) = \sum_{k=1}^{H} w_k \, u_{S_k}(z) \tag{25}$$

The interaction indices for a unanimity game can be calculated as:

$$I_{\text{CII}}(T) = \sum_{h: T \subseteq S_h} w_h \left[\sum_{j=0}^{n-|S_h|} \binom{n - |S_h|}{j} W\big(j + |S_h| - |T|\big) \right] \tag{26}$$

We configure the model as follows:

- S_k **cardinality:** $|S_k|$ is constrained to $|S_k| \in [2, 4]$.
- **Number of S_k:** In the fixed-dimension benchmark ($n = 30$), $H = 18$. In the dimension-scaling benchmark, $H = \max(12, \lfloor n/3 \rfloor)$.
- S_k **Weights:** Weights $w_k = s_k \cdot \alpha_k$, where signs $s_k \in \{-1.0, 1.0\}$ and magnitudes $\alpha_k \sim \mathcal{U}(0.6, 1.4)$

B Pseudocode of Monte Carlo Estimate of Cardinal and Probabilistic Interaction Indices

The pseudo-code of the proposed method is as follows.

Algorithm 1 Coordinate-Ray Interaction Quantification (CRIQ)

Require: Model f, Input instance $x \in \mathbb{R}^n$, Target subset $T \subseteq N$ ($|T| = k$), Baseline distribution $\mathcal{D}$, Index specification (PII with trajectory $(\mathbf{u}_{-T}, \mu)$ or CII with weights $W(s)$ and sampling distribution $P(s)$), Query budget M, Architecture type $\mathcal{A} \in \{\texttt{Smooth}, \texttt{CPWL}\}$.

Initialize interaction accumulator $\widehat{I} \leftarrow 0$
for $m = 1$ **to** M **do**
  Sample baseline reference vector $R \sim \mathcal{D}$
  if Index is PII **then**
    Sample trajectory parameter $t \sim \mu(t)$ and draw mask $S \sim \text{Bernoulli}(\mathbf{u}_{-T}(t))$
    Set importance weight multiplier $K_{\text{weight}} \leftarrow 1.0$
  else ▷ Index is CII
    Sample cardinality $s \sim P(s)$ and draw context $S \subseteq N \setminus T$ with $|S| = s$ uniformly at random
    Compute sample weight multiplier $K_{\text{weight}} \leftarrow \frac{W(s)\binom{n-k}{s}}{P(s)}$ ▷ $\equiv 1.0$ for Shapley under uniform $P(s)$
  end if
  $\widehat{\Delta}_T \leftarrow$ ComputeDiscreteDerivative$(f, x, R, T, S, \mathcal{A})$ ▷ Algorithm 2
  $\widehat{I} \leftarrow \widehat{I} + \frac{K_{\text{weight}}}{M}\widehat{\Delta}_T$
end for
return $\widehat{I}$

Algorithm 2 Discrete Derivative Estimation ComputeDiscreteDerivative$(f, x, R, T, S, \mathcal{A})$

Require: Model f, Input x, Baseline R, Target subset $T = \{i_1, \ldots, i_k\}$, Context subset $S \subseteq N \setminus T$, Architecture $\mathcal{A} \in \{\texttt{Smooth}, \texttt{CPWL}\}$.

Form base context state $Z_S \in \mathbb{R}^n$: $(Z_S)_j \leftarrow x_j$ if $j \in S$, and $(Z_S)_j \leftarrow R_j$ if $j \notin S$ ▷ $(Z_S)_j = R_j$ for all $j \in T$
if $\mathcal{A} = \texttt{Smooth}$ **then** ▷ k-times continuously differentiable architectures
  Sample integration coordinates $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k) \sim \text{Uniform}([0,1]^k)$
  Construct coordinate-ray evaluation point:
  $\mathbf{r}_T(\boldsymbol{\xi}; Z_S) \leftarrow Z_S + \sum_{m=1}^{k} \xi_m (x_{i_m} - R_{i_m}) \mathbf{e}_{i_m}$
  Compute mixed partial derivative with automatic differentiation:
  $g_k \leftarrow \left. \frac{\partial^k f}{\prod_{j \in T} \partial z_j} \right|_{\mathbf{r}_T(\boldsymbol{\xi}; Z_S)}$
  return $\left(\prod_{j \in T} (x_j - R_j)\right) \cdot g_k$
else ▷ $\mathcal{A} = \texttt{CPWL}$: Iterated Dimensional Reduction
  Select pivot coordinate $p \in T$
  Initialize directional gradient accumulator $\Sigma_{\text{rays}} \leftarrow 0$
  for each sub-coalition $V \subseteq T \setminus \{p\}$ **do**
    Form vertex state $Z_{V \cup S} \in \mathbb{R}^n$: $(Z_{V \cup S})_j \leftarrow x_j$ if $j \in V \cup S$, and $(Z_{V \cup S})_j \leftarrow R_j$ if $j \notin V \cup S$
    Evaluate first-order directional derivative at base vertex $g_1 \leftarrow \frac{\partial f}{\partial z_p}|_{Z_{V \cup S}}$
    $\Sigma_{\text{rays}} \leftarrow \Sigma_{\text{rays}} + (-1)^{k-1-|V|} \cdot g_1$
  end for
  return $(x_p - R_p) \cdot \Sigma_{\text{rays}}$
end if

B.1 PRACTICAL IMPLEMENTATIONS

While the pseudo-codes formalize formalize the Monte Carlo estimation for a single coalition T evaluated with budget M, practical pipelines require computing interactions across $\mathcal{T} = \{T|T \subseteq N, |T| = k\}$ with $\binom{n}{k}$ terms. Evaluating each coalition independently would require $\mathcal{O}(M\binom{n}{k})$ evaluations, rendering high-dimensional analysis computationally prohibitive. We address this concern through reverse-mode amortization.

Reverse-mode amortization. In deep neural networks, reverse-mode automatic differentiation computes the full gradient $\nabla_z f(z) \in \mathbb{R}^n$ in a single backward pass. We leverage this property to evaluate interactions by slices. For pairwise interactions ($k = 2$), we select a pivot feature $i \in N$ and a sampled background coalition $S \subseteq N \setminus \{i\}$. Evaluating the gradient difference:$\Delta g_i(Z_S) = \nabla_z f(Z_{S\cup\{i\}}) - \nabla_z f(Z_S) \in \mathbb{R}^n$ simultaneously provides derivative shifts along all candidate coordinates $j \in N \setminus \{i\}$. More generally, for k-th order interactions, fixing a pivot coalition $P \subset N$ with $|P| = k-1$ evaluates the alternating gradient vector $\Delta g_P(Z_S) = \nabla_z \left[\sum_{V\subseteq P}(-1)^{k-1-|V|} f(Z_{V\cup S})\right]$ across 2^{k-1} backward passes, simultaneously updating target coalitions $T = P \cup \{j\}$ for all $n-k+1$ features $j \in N \setminus P$ in parallel.

Geometric Necessity of Base-Vertex Evaluation. In Proposition 2, IDR expresses interaction terms with 1D integrals $\int_0^1 \frac{\partial f}{\partial z_p}(r_p(\xi))\, d\xi$. Along an arbitrary ray position $\xi > 0$, the evaluation point for feature j is given by

$$r_j(\xi_j; Z_S) = Z_S + \xi_j(x_j - R_j)e_j$$

For two distinct candidate features $j \neq l$, the points $r_j(\xi_j; Z_S)$ and $r_l(\xi_l; Z_S)$ lie along orthogonal coordinate axes and represent distinct spatial locations in $\mathbb{R}^n$. Since a single reverse-mode backward pass can only evaluate the gradient field at a single point, evaluating $\xi > 0$ along rays independently would require $n-1$ distinct backward passes, preventing reverse-mode parallelism. More generally, for a pivot coalition P, candidate rays along feature $j \in N\setminus P$ originate at $r_j(0; Z_{V\cup S}) = Z_{V\cup S}$ for each $V \subseteq P$. Since these 2^{k-1} vertices are identical for all $j \in N \setminus P$, evaluating the full gradient field at each vertex allows a total of 2^{k-1} backward passes to simultaneously update all $n-k+1$ candidate interactions in parallel.

However, at the base vertex ($\xi = 0$), all candidate rays share the exact same origin point:

$$r_j(0; Z_S) = Z_S \quad \forall j \in N \setminus \{i\}.$$

By evaluating derivatives strictly at $Z_{S\cup\{i\}}$ and Z_S, single backward passes at $Z_{S\cup\{i\}}$ and Z_S (two passes in total) simultaneously update the interaction terms for all $n-1$ features $j \in N \setminus \{i\}$ in parallel.

Exact induced context distributions. Under reverse-mode amortization, we fix a pivot subcoalition $P \subset N$ of size $|P| = k-1$ and sample a background coalition $S \subseteq N \setminus P$. For feature $j \in N \setminus P$, evaluating the j-th component of $\Delta g_P(Z_S)$ updates the discrete derivative for $T = P \cup \{j\}$ ($|T| = k$). Since $S \subseteq N \setminus P$ is sampled over all features outside P, j may be absent ($j \notin S$, meaning $z_j = R_j$) or present ($j \in S$, meaning $z_j = x_j$). We show that this marginalization preserves the exact context distributions for any target context $C \subseteq N \setminus (P \cup \{j\}) = N \setminus T$. First, we show the preservation for PIIs.

Proposition 4 *Let $P \subset N$ be a pivot coalition of size $|P| = k-1$, and let background coalition $S \subseteq N\setminus P$ be drawn under an independent Bernoulli distribution parameterized by $u \in [0,1]^{n-k+1}$. For feature $j \in N \setminus P$ and target context $C \subseteq N \setminus (P \cup \{j\})$, evaluating the j-th component of $\Delta g_P(Z_S)$ generates context C with probability:*

$$\mathbb{P}(C \mid u) = \prod_{m\in C} u_m \prod_{l\in(N\setminus T)\setminus C} (1-u_l). \tag{27}$$

For $C \subseteq N \setminus (P \cup \{j\})$, coordinate j evaluates context C under two mutually exclusive cases:

1. $j \notin S$: Feature j is at baseline ($z_j = R_j$) and $S = C$. This case occurs with probability:

$$\mathbb{P}(j \notin S \wedge S = C) = (1 - u_j) \prod_{m \in C} u_m \prod_{l \in (N \setminus T) \setminus C} (1 - u_l). \tag{28}$$

2. $j \in S$: Feature j is active ($z_j = x_j$) and $S = C \cup \{j\}$. This case occurs with probability:

$$\mathbb{P}(j \in S \wedge S = C \cup \{j\}) = u_j \prod_{m \in C} u_m \prod_{l \in (N \setminus T) \setminus C} (1 - u_l). \tag{29}$$

Summing these probabilities:

$$\begin{aligned} \mathbb{P}(C \mid u) &= \Big[(1 - u_j) + u_j\Big] \prod_{m \in C} u_m \prod_{l \in (N \setminus T) \setminus C} (1 - u_l) \\ &= \prod_{m \in C} u_m \prod_{l \in (N \setminus T) \setminus C} (1 - u_l). \end{aligned} \tag{30}$$

Since $(1 - u_j) + u_j = 1$, feature j's inclusion cancels out. Integrating over trajectory measure $t \sim \mu(t)$ recovers the exact PII definition for coalition $T = P \cup \{j\}$.

The preservation also holds for CIIs.

Proposition 5 *Let $P \subset N$ be a pivot coalition of size $|P| = k - 1$, and let background cardinality $s = |S| \in \{0, 1, \ldots, n - k + 1\}$ be sampled from $P(s)$, followed by drawing $S \subseteq N \setminus P$ uniformly from all $\binom{n-k+1}{s}$ subsets of size s. For feature $j \in N \setminus P$, the evaluation along coordinate j induces a context $C \subseteq N \setminus (P \cup \{j\})$ with cardinality $c = |C| \in \{0, 1, \ldots, n - k\}$ governed by:*

1. ***Conditional Uniformity:*** *For cardinality c, context C is distributed strictly uniformly over all subsets of cardinality c: $C \mid c \sim \mathcal{U}_c$.*
2. ***Induced Cardinality Distribution:*** *The induced context size distribution $P_{\text{ind}}(c) = \mathbb{P}(|C| = c)$ satisfies:*

$$P_{\text{ind}}(c) = \left(1 - \frac{c}{n - k + 1}\right) P(c) + \left(\frac{c + 1}{n - k + 1}\right) P(c + 1). \tag{31}$$

3. ***Unbiased Importance Reweighting:*** *Reweighting each evaluated context with:*

$$K_{\text{weight}}(c) = \frac{W(c)\binom{n-k}{c}}{P_{\text{ind}}(c)} \tag{32}$$

reconstructs the exact CII.

For context $C \subseteq N \setminus (P \cup \{j\})$ of cardinality $c = |C|$, coordinate j evaluates context C under two mutually exclusive cases:

- $j \notin S$: Requires $S = C$ (size $s = c$). Its probability is:

$$\mathbb{P}(j \notin S \wedge S = C) = P(c) \cdot \frac{1}{\binom{n-k+1}{c}}. \tag{33}$$

- $j \in S$: Requires $S = C \cup \{j\}$ (size $s = c + 1$). Its probability is:

$$\mathbb{P}(j \in S \wedge S = C \cup \{j\}) = P(c + 1) \cdot \frac{1}{\binom{n-k+1}{c+1}}. \tag{34}$$

Summing these probabilities, the probability of drawing specific context C depends strictly on its cardinality c:

$$\mathbb{P}(C) = \frac{P(c)}{\binom{n-k+1}{c}} + \frac{P(c + 1)}{\binom{n-k+1}{c+1}} = \mathcal{Q}(c). \tag{35}$$

Since $\mathbb{P}(C)$ depends only on c, summing over $\binom{n-k}{c}$ contexts yields the total induced cardinality probability:

$$P_{\text{ind}}(c) = \sum_{C \subseteq N \setminus T : |C| = c} \mathbb{P}(C) = \binom{n-k}{c} \left[\frac{P(c)}{\binom{n-k+1}{c}} + \frac{P(c+1)}{\binom{n-k+1}{c+1}} \right]. \tag{36}$$

Evaluating the binomial ratios:

$$\frac{\binom{n-k}{c}}{\binom{n-k+1}{c}} = \frac{(n-k)!}{c!(n-k-c)!} \cdot \frac{c!(n-k+1-c)!}{(n-k+1)!} = \frac{n-k+1-c}{n-k+1} = 1 - \frac{c}{n-k+1}, \tag{37}$$

$$\frac{\binom{n-k}{c}}{\binom{n-k+1}{c+1}} = \frac{(n-k)!}{c!(n-k-c)!} \cdot \frac{(c+1)!(n-k-c)!}{(n-k+1)!} = \frac{c+1}{n-k+1}. \tag{38}$$

Substituting these ratios:

$$P_{\text{ind}}(c) = \left(1 - \frac{c}{n-k+1}\right) P(c) + \left(\frac{c+1}{n-k+1}\right) P(c+1). \tag{39}$$

Furthermore, the conditional distribution is:

$$\mathbb{P}(C \mid |C| = c) = \frac{\mathbb{P}(C)}{P_{\text{ind}}(c)} = \frac{\mathcal{Q}(c)}{\binom{n-k}{c}\mathcal{Q}(c)} = \frac{1}{\binom{n-k}{c}}, \tag{40}$$

confirming strict conditional uniformity $C \mid c \sim \mathcal{U}_c$. Weighting by $K_{\text{weight}}(c) = \frac{W(c)\binom{n-k}{c}}{P_{\text{ind}}(c)}$ yields:

$$\begin{aligned} \mathbb{E}\Big[K_{\text{weight}}(c)\Delta_T f(Z_C)\Big] &= \sum_{c=0}^{n-k} P_{\text{ind}}(c) \frac{W(c)\binom{n-k}{c}}{P_{\text{ind}}(c)} \mathbb{E}_{C\sim\mathcal{U}_c}\big[\Delta_T f(Z_C)\big] \\ &= \sum_{c=0}^{n-k} W(c)\binom{n-k}{c} \mathbb{E}_{C\sim\mathcal{U}_c}\big[\Delta_T f(Z_C)\big] = I_{\text{CII}}(T), \end{aligned} \tag{41}$$

proving preservation.

In particular, using cardinality sampling $P(s) = \frac{1}{n}$ yields $P_{\text{ind}}(c) = \frac{1}{n-1}$ for all context sizes $c \in \{0, \ldots, n-2\}$. Since 2nd-order SII weight satisfies $W(c)\binom{n-2}{c} = \frac{1}{n-1}$, the induced distribution matches the target distribution ($P_{\text{ind}}(c) = W(c)\binom{n-2}{c}$). Therefore, the importance weight multiplier reduces to 1 across all coordinates.

Discretization bias vs. computational efficiency. In Continuous Piecewise-Linear (CPWL) models such as deep ReLU networks, $\nabla_z f(z)$ is locally constant almost everywhere within each polytope. When an integration ray between R_j and x_j remains within a polytope, evaluating the gradient at the base vertex $\xi = 0$ is mathematically exact.

When an integration ray crosses one or more boundaries, evaluating strictly at $\xi = 0$ acts as a one-point quadrature approximation that omits intermediate slope transitions. The approximation introduces an asymptotic discretization bias as global query budget $Q \to 2^n$. However, this trade-off has several benefits:

1. It eliminates inner-loop ray scans, reducing the cost of evaluating all $n-1$ pairs to exactly two backward passes.
2. In practical regimes where $Q \ll 2^n$, this formulation exhibits lower variance than combinatorial subset sampling and can correctly recovering interaction rank orderings.

When asymptotic precision is required, practitioners can replace the 1-point estimator with standard M_ξ-point 1D quadrature at the cost of $2M_\xi$ backward passes.

C Proof of Proposition 1: Model-Induced Discrete Derivative as Integral of Mixed Partial Derivatives

To begin the proof, we fix target coalition $T = \{i_1, \ldots, i_k\}$ and context $S \subseteq N \setminus T$. Features outside T are set to the context values $Z_S(x, R)_{-T} \in \mathbb{R}^{n-k}$:

$$(Z_S(x,R)_{-T})_j = \begin{cases} x_j & \text{if } j \in S \subseteq N \setminus T, \\ R_j & \text{if } j \notin S \end{cases} \tag{42}$$

Let $B = \prod_{m=1}^{k} [R_{i_m}, x_{i_m}] \subset \mathbb{R}^k$ be the k-dimensional hyper-rectangle spanned by features in T.

By Theorem 2, the integral of the k-th order mixed partial derivative of $f(z_T, (Z_S)_{-T})$ over B equals the alternating sum evaluated across the 2^k vertices of B:

$$\int_{R_{i_1}}^{x_{i_1}} \cdots \int_{R_{i_k}}^{x_{i_k}} \frac{\partial^k f(z_T, (Z_S)_{-T})}{\prod_{j \in T} \partial z_j} \, dz_{i_k} \ldots dz_{i_1} = \sum_{U \subseteq T} (-1)^{k-|U|} f(z_T^U, (Z_S)_{-T}) \tag{43}$$

where each vertex vector $z_T^U \in \mathbb{R}^k$ is defined coordinate-wise by:

$$(z_T^U)_j = \begin{cases} x_j & \text{if } j \in U, \\ R_j & \text{if } j \in T \setminus U \end{cases} \tag{44}$$

Combining the vertex coordinates z_T^U with context $(Z_S(x,R))_{-T}$ yields:

$$(z_T^U, (Z_S)_{-T})_j = \begin{cases} x_j & \text{if } j \in U \cup S, \\ R_j & \text{if } j \notin U \cup S \end{cases} \quad \Longrightarrow \quad (z_T^U, (Z_S)_{-T}) \equiv Z_{U \cup S}(x, R) \tag{45}$$

Therefore, evaluating f at vertex $(z_T^U, (Z_S)_{-T})$ corresponds to evaluating f at $Z_{U \cup S}(x, R)$. Substituting back into Equation 43:

$$\sum_{U \subseteq T} (-1)^{k-|U|} f(z_T^U, (Z_S)_{-T}) = \sum_{U \subseteq T} (-1)^{k-|U|} f\big(Z_{U \cup S}(x, R)\big) \equiv \Delta_T f(Z_S(x, R)) \tag{46}$$

Deriving Coordinate-Ray Form. To get a representation using the unit hypercube $[0,1]^k$, consider the ray representation along a single feature $j \in T$:

$$z_j(\xi_j) = R_j + \xi_j (x_j - R_j), \quad \xi_j \in [0,1] \tag{47}$$

Differentiating with respect to ξ_j gives $dz_j = (x_j - R_j)\, d\xi_j$, which factors out interval length $(x_j - R_j)$.

Repeating this substitution independently for each coordinate $i_1, \ldots, i_k \in T$, we can extract the volume of the hyper-rectangle:

$$dz_{i_1} \ldots dz_{i_k} = \left(\prod_{j \in T} (x_j - R_j) \right) d\xi_1 \ldots d\xi_k \tag{48}$$

Substituting to the original equation results in the coordinate-ray integral form of discrete derivative:

$$\Delta_T f(Z_S(x,R)) = \left(\prod_{j \in T} (x_j - R_j) \right) \int_{[0,1]^k} \frac{\partial^k f(r_T(\xi; Z_S))}{\prod_{j \in T} \partial z_j} \, d\xi_1 \ldots d\xi_k \tag{49}$$

D Proof of Proposition 2: Iterated Dimensionality Reduction

To begin the proof, we first choose a feature p, and show the coordinate ray form of 1st order discrete derivative.

$$f(x_p, Z_{-p}) - f(R_p, Z_{-p}) = (x_p - R_p) \int_0^1 \frac{\partial f}{\partial z_p}\big(r_p(\xi; Z)\big)\, d\xi \tag{50}$$

We then get the following lemma:

Lemma 1 *For coalition $T \subseteq N$ with $k = |T| \geq 1$ and context $S \subseteq N \setminus T$, the discrete derivative factorizes along any chosen pivot feature $p \in T$ as:*

$$\Delta_T f(Z_S(x, R)) = \sum_{V \subseteq T \setminus \{p\}} (-1)^{k-1-|V|} \Big(f(Z_{V \cup \{p\} \cup S}(x, R)) - f(Z_{V \cup S}(x, R)) \Big) \tag{51}$$

This lemma comes from the fact that 1st order discrete derivative operator commutes, $\Delta_T f(Z_S) = \Delta_{T \setminus \{p\}}(\Delta_{\{p\}} f(Z_S))$. Expanding $\Delta_{T \setminus \{p\}}$ over sub-coalitions $V \subseteq T \setminus \{p\}$ yields Equation 51. It should be noted that $\Delta_T f(Z_S)$ does not have any dependency on p, i.e., is invariant to the choice of pivot feature.

By Lemma 1, we expand $\Delta_T f(Z_S)$ along $p \in T$. For each sub-coalition $V \subseteq T \setminus \{p\}$, the evaluation points $Z_{V \cup \{p\} \cup S}(x, R)$ and $Z_{V \cup S}(x, R)$ differ solely along p. Using the coordinate ray form along $r_p(\xi; Z_{V \cup S})$ expresses this marginal difference as a 1D line integral:

$$f\big(Z_{V \cup \{p\} \cup S}(x, R)\big) - f\big(Z_{V \cup S}(x, R)\big) = (x_p - R_p) \int_0^1 \frac{\partial f}{\partial z_p}\big(r_p(\xi; Z_{V \cup S}(x, R))\big)\, d\xi \tag{52}$$

Substituting this integral into the summation over $V \subseteq T \setminus \{p\}$ and factoring out $(x_p - R_p)$ yields Equation 14.

E Proof of Convergence Guarantees

We begin our proof by reiterating the bounding constants. We define the maximum feature displacement $D_{\max} = \sup_{R \in D, j \in N} |x_j - R_j| < \infty$, the maximum mixed partial derivative $G^{(k)}_{\max} = \sup_{z \in B_T} \left| \frac{\partial^k f(z)}{\prod_{j \in T} \partial z_j} \right| < \infty$ over hyper-rectangle B_T, and the maximum sample weight multiplier $K_{\max} = \sup_s |K_{\text{weight}}| < \infty$.

The estimator $\hat{I}_M$ is the average of M independent identically distributed random variables $Y^{(1)}, \ldots, Y^{(M)}$ drawn from $Y = K_{\text{weight}}\big(\prod_{j \in T}(x_j - R_j)\big) \frac{\partial^k f(r_T(\xi; Z_S))}{\prod_{j \in T} \partial z_j}$. By the law of total expectation, Y satisfies $\mathbb{E}[Y] = I(T)$.

Applying the triangle inequality to Y:

$$|Y| \leq |K_{\text{weight}}| \cdot \left(\prod_{j \in T} |x_j - R_j| \right) \cdot \left| \frac{\partial^k f(r_T(\xi; Z_S))}{\prod_{j \in T} \partial z_j} \right| \leq K_{\max} (D_{\max})^k G^{(k)}_{\max} \equiv C \tag{53}$$

Thus, $Y \in [-C, C]$, with length $2C = 2K_{\max}(D_{\max})^k G^{(k)}_{\max}$. By Hoeffding's inequality:

$$\mathbb{P}\big(|\hat{I}_M - I(T)| \geq \epsilon\big) \leq 2 \exp\left(-\frac{2M^2\epsilon^2}{M(2C)^2} \right) = 2 \exp\left(-\frac{M\epsilon^2}{2C^2} \right) \tag{54}$$

Setting $2 \exp\left(-\frac{M\epsilon^2}{2C^2}\right) \leq \delta$ and solving for M, we get:

$$-\frac{M\epsilon^2}{2C^2} \leq \ln\left(\frac{\delta}{2}\right) \iff \frac{M\epsilon^2}{2C^2} \geq \ln\left(\frac{2}{\delta}\right) \iff M \geq \frac{2C^2 \ln(2/\delta)}{\epsilon^2} \tag{55}$$

The proof for IDR is essentially identical. A single-sample draw is $Y_{\text{IDR}} = K_{\text{weight}}(x_p - R_p) \sum_{V \subseteq T \setminus \{p\}} (-1)^{k-1-|V|} \frac{\partial f}{\partial z_p}(r_p(\xi; Z_{V \cup S}))$ under IDR. Therefore, we get the same bound except that C is defined as:

$$|Y_{\text{IDR}}| \leq |K_{\text{weight}}| \cdot |x_p - R_p| \sum_{V \subseteq T \setminus \{p\}} \left| \frac{\partial f}{\partial z_p} \right| \leq K_{\max} D_{\max} \cdot \left(2^{k-1} G^{(1)}_{\max}\right) \equiv C_{\text{IDR}} \tag{56}$$

E.1 Proof of Global Query Budget

Based on the bound on M, we can derive the bound on global query budget Q under reverse mode amortization.

Corollary 1 *Let $\mathcal{T} = \{T \subseteq N : |T| = k\}$ be set of coalitions of order k. For $\epsilon > 0$ and $\delta \in (0, 1)$, the bound of maximum deviation converges with $\mathbb{P}(\max_{T \in \mathcal{T}} |\hat{I}(T) - \mathbb{E}[\hat{I}(T)]| \geq \epsilon) \leq \delta$ provided the total global query budget Q satisfies:*

$$Q \geq \frac{2^k \binom{n}{k-1}}{k\epsilon^2} C^2 \left(\ln(2/\delta) + k \ln n \right) \tag{57}$$

Applying Boole's inequality to the single-coalition tail bound across all $|\mathcal{T}| = \binom{n}{k}$ target coalitions:

$$\mathbb{P}\left(\max_{T \in \mathcal{T}} \left| \hat{I}(T) - I(T) \right| \geq \epsilon \right) \leq \sum_{T \in \mathcal{T}} \mathbb{P}\left(\left| \hat{I}(T) - I(T) \right| \geq \epsilon \right) \leq 2 \binom{n}{k} \exp\left(-\frac{M\epsilon^2}{2C^2} \right). \tag{58}$$

Solving for M:

$$M \geq \frac{2C^2}{\epsilon^2}\left(\ln\left(\frac{2}{\delta}\right) + \ln\binom{n}{k}\right). \tag{59}$$

Since each T needs M queries, we have to allocate global budget Q over $\mathcal{T}$. Observe from the implementation trick in Appendix B.1 that reverse-mode automatic differentiation iterates across all $\binom{n}{k-1}$ pivot configurations of size $k-1$. For each configuration, evaluating the gradient shift across coordinate rays simultaneously updates discrete derivative estimates for all $n-k+1$ coalitions using 2^{k-1} backward passes. Since each coalition T contains $\binom{k}{k-1} = k$ sub-coalitions of size $k-1$, allocating M_{pivot} samples per configuration yields $M = kM_{\text{pivot}}$ updates per coalition under a global query budget of $Q = 2^{k-1}\binom{n}{k-1}M_{\text{pivot}}$.

Substituting $M = \frac{kQ}{2^{k-1}\binom{n}{k-1}}$ into Equation 59 yields

$$Q \geq \frac{2^k\binom{n}{k-1}}{k\epsilon^2}C^2\left(\ln\left(\frac{2}{\delta}\right) + \ln\binom{n}{k}\right). \tag{60}$$

Using $\ln\binom{n}{k} \leq k\ln n$, we find $Q = \mathcal{O}\left(\binom{n}{k-1}\frac{\ln(1/\delta)+k\ln n}{\epsilon^2}\right)$. For $k=1$, this limit is $Q = \mathcal{O}\left(\frac{\ln(1/\delta)+\ln n}{\epsilon^2}\right)$, scaling logarithmically with feature dimension n. For $k=2$, the limit evaluates to $Q = \mathcal{O}\left(\frac{n(\ln(1/\delta)+\ln n)}{\epsilon^2}\right)$,scaling quasi-linearly.

F FULL RESULTS

In this section, we report the full results for Shapley Interaction Index (SII) and Banzhaf Interaction Index (BII) at $k = 1, 2$.

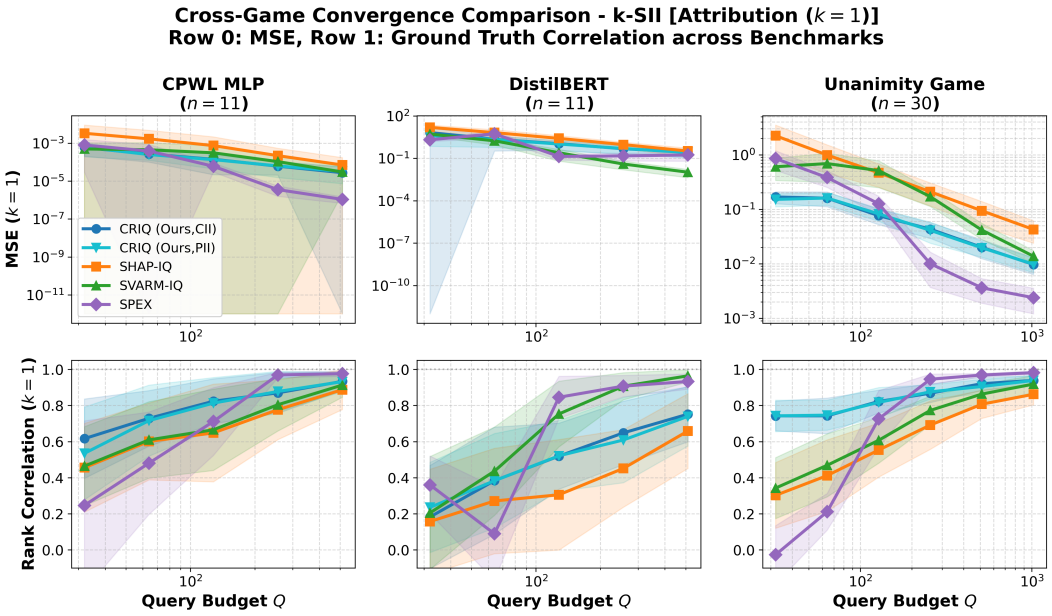

Figure 4: MSE and rank correlation of Shapley value estimated by CRIQ and baseline methods over query budget Q for unanimity network ($n = 30$), MLP network ($n = 11$) and DistilBERT ($n = 11$).

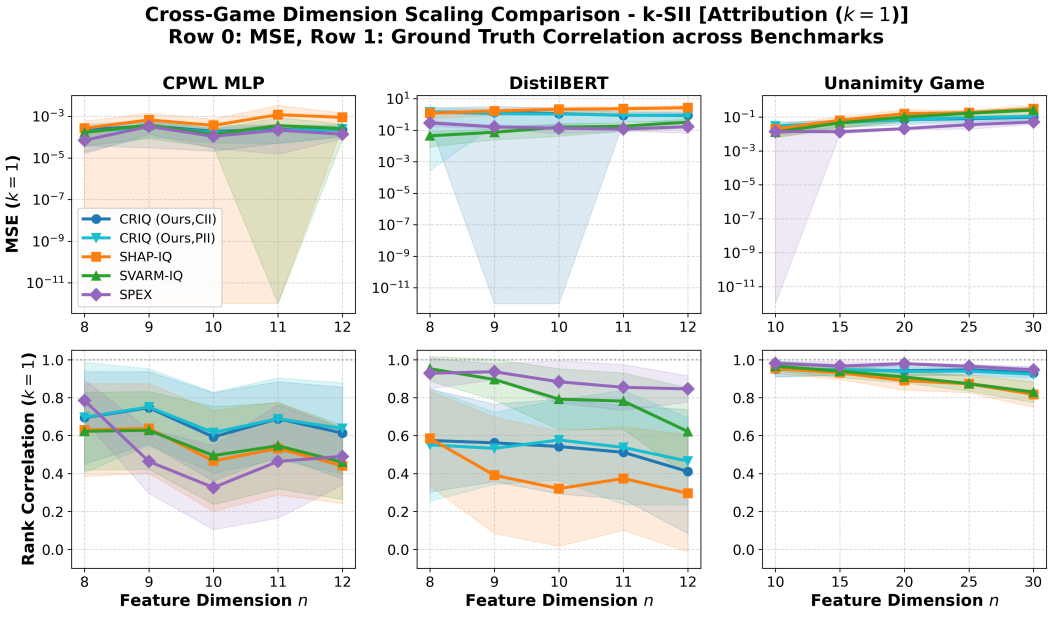

Figure 5: MSE and rank correlation of Shapley value estimated by CRIQ and baseline methods over feature dimension n for unanimity network, MLP network and DistilBERT for fixed Q.

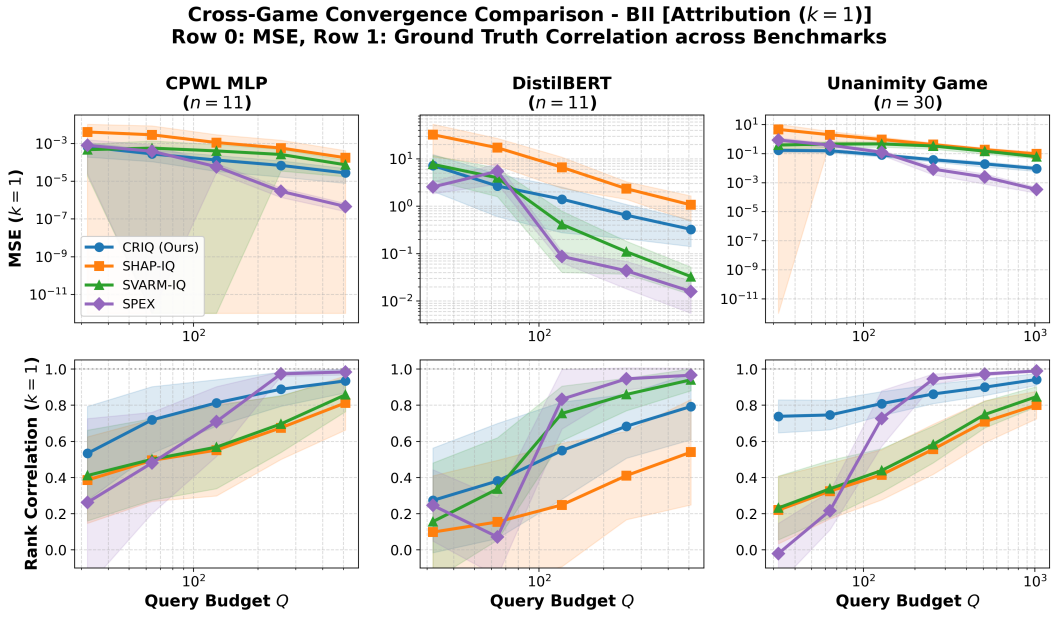

Figure 6: MSE and rank correlation of Banzhaf value estimated by CRIQ and baseline methods over query budget Q for unanimity network ($n = 30$), MLP network ($n = 11$) and DistilBERT ($n = 11$).

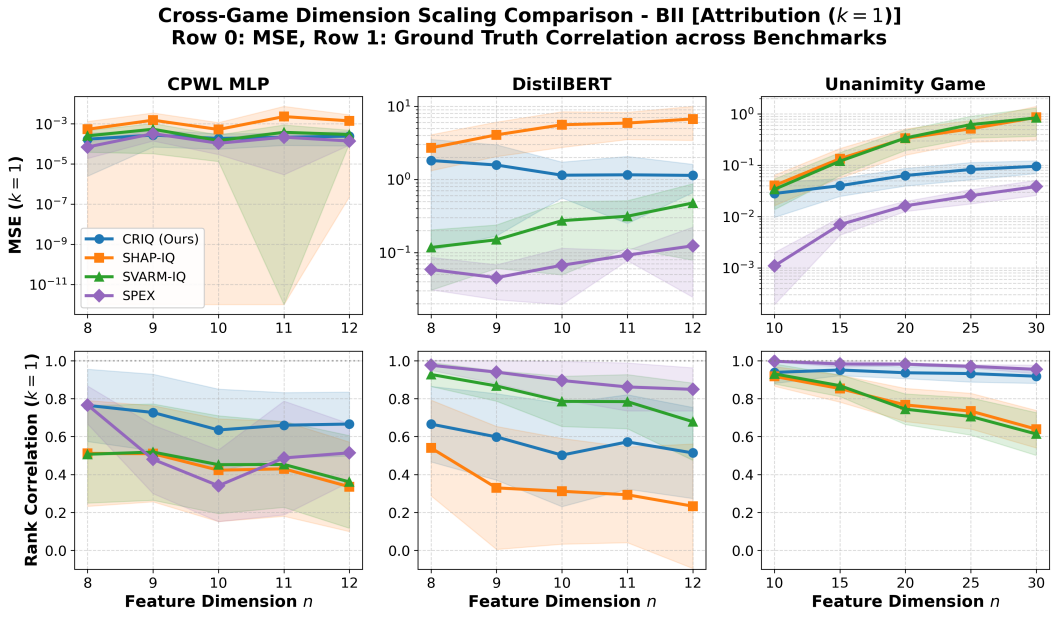

Figure 7: MSE and rank correlation of Banzhaf value estimated by CRIQ and baseline methods over feature dimension n for unanimity network, MLP network and DistilBERT for fixed Q.

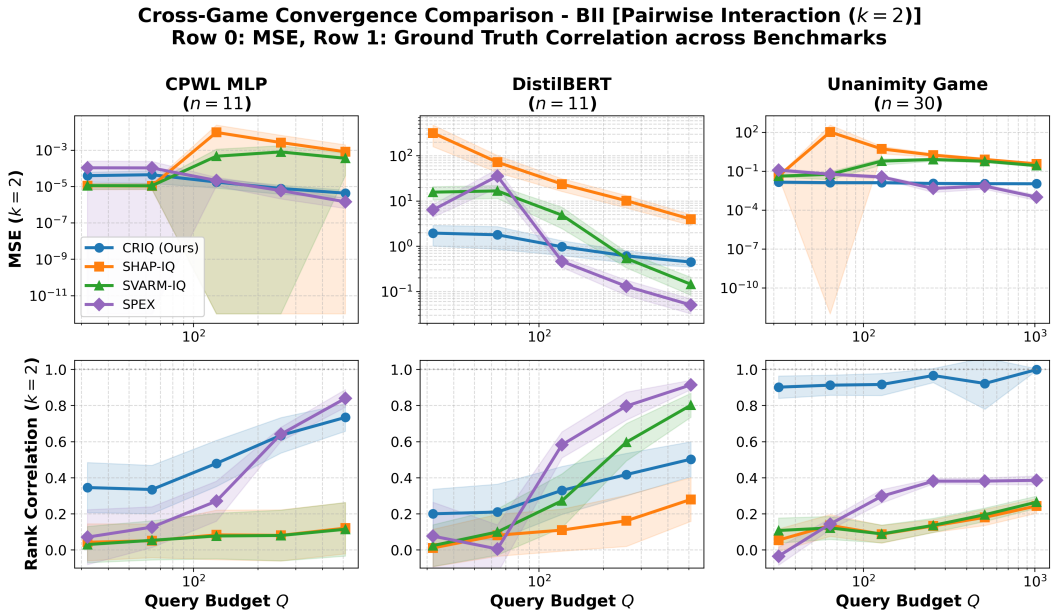

Figure 8: MSE and rank correlation of 2nd order Banzhaf Interaction Index estimated by CRIQ and baseline methods over query budget Q for unanimity network ($n = 30$), MLP network ($n = 11$) and DistilBERT ($n = 11$).

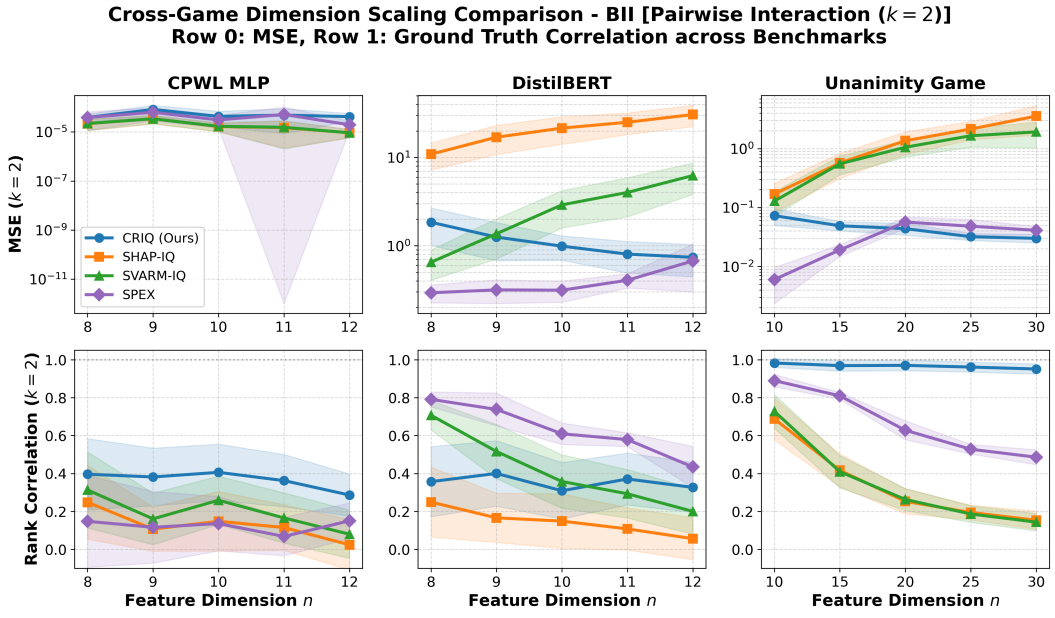

Figure 9: MSE and rank correlation of 2nd order Banzhaf Interaction Index estimated by CRIQ and baseline methods over feature dimension n for unanimity network, MLP network and DistilBERT for fixed Q.