ON THE HIDDEN INTERACTION STRUCTURE OF SHAPLEY VALUE EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

The Shapley value summarizes a feature's marginal contribution by averaging it over different contexts, but this scalar attribution does not reveal how the underlying effect varies with the surrounding features. We study the higher-order interaction structure compressed by this averaging. We derive an order-wise representation in which the ordinary Shapley value, at any order k, is expressed through lower-order empty-context effects and expected context-dependent order-k effects. This makes explicit how higher-order interaction structure is aggregated into scalar Shapley summaries. For cardinal interaction indices, we further show that efficiency is equivalent to Shapley recovery under uniform member-wise allocation. We then ask when these scalar expectations adequately represent their underlying context-wise effects. We quantify the remaining variability by *Shapley context variance*. We show that conditional context variance decomposes into nonnegative squared contributions from higher-order interactions, and that full context variance implies a lower bound on the aggregate second moment of immediate next-order extensions. Thus, large context variance requires substantial higher-order effects, whereas small variance indicates that the scalar expectation closely represents the underlying context-wise effects. Analytical and empirical studies further show that similar scalar summaries can mask markedly different context dependence and that context variance identifies cases where higher-order inspection is informative.

1 INTRODUCTION

The Shapley value is a fundamental solution concept in cooperative game theory and has become a widely used framework for feature attribution in machine learning (Shapley, 1953; Štrumbelj & Kononenko, 2014; Lundberg & Lee, 2017; Sundararajan & Najmi, 2020; Lundberg et al., 2020; Rozemberczki et al., 2022). In model explanation, it summarizes a feature's contribution by averaging its marginal effect over different surrounding coalitions. This provides a compact scalar attribution, but does not consider how the underlying effect varies across the contexts.

Figure 1 illustrates this distinction in a simple image-classification setting. In the first classroom scenario, a region of the board belongs to a larger, highly informative visual pattern, so its contribution to the target-class score can change substantially depending on which other board regions are present. In the second, the backpack provides weaker but largely isolated evidence whose contribution remains relatively stable across contexts. Nevertheless, averaging these effects can yield similar scalar attributions. The same scalar value can therefore represent either a strongly context-dependent effect or a consistently weak one.

Such context dependence is induced by interactions. When the effect of a feature changes as another feature enters the context, their interaction accounts for that change. If the joint effect itself changes with additional context, higher-order interactions are involved. Therefore, context variation provides a direct view of interaction structure that is compressed when Shapley explanations are reduced to scalar averages.

> *What higher-order interaction structure is compressed by scalar Shapley explanations, and when does the scalar expectation adequately represent the underlying context-wise effects?*

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.

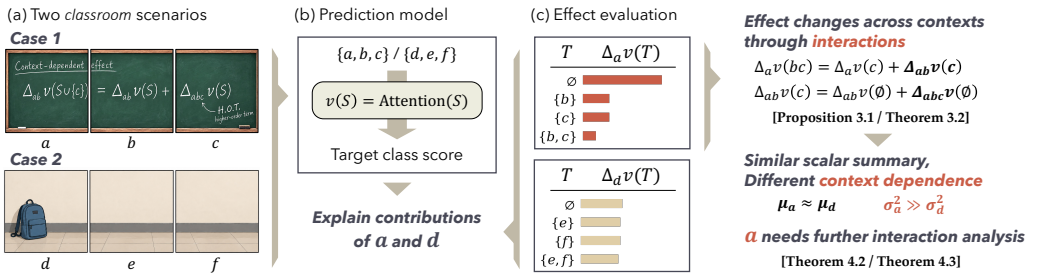


Figure 1: Schematic image-classification example illustrating context dependence hidden by scalar Shapley summaries. Under the same attention model, visual features a and d can have similar scalar summaries despite markedly different context-wise effects. Changes across contexts arise through interactions, motivating context variance as a diagnostic of hidden higher-order structure.

Prior work has identified related limitations of scalar Shapley explanations. Huang & Marques-Silva (2024) show that SHAP can assign zero importance to features relevant to a prediction, while Chang et al. (2025) show how negative interactions can reduce feature-level Shapley values in non-convex games. Kumar et al. (2021) introduce Shapley residuals through a geometric projection of the cooperative game, providing a different diagnostic of when singleton Shapley values incompletely represent the underlying game. Regional explanations also note cancellation of marginal contributions as one failure mode of local Shapley values, within a broader study of local variable importance (I Amoukou & Brunel, 2025). Our focus is specifically on the context-wise effects being averaged and the higher-order interaction structure underlying their variation.

Another line of work extends Shapley attribution from individual features to feature coalitions. SII, Shapley-Taylor, Faith-Shap, and n-Shapley define scalar interaction scores under different allocation principles (Grabisch & Roubens, 1999; Sundararajan et al., 2020; Tsai et al., 2023; Bordt & von Luxburg, 2023), while methods such as SHAP-IQ address their efficient approximation (Fumagalli et al., 2023). Our question is complementary: what information is lost when a context-dependent interaction effect is reduced to a scalar? By making these underlying effects explicit, we characterize the higher-order structure compressed by scalar summaries and assess when those summaries adequately represent their context-wise effects.

We derive an order-wise representation of context-dependent effects and the ordinary Shapley value. At order k, the latter is expressed through lower-order empty-context effects and expected context-dependent order-k effects. Each expected order-k effect is a scalar summary of a context-dependent interaction effect whose variation reflects higher-order interactions. The representation recovers the standard Shapley form at $k = 1$ and the Harsanyi representation at full order, and its expected order-k effects coincide with the corresponding maximal-order Shapley-Taylor attributions. More generally, for cardinal interaction indices, we show that efficiency is equivalent to recovering the ordinary Shapley value by uniformly distributing each coalition attribution among its members, even though different indices allocate higher-order contributions differently across interaction orders.

This representation makes explicit the interaction structure and context-dependent effects underlying scalar Shapley explanations. We then ask when these scalar expectations adequately represent the underlying effects, which may vary substantially across contexts in complex models. We quantify this remaining variability by the variance, which we call *Shapley context variance*. We show that conditional context variance decomposes into nonnegative squared contributions from higher-order interactions, and that full context variance implies a lower bound on the aggregate second moment of immediate next-order extensions. These results connect variability directly to higher-order structure: small variance indicates that the scalar expectation closely represents the underlying effects, whereas large variance signals substantial higher-order effects and motivates further interaction analysis.

Through an analytical example and empirical studies, we show that similar scalar summaries can mask markedly different context-wise effects and that context variance identifies cases where higher-order inspection is informative. Together, these results link scalar Shapley explanations to the

higher-order structure they compress and provide a diagnostic for assessing when further interaction analysis is warranted.

2 CONTEXT-WISE EFFECTS AND SCALAR SUMMARIES

This section formalizes the quantities illustrated in Figure 1. We explain the Shapley value as an expectation of context-wise marginal effects, and then show how changes in these effects across contexts correspond to higher-order interactions. This relation provides the basis for the order-wise representation developed in Section 3.

2.1 SHAPLEY VALUE AS EXPECTED MARGINAL EFFECTS

For convenience, we omit braces for singletons (e.g., $T \cup i$) and denote set cardinalities by lowercase letters (e.g., $s = |S|$) when it is clear. Let $N = \{1, \dots, n\}$ be the player set and $v : 2^N \to \mathbb{R}$ a set function. In model explanation, players represent input features and $v(S)$ evaluates a feature subset S. Figure 1(b) schematically depicts one such target-class score.

For a player $i \in N$ and a context $S \subseteq N \setminus i$, its marginal effect is $\Delta_i v(S) \coloneqq v(S \cup i) - v(S)$. The Shapley value assigns player i its expected marginal effect under the Shapley context distribution:

$$\phi_i(v) = \sum_{S \subseteq N \setminus i} \underbrace{\frac{1}{n}\binom{n-1}{s}^{-1}}_{p_i(S)} \Delta_i v(S) = \mathbb{E}_{S \sim p_i}\left[\Delta_i v(S)\right]. \tag{1}$$

Equivalently, p_i is the distribution of players preceding i in a uniformly random permutation of N (Castro et al., 2009). The bars in Figure 1(c) represent such context-wise effects, and their scalar Shapley summaries are the corresponding weighted averages.

2.2 CONTEXT DEPENDENCE AS HIGHER-ORDER INTERACTION

Consider the effect of a feature i under two contexts that differ only by one additional player j. Their difference, $\Delta_i v(T \cup j) - \Delta_i v(T)$, measures how the presence of j changes the effect of i. We denote this pairwise interaction effect by $\Delta_{ij} v(T)$. More generally, for a nonempty coalition $C \subseteq N$ and a context $T \subseteq N \setminus C$, we define its interaction effect by the discrete derivative

$$\Delta_C v(T) \coloneqq \sum_{L \subseteq C} (-1)^{c-l} v(T \cup L). \tag{2}$$

At the empty context, $\alpha_C \coloneqq \Delta_C v(\emptyset)$ is the Harsanyi dividend associated with C (Harsanyi, 1982; Dehez, 2017).

For any $j \notin C \cup T$, these interaction effects satisfy

$$\Delta_C v(T \cup j) - \Delta_C v(T) = \Delta_{C \cup j} v(T). \tag{3}$$

Equation (3) shows that changing the context of an effect by one player produces an interaction of one higher order. If this higher-order effect also varies with context, the same relation applies again. The relations highlighted in Figure 1 illustrate this recursion from a feature effect to pairwise and then three-way interactions. Repeated application therefore expresses context dependence through interactions of progressively higher order, which motivates the order-wise representation developed in Section 3.

3 INTERACTION STRUCTURE IN SHAPLEY EXPLANATIONS

Starting from the recursive relation in Equation (3), we derive an order-k representation of a context-wise effect $\Delta_R v(S)$ and then obtain the corresponding representation of the ordinary Shapley value by averaging over its contexts. This makes explicit the context-dependent order-k effects whose expectations appear as scalar interaction terms.

3.1 Higher-Order Interactions in Context-Wise Effects

Consider a context-wise effect $\Delta_R v(S)$ for nonempty disjoint coalitions $R, S \subseteq N$. Repeatedly applying Equation (3) to its context yields the following closed-form representation at an arbitrary order k.

Proposition 3.1 (Order-k representation of a context-wise effect)**.** *Let $R, S \subseteq N$ be nonempty disjoint coalitions, with $r = |R|$ and $s = |S|$. For any integer k such that $r < k \leq n$,*

$$\Delta_R v(S) = \sum_{\substack{U \subseteq S \\ r+u<k}} \Delta_{R\cup U} v(\emptyset) + \sum_{\substack{U \subseteq S \\ r+u=k}} \frac{k-r}{s} \sum_{T \subseteq S\setminus U} \binom{s-1}{t}^{-1} \Delta_{R\cup U} v(T). \tag{4}$$

The first term in Proposition 3.1 contains effects involving R below order k, evaluated at the empty context. The second expresses the remaining contribution through context-dependent order-k effects $\Delta_{R\cup U} v(T)$. Since these effects can themselves vary with context through higher-order interactions, they aggregate contributions from interaction orders k and above. The full derivation is provided in Appendix A.1.

3.2 Interaction Representation of Shapley Value

We next extend the order-wise representation in Proposition 3.1 to the Shapley value by averaging over its contexts, yielding the following result.

Theorem 3.2 (Order-k representation of the Shapley value)**.** *For any player $i \in N$ and interaction order $1 \leq k \leq n$,*

$$\phi_i(v) = \sum_{u=0}^{k-2} \frac{1}{u+1} \sum_{\substack{U \subseteq N\setminus i \\ |U|=u}} \Delta_{i\cup U} v(\emptyset) + \frac{1}{k} \sum_{\substack{U \subseteq N\setminus i \\ |U|=k-1}} \mathbb{E}_{T\sim p_{i\cup U}} \left[\Delta_{i\cup U} v(T)\right], \tag{5}$$

where the first sum is empty for $k = 1$. For any nonempty coalition $C \subseteq N$, p_C denotes the distribution over contexts $T \subseteq N \setminus C$, defined by

$$p_C(T) = \frac{|C|}{n} \binom{n-1}{|T|}^{-1}. \tag{6}$$

Equivalently, if π is a uniformly random permutation of N, p_C is the distribution of the players preceding the first member of C in π. The proof and permutation characterization are given in Appendix A.2.

At $k = 1$, Equation (5) reduces to the standard expected marginal-effect form of the Shapley value. At $k = n$, it recovers the Harsanyi representation. At intermediate orders, the scalar terms are expectations of context-dependent order-k interaction effects whose context dependence reflects higher-order interactions.

For a nonempty coalition C, we denote the expected interaction effect in Equation (5) by

$$\mu_C := \mathbb{E}_{T\sim p_C} \left[\Delta_C v(T)\right]. \tag{7}$$

The quantity μ_C summarizes the context-dependent effect $\Delta_C v(T)$ by its expectation, but does not describe how the effect varies across the contexts being averaged. We next relate these terms to existing interaction indices before examining the information beyond their expectations in Section 4.

3.3 Relation to Existing Interaction Indices

Interaction indices extend attribution from individual players to coalitions, often by aggregating context-wise effects $\Delta_C v(T)$ into scalar coalition attributions. The representation in Theorem 3.2 has a direct connection to the Shapley-Taylor interaction index (STI) (Sundararajan et al., 2020). For a maximum order k, STI assigns $\Delta_C v(\emptyset)$ to coalitions with $|C| < k$ and $\mathbb{E}_{T\sim p_C}[\Delta_C v(T)]$ to coalitions with $|C| = k$. Equation (5) therefore recovers the ordinary Shapley value by uniformly

distributing each STI attribution among its members. This recovery is not unique to STI. For example, n-Shapley recovers the ordinary Shapley value through the same member-wise aggregation despite using different coalition attributions (Bordt & von Luxburg, 2023). We characterize when this recovery holds more generally.

Let $I^k = \{I^k_C\}_{1\le|C|\le k}$ denote an interaction index of maximum order k satisfying linearity, symmetry, and dummy, which in the classical setting belongs to the cardinal interaction index (CII) class (Grabisch & Roubens, 1999; Fumagalli et al., 2023). Define its corresponding player attribution as

$$\psi^k_i(v) = \sum_{\substack{C\ni i\\ 1\le|C|\le k}} \frac{1}{|C|} I^k_C(v). \tag{8}$$

Proposition 3.3 (Efficiency and Shapley recovery)**.** *Let I^k satisfy linearity, symmetry, and dummy. Then, the following are equivalent:*

$$\sum_{\substack{C\subseteq N\\ 1\le|C|\le k}} I^k_C(v) = v(N) - v(\emptyset) \quad \textit{for every } v, \tag{9}$$

and

$$\psi^k_i(v) = \phi_i(v) \quad \textit{for every } v \textit{ and } i \in N. \tag{10}$$

Thus, under these axioms, efficiency is equivalent to recovering the ordinary Shapley value through the member-wise aggregation in Equation (8). Recovery therefore does not uniquely determine the coalition-level representation. STI, Faith-Shap, and n-Shapley are all efficient and recover the same Shapley values, but allocate higher-order contributions differently across interaction orders (Sundararajan et al., 2020; Tsai et al., 2023; Bordt & von Luxburg, 2023). STI assigns contributions above order k only to order-k coalitions, whereas Faith-Shap and n-Shapley also allocate part of them to lower-order coalitions. The corresponding allocation coefficients and an illustrative example are given in Appendix B.

The representation in Theorem 3.2 is particularly useful for our diagnostic question because each order-k scalar term is the expectation μ_C of the context-dependent effect $\Delta_C v(T)$. We can therefore ask directly how well μ_C represents this effect across contexts. Section 4 studies this question through context variance and relates the remaining context dependence to higher-order interactions.

4 Diagnosing Context Dependence in Shapley Explanations

At each order, the representation in Section 3 contains scalar expectations $\mu_C = \mathbb{E}_{T\sim p_C}[\Delta_C v(T)]$ of context-dependent interaction effects. We now ask when these scalar expectations adequately represent the context-wise effects they summarize. We quantify the remaining variation by *Shapley context variance* and introduce *Shapley interaction energy* to measure the overall squared magnitude of the same context-wise effect. Small context variance indicates that the scalar expectation closely represents the underlying effects, whereas large variance indicates substantial context dependence. We then relate this variation directly to higher-order structure. Conditional context variance decomposes into nonnegative higher-order contributions. Extending this connection to full context variance, we further show that full context variance implies a lower bound on the aggregate interaction energy of immediate next-order extensions. Together, these results provide a theoretical basis for using context variance to identify scalar summaries for which further higher-order interaction analysis may be informative.

4.1 Context Variance and Interaction Energy

For any nonempty coalition $C \subseteq N$, define

$$\begin{aligned} &\mu_C \coloneqq \mathbb{E}_{T\sim p_C}[\Delta_C v(T)], \quad \sigma^2_C \coloneqq \mathrm{Var}_{T\sim p_C}[\Delta_C v(T)],\\ &\mathcal{E}_C \coloneqq \mathbb{E}_{T\sim p_C}[\Delta_C v(T)^2] = \mu^2_C + \sigma^2_C. \end{aligned} \tag{11}$$

Here, μ_C is the expected interaction effect in Equation (7), σ^2_C is its *Shapley context variance*, and $\mathcal{E}_C$ is the *Shapley interaction energy*.

Shapley context variance quantifies the squared deviation of context-wise effects from their scalar expectation μ_C, and thus directly reflects how representative this summary is across contexts. Small variance indicates effects concentrated around their mean, whereas large variance indicates heterogeneous effects obscured by averaging, regardless of the magnitude of μ_C. Shapley interaction energy measures the overall squared magnitude of the context-wise interaction under p_C. The decomposition $\mathcal{E}_C = \mu_C^2 + \sigma_C^2$ shows that it accounts for both the squared magnitude of the scalar expectation and the remaining variation across contexts. These quantities can be estimated using the permutation estimators in Appendix D.

4.2 Higher-Order Structure of Context Variance

Equation (3) also induces a relation between expected effects at adjacent interaction orders.

Proposition 4.1 (Mean effects across interaction orders)**.** *For every nonempty coalition $C \subsetneq N$, with $c = |C|$,*

$$\mu_C = \Delta_C v(\emptyset) + \frac{1}{c+1} \sum_{j \notin C} \mu_{C \cup j}. \tag{12}$$

This relation expresses the mean through a signed aggregation of next-order mean effects. These effects can cancel in their sum, so μ_C alone need not reveal their magnitude. Such cancellation can arise, for example, when complex models contain redundant or substitutable evidence that induces next-order effects of opposing signs. Context variance captures structure beyond this aggregation.

To characterize this variation, we use the permutation representation of p_C. Let $U_i \sim \mathrm{Unif}(0,1)$ independently and define $Q_C := \min_{i \in C} U_i$ and $T = \{j \notin C : U_j < Q_C\}$. Then $T \sim p_C$ and $Q_C \sim \mathrm{Beta}(1, c)$. Conditional on $Q_C = q$, each player outside C belongs to T independently with probability q. Thus, conditioning on $Q_C = q$ yields an independent Bernoulli(q) context. Define the corresponding conditional mean effect as

$$\mu_C(q) \coloneqq \mathbb{E}[\Delta_C v(T) \mid Q_C = q] = \sum_{T \subseteq N \setminus C} q^t (1-q)^{n-c-t} \Delta_C v(T). \tag{13}$$

Since $\mu_C = \mathbb{E}_{Q_C}[\mu_C(Q_C)]$, the law of total variance gives

$$\sigma_C^2 = \mathbb{E}_{Q_C}[\mathrm{Var}[\Delta_C v(T) \mid Q_C]] + \mathrm{Var}_{Q_C}[\mu_C(Q_C)]. \tag{14}$$

The first term measures variation across contexts at fixed inclusion probability. The second measures how the conditional mean varies with q. The first term admits the following interaction structure.

Theorem 4.2 (Higher-order structure of conditional context variance)**.** *For any nonempty coalition $C \subsetneq N$ and $q \in (0,1)$,*

$$\mathrm{Var}\left(\Delta_C v(T) \mid Q_C = q\right) = \sum_{\emptyset \neq R \subseteq N \setminus C} [q(1-q)]^{|R|} \mu_{C \cup R}(q)^2. \tag{15}$$

Each term in Equation (15) corresponds to a strict superset $C \cup R$ and is nonnegative. Conditional context variance therefore decomposes into squared higher-order conditional effects without cancellation. In particular, small conditional variance constrains these weighted higher-order contributions.

Theorem 4.2 gives an exact decomposition for conditional variance. For the full Shapley context variance used as our diagnostic, we obtain a direct implication for immediate next-order extensions.

Theorem 4.3 (Context variance implies next-order interaction energy)**.** *For any nonempty coalition $C \subsetneq N$,*

$$\sum_{j \notin C} \mathcal{E}_{C \cup j} = \sum_{j \notin C} \mathbb{E}_{T \sim p_{C \cup j}}\left[\Delta_{C \cup j} v(T)^2\right] \geq \frac{(c+1)^2}{c(n-c)} \sigma_C^2. \tag{16}$$

Thus, large context variance certifies the presence of interaction energy at the next order. Since $\mathcal{E}_{C \cup j} = \mu_{C \cup j}^2 + \sigma_{C \cup j}^2$, this energy may appear as a large expected interaction, continued context dependence, or both. Context variance therefore provides a theoretical basis for identifying scalar summaries for which higher-order inspection may be informative, and the same diagnostic can be applied recursively. Proofs and details of the conditional permutation representation are provided in Appendix C.

5 EVALUATION

Section 4 related context variance to the representativeness of scalar Shapley summaries and to higher-order interaction energy. We now examine these relationships in model explanations. We first illustrate them analytically in a simple attention model, then evaluate the relation between context variance and higher-order structure in exactly enumerated tabular games, and finally examine selective higher-order inspection in a language model.

5.1 ANALYTICAL ATTENTION EXAMPLE

We instantiate the two cases in Figure 1 with a simple analytical attention model to show that similar scalar Shapley values can correspond to markedly different context dependence and higher-order interaction structure. Consider a single-query softmax attention layer with attention score s_j and scalar projected value r_j for feature j. The coalition value is

$$v(S) = \frac{\sum_{j\in S} w_j r_j}{1 + \sum_{j\in S} w_j}, \qquad w_j = \exp(s_j), \tag{17}$$

where w_j is the unnormalized attention weight. The denominator includes a fixed reference token with weight 1 and value 0, so $v(\emptyset) = 0$.

For the board region, we set $w_a = w_b = w_c = 1$ and $r_a = r_b = r_c = 6$, so that the three features provide similar evidence through the same attention mechanism. The context-wise effects of a are

$$\Delta_a v(\emptyset) = 3, \qquad \Delta_a v(b) = \Delta_a v(c) = 1, \qquad \Delta_a v(bc) = 0.5,$$

giving $\mu_a = 1.500$ and $\sigma_a^2 = 1.167$. Thus, a has a strong effect when considered alone, but its marginal contribution decreases substantially once the neighboring board features are present. Its scalar Shapley value averages over this substantial change across contexts.

For the backpack region, we set $w_d = 1$, $r_d = 3.15$, and $w_e = w_f = 0.1$, $r_e = r_f = 0$, representing more isolated target evidence with weak surrounding contributions. The context-wise effects of d are 1.575, 1.500, 1.500, and 1.432 across $\emptyset$, $\{e\}$, $\{f\}$, and $\{e, f\}$, respectively, yielding $\mu_d = 1.502$ and $\sigma_d^2 = 0.0034$. Although μ_a and μ_d are nearly identical, a exhibits substantially larger context variance than d. Thus, the scalar summary is much more representative of the underlying context-wise effects for d than for a.

The context dependence of a is accounted for by higher-order interactions. In particular, $\Delta_{ab} v(\emptyset) = -2$, $\Delta_{ab} v(c) = -0.5$, and $\Delta_{abc} v(\emptyset) = 1.5$, which instantiate the recursive relations illustrated in Figure 1. The aggregate interaction energy of the immediate pairwise extensions is 5.50 for a and 0.0106 for d, showing that the large difference in context variance is accompanied by a correspondingly large difference in next-order interaction structure.

5.2 CONTEXT VARIANCE BEYOND THE SCALAR EXPECTATION

We examine systematically whether the relation observed in Section 5.1 appears in model explanations. We train an XGBoost regressor (Chen & Guestrin, 2016) on California Housing with $n = 8$ features and evaluate 200 test instances. For each instance x, we use the interventional game $v_x(S) = 1/|\mathcal{B}| \cdot \sum_{b\in\mathcal{B}} f(x_S, b_{\bar{S}})$, where $\mathcal{B}$ denotes the background set. All 2^8 coalition values are exactly enumerated. Experimental details are given in Appendix E.

Context variance reveals higher-order structure beyond the scalar expectation. For a coalition C, let $\mathcal{E}_{\text{next}}(C) := \sum_{j\notin C} \mathcal{E}_{C\cup\{j\}}$ denote the aggregate Shapley interaction energy of its immediate next-order extensions. Figures 2(a,b) plot the scalar expectation μ_C against its context variance σ_C^2 for all singleton and pair coalitions across the 200 local games, with darker points indicating larger $\mathcal{E}_{\text{next}}(C)$. For singletons, μ_C is the ordinary Shapley value, and for pairs, it is the corresponding second-order scalar interaction.

At both orders, coalitions with similar scalar expectations exhibit substantially different context variances and next-order interaction energies, including when μ_C is near zero. Thus, the scalar expectation μ_C alone does not establish whether the corresponding Shapley summary adequately represents the underlying context-wise effects.

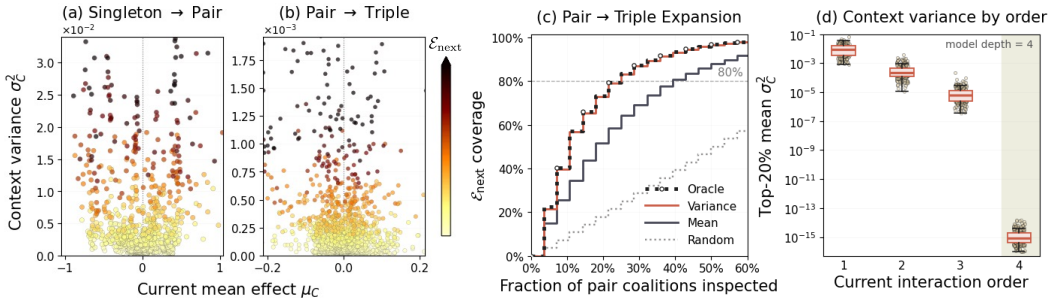


Figure 2: Context variance reveals higher-order structure beyond the scalar expectation. (a,b) Across 200 local games, singleton and pair coalitions with similar scalar expectations μ_C can have substantially different context variances σ_C^2 and next-order interaction energies $\mathcal{E}_{\text{next}}$. (c) For pair-to-triple inspection, ranking pairs by context variance captures next-order interaction energy with fewer inspected pairs than ranking by $|\mu_C|$, and nearly matches the oracle ranking by $\mathcal{E}_{\text{next}}$. (d) In the depth-4 XGBoost model, the upper tail of context variance collapses at order four, where interactions beyond order four are absent.

Selective higher-order inspection. To test whether context variance can identify scalar summaries for which further higher-order analysis is informative, we rank the 28 pair coalitions in each game and measure the fraction of total next-order interaction energy captured by the inspected pairs. Figure 2(c) compares rankings by context variance σ_C^2, the magnitude of the scalar expectation $|\mu_C|$ ("Mean" in the figure), random selection, and an oracle ranking by $\mathcal{E}_{\text{next}}$.

The variance ranking closely matches the oracle across inspection budgets. On average, capturing 80% of the next-order interaction energy requires inspecting 23.8% of the pairs under variance ranking, compared with 23.6% for the oracle and 41.2% for $|\mu_C|$. Thus, in these games, context variance identifies pairwise summaries whose further expansion reveals substantial higher-order effects more efficiently than ranking by the magnitude of their scalar expectations alone. We also evaluate SII, BII, and Faith-Shap, which show coverage similar to $|\mu_C|$ in this experiment. These results are reported in Appendix E.1.

Context variance at the interaction-order boundary. We further examine how context variance changes across interaction orders in the depth-4 XGBoost model. To summarize the upper tail of remaining context dependence, for each game and order c, we average the largest 20% of σ_C^2 among coalitions with $|C| = c$. Figure 2(d) shows the distribution of this quantity over the 200 local games. Context variance remains clearly nonzero through orders one to three, but becomes numerically negligible at order four, where the median is 8.27×10^{-16}.

This does not mean that the fourth-order interactions themselves vanish. Rather, their effects no longer vary with additional context because no higher-order interaction can arise beyond order four. This agreement with the model structure supports the interpretation of context variance as a measure of remaining context-dependent higher-order structure.

5.3 Higher-Order Interaction Structure in Language Explanations

We consider a language setting where exhaustive higher-order interaction analysis becomes impractical as the number of players grows. We use a BERT sentiment classifier (Devlin et al., 2019) on IMDB reviews (Maas et al., 2011) and treat words as players. For each sentence, absent words are replaced by `[MASK]` at their original positions, and the game value is defined by the positive-minus-negative logit margin. The sentence-level games considered here contain 16–24 players. Further implementation details are provided in Appendix E.2.

From a scalar word effect to higher-order interactions. Figure 3(a) compares each word's Shapley value with its context variance. The word `disappoint` has $\mu_i = -2.626$ and the largest context variance in the sentence, $\sigma_i^2 = 62.207$. Thus, its scalar Shapley value averages effects that

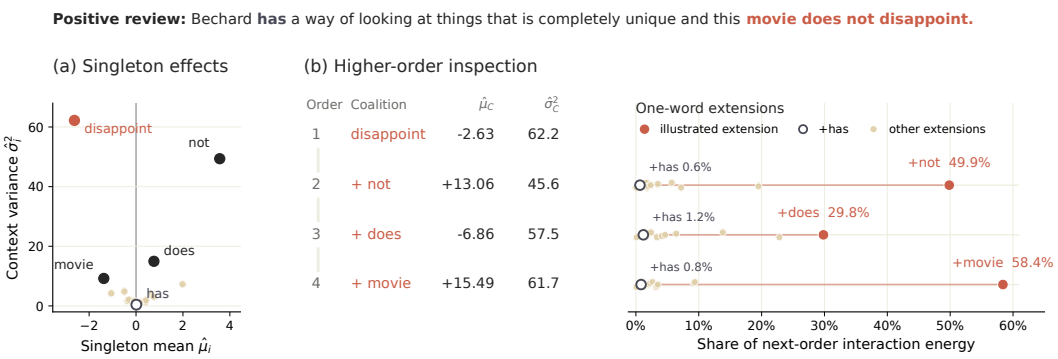


Figure 3: Context-dependent higher-order interactions in a BERT sentiment explanation. (a) Singleton Shapley values and context variances for a positive IMDB review, where `disappoint` has the largest context variance. (b) Higher-order inspection starting from textttdisappoint: the table reports $\hat{\mu}_C$ and $\hat{\sigma}_C^2$, and the right panels show each one-word extension's share of next-order interaction energy. Highlighted extensions form `movie does not disappoint`, while `+has` provides a low-energy contrast.

vary substantially with the surrounding words, indicating that further interaction inspection may be informative.

Figure 3(b) follows this high-variance effect across higher orders. Among the pairwise extensions of `disappoint`, `not` × `disappoint` accounts for 49.9% of its next-order interaction energy and retains substantial context variance. Continuing the inspection yields `does` × `not` × `disappoint` and `movie` × `does` × `not` × `disappoint`, which account for 29.8% and 58.4% of the corresponding next-order energy, respectively. The resulting path forms the phrase "movie does not disappoint," showing that the context dependence of the scalar word effect is associated with specific higher-order linguistic interactions. For comparison, extensions with `has` account for only 0.6%, 1.2%, and 0.8% of the corresponding next-order interaction energy, suggesting little additional interaction structure to inspect along this branch.

Potential for scalable higher-order exploration. The recursive pattern in Figure 3(b) suggests a way to narrow higher-order exploration in large interaction spaces. Rather than considering the combinatorial number of coalitions at each order, one can inspect the immediate extensions of a coalition with substantial context variance and continue where context dependence remains. In this example, such local expansion reveals the interaction structure associated with "movie does not disappoint" through a small sequence of candidate extensions. This illustrates the potential of context variance to support scalable higher-order exploration by progressively restricting the candidate space.

6 Conclusion

We studied what higher-order interaction structure is compressed when Shapley explanations summarize context-dependent effects by scalar expectations, and when these summaries adequately represent the underlying effects. Our order-wise representation makes this structure explicit by expressing the ordinary Shapley value through lower-order empty-context effects and expected higher-order interaction effects. It also connects the standard Shapley value to existing interaction indices.

We then introduced Shapley context variance to quantify variation beyond the scalar expectation. Its conditional form decomposes into nonnegative higher-order contributions, while full context variance lower-bounds the aggregate interaction energy of immediate next-order extensions. Empirically, similar scalar summaries can exhibit substantially different context dependence, and context variance helps identify where higher-order inspection is informative. Context variance is diagnostic rather than a direct recovery of the specific higher-order interactions responsible for this variation. Combining this diagnostic with efficient interaction estimation or adaptive exploration may provide a scalable path from scalar Shapley explanations to selective higher-order analysis.

AI USE STATEMENT

In this work, we used generative AI tools primarily to improve the fluency, clarity, and organization of the manuscript, and to assist with literature retrieval and analysis. We also used these tools as auxiliary support for mathematical derivations, including identifying relevant mathematical tools and developing intermediate proof steps, and for refining experimental implementations based on code and formulations developed by the authors. The research questions and target mathematical relationships were formulated by the authors, and the experimental settings and interpretation of results were determined by the authors. We did not use generative AI tools for other applicable tasks requiring disclosure, and the remaining required-disclosure tasks were not applicable to this work. We reviewed all AI-assisted work: final mathematical arguments and proofs were verified and revised by the authors, experimental code was checked and tested, and retrieved literature was verified against the original sources. 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 work does not involve human-subject studies or the collection of personal or sensitive data. The proposed analysis concerns the mathematical properties and interpretation of Shapley-based model explanations. We are not aware of specific ethical concerns arising from the methodology or experiments beyond the general need for caution when interpreting model explanations in real-world applications.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed proofs and derivations of the theoretical results in the appendix. We also describe the experimental settings, datasets and models, and additional implementation details required to reproduce the empirical analyses. Additional experimental results and supporting analyses are included in the appendix.

REFERENCES

Sebastian Bordt and Ulrike von Luxburg. From shapley values to generalized additive models and back. In *International Conference on Artificial Intelligence and Statistics*, pp. 709–745. PMLR, 2023.

Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & operations research*, 36(5):1726–1730, 2009.

Wonjoon Chang, Myeongjin Lee, and Jaesik Choi. Rethinking shapley value for negative interactions in non-convex games. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Pierre Dehez. On harsanyi dividends and asymmetric values. *International Game Theory Review*, 19(03):1750012, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

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.

Michel Grabisch. The interaction and möbius representations of fuzzy measures on finite spaces, k-additive measures: a survey. *Fuzzy measures and integrals—theory and applications*, pp. 70–93, 2000.

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.

John C Harsanyi. A simplified bargaining model for the n-person cooperative game. In *Papers in game theory*, pp. 44–70. Springer, 1982.

Xuanxiang Huang and Joao Marques-Silva. On the failings of shapley values for explainability. *International Journal of Approximate Reasoning*, 171:109112, 2024.

Salim I Amoukou and Nicolas Brunel. Regional explanations: Bridging local and global variable importance. *Advances in Neural Information Processing Systems*, 38:151601–151631, 2025.

Indra Kumar, Carlos Scheidegger, Suresh Venkatasubramanian, and Sorelle Friedler. Shapley residuals: Quantifying the limits of the shapley value for explanations. *Advances in Neural Information Processing Systems*, 34:26598–26608, 2021.

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.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5-part-2):64–79, 1972.

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.

Adrien Saumard. Weighted poincaré inequalities, concentration inequalities and tail bounds related to stein kernels in dimension one. *Bernoulli*, 25(4B):3978–4006, 2019.

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

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

Mukund Sundararajan, 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.

A Proofs for the Interaction Representation

A.1 Proof of Proposition 3.1

We derive Proposition 3.1 by repeatedly applying the relation

$$\Delta_R v(T \cup j) - \Delta_R v(T) = \Delta_{R \cup j} v(T)$$

along an ordering of the context. We first establish the resulting representation for a fixed permutation.

For a nonempty $U \subseteq S$ and a permutation $\pi \in \Pi(S)$, let π^U denote the set of players in $S \setminus U$ that precede every member of U in π.

Lemma A.1 (Order expansion for a fixed permutation)**.** *Let $R \subseteq N$ be a nonempty coalition and $S \subseteq N \setminus R$. For any integer $d \geq 1$ and every permutation $\pi \in \Pi(S)$,*

$$\Delta_R v(S) = \sum_{\substack{U \subseteq S \\ |U| < d}} \Delta_{R \cup U} v(\emptyset) + \sum_{\substack{U \subseteq S \\ |U| = d}} \Delta_{R \cup U} v(\pi^U). \tag{18}$$

Proof. We use induction on d.

For $d = 1$, write $\pi = (\pi_1, \ldots, \pi_s)$ and define $S_\ell = \{\pi_1, \ldots, \pi_\ell\}$, with $S_0 = \emptyset$. Adding the players in the order specified by π gives

$$\begin{aligned} \Delta_R v(S) &= \Delta_R v(\emptyset) + \sum_{\ell=1}^{s} \left(\Delta_R v(S_\ell) - \Delta_R v(S_{\ell-1})\right) \\ &= \Delta_R v(\emptyset) + \sum_{\ell=1}^{s} \Delta_{R \cup \pi_\ell} v(S_{\ell-1}). \end{aligned} \tag{19}$$

For each ℓ, $S_{\ell-1}$ is precisely the set of players preceding π_ℓ, i.e., $S_{\ell-1} = \pi^{\pi_\ell}$. Thus, Equation (19) is Equation (18) for $d = 1$.

Assume that Equation (18) holds for some $d < s$. For each $U \subseteq S$ with $|U| = d$, apply the $d = 1$ case to $\Delta_{R \cup U} v(\pi^U)$, following the order induced by π on π^U:

$$\Delta_{R \cup U} v(\pi^U) = \Delta_{R \cup U} v(\emptyset) + \sum_{j \in \pi^U} \Delta_{R \cup U \cup j} v(\pi^j). \tag{20}$$

Consider any $U' \subseteq S$ with $|U'| = d + 1$. Let j be the first member of U' in π and set $U = U' \setminus j$. Then $j \in \pi^U$, and the players preceding j are exactly $\pi^{U'}$, so $\pi^j = \pi^{U'}$. Conversely, every term in the double sum of Equation (20) determines such a unique pair (U', j). Therefore,

$$\sum_{\substack{U \subseteq S \\ |U| = d}} \sum_{j \in \pi^U} \Delta_{R \cup U \cup j} v(\pi^j) = \sum_{\substack{U' \subseteq S \\ |U'| = d+1}} \Delta_{R \cup U'} v(\pi^{U'}). \tag{21}$$

Substituting Equation (20) into the induction hypothesis yields

$$\Delta_R v(S) = \sum_{\substack{U \subseteq S \\ |U| < d+1}} \Delta_{R \cup U} v(\emptyset) + \sum_{\substack{U \subseteq S \\ |U| = d+1}} \Delta_{R \cup U} v(\pi^U), \tag{22}$$

which proves the result for $d + 1$.

For $d < s$, the induction step above gives the representation at $d + 1$. When $d = s$, the only term in the second sum is $\Delta_{R \cup S} v(\pi^S) = \Delta_{R \cup S} v(\emptyset)$, so it becomes part of the first sum at $d + 1$. For $d > s$, the second sum is empty and the representation is unchanged.

□

We now prove Proposition 3.1.

Proposition 3.1 (Order-k representation of a context-wise effect). Let $R, S \subseteq N$ be nonempty disjoint coalitions, with $r = |R|$ and $s = |S|$. For any integer k such that $r < k \leq n$,

$$\Delta_R v(S) = \sum_{\substack{U \subseteq S \\ r+u<k}} \Delta_{R\cup U} v(\emptyset) + \sum_{\substack{U \subseteq S \\ r+u=k}} \frac{k-r}{s} \sum_{T \subseteq S\setminus U} \binom{s-1}{t}^{-1} \Delta_{R\cup U} v(T). \tag{23}$$

Proof. Let $d = k - r$. Taking the expectation of Equation (18) over a uniformly random $\pi \in \Pi(S)$ gives

$$\Delta_R v(S) = \sum_{\substack{U \subseteq S \\ |U|<d}} \Delta_{R\cup U} v(\emptyset) + \sum_{\substack{U \subseteq S \\ |U|=d}} \mathbb{E}_\pi \left[\Delta_{R\cup U} v(\pi^U)\right]. \tag{24}$$

For $U \subseteq S$ with $|U| = d$ and $T \subseteq S \setminus U$, let $t = |T|$. The event $\pi^U = T$ occurs when the first t positions contain exactly the players in T, followed by one member of U. Thus,

$$\Pr(\pi^U = T) = \frac{t!\, d\, (s-t-1)!}{s!} = \frac{d}{s}\binom{s-1}{t}^{-1}. \tag{25}$$

Therefore,

$$\mathbb{E}_\pi \left[\Delta_{R\cup U} v(\pi^U)\right] = \frac{d}{s} \sum_{T \subseteq S\setminus U} \binom{s-1}{t}^{-1} \Delta_{R\cup U} v(T). \tag{26}$$

Substituting $d = k - r$ into Equation (24) gives Equation (23). □

A.2 Proof of Theorem 3.2

Theorem 3.2 (Order-k representation of the Shapley value). For any player $i \in N$ and interaction order $1 \leq k \leq n$,

$$\phi_i(v) = \sum_{u=0}^{k-2} \frac{1}{u+1} \sum_{\substack{U \subseteq N\setminus i \\ |U|=u}} \Delta_{i\cup U} v(\emptyset) + \frac{1}{k} \sum_{\substack{U \subseteq N\setminus i \\ |U|=k-1}} \mathbb{E}_{T\sim p_{i\cup U}} \left[\Delta_{i\cup U} v(T)\right], \tag{27}$$

where the first sum is empty for $k = 1$. For any nonempty coalition $C \subseteq N$, p_C denotes the distribution over contexts $T \subseteq N \setminus C$, defined by

$$p_C(T) = \frac{|C|}{n}\binom{n-1}{|T|}^{-1}. \tag{28}$$

Proof. We first verify the permutation interpretation of p_C stated in Section 3.2. For a uniformly random permutation $\pi \in \Pi(N)$ and a nonempty coalition $C \subseteq N$, let π^C denote the set of players preceding every member of C. For any $T \subseteq N \setminus C$, with $c = |C|$ and $t = |T|$, the event $\pi^C = T$ occurs when the first t positions contain exactly the players in T, followed by one member of C. Therefore,

$$\Pr(\pi^C = T) = \frac{t!\, c\, (n-t-1)!}{n!} = \frac{c}{n}\binom{n-1}{t}^{-1} = p_C(T). \tag{29}$$

The Shapley value has the permutation representation (Castro et al., 2009)

$$\phi_i(v) = \mathbb{E}_\pi \left[\Delta_i v(\pi^i)\right]. \tag{30}$$

For $k = 1$, Equation (27) follows directly from Equations (30) and (29).

For $k > 1$, applying Lemma A.1 to $\Delta_i v(\pi^i)$ with $d = k - 1$, using the order induced by π, gives for every permutation π

$$\Delta_i v(\pi^i) = \sum_{\substack{U \subseteq \pi^i \\ |U|<k-1}} \Delta_{i\cup U} v(\emptyset) + \sum_{\substack{U \subseteq \pi^i \\ |U|=k-1}} \Delta_{i\cup U} v(\pi^{i\cup U}). \tag{31}$$

In the second sum, the context in Lemma A.1 is $\pi^{i\cup U}$, because every member of π^i already precedes i.

Taking the expectation over π and grouping terms by U yields

$$\phi_i(v) = \sum_{u=0}^{k-2} \sum_{\substack{U\subseteq N\setminus i \\ |U|=u}} \Pr(U \subseteq \pi^i)\Delta_{i\cup U}v(\emptyset) + \sum_{\substack{U\subseteq N\setminus i \\ |U|=k-1}} \mathbb{E}_\pi\left[\mathbf{1}\{U\subseteq \pi^i\}\Delta_{i\cup U}v(\pi^{i\cup U})\right]. \tag{32}$$

For a fixed U with $|U| = u < k-1$, the event $U \subseteq \pi^i$ occurs exactly when i appears after every member of U. Since each of the $u+1$ players in $U \cup i$ is equally likely to appear last,

$$\Pr(U \subseteq \pi^i) = \frac{1}{u+1}. \tag{33}$$

For a fixed U with $|U| = k-1$, let $C = i \cup U$, so $|C| = k$. The event $U \subseteq \pi^i$ occurs exactly when i is the last member of C in π. Conditioned on $\pi^C = T$, the relative ordering of the members of C is uniform, and hence

$$\Pr(U \subseteq \pi^i \mid \pi^C = T) = \frac{1}{k}. \tag{34}$$

Using Equation (29),

$$\begin{aligned}\mathbb{E}_\pi\left[\mathbf{1}\{U\subseteq\pi^i\}\Delta_C v(\pi^C)\right] &= \frac{1}{k}\sum_{T\subseteq N\setminus C} p_C(T)\Delta_C v(T) \\ &= \frac{1}{k}\mathbb{E}_{T\sim p_C}\left[\Delta_C v(T)\right].\end{aligned} \tag{35}$$

Substituting Equations (33) and (35) into Equation (32) gives Equation (27).

Equation (31) also provides a permutation-sampling form of the order-k representation. Averaging its right-hand side over uniformly sampled permutations yields $\phi_i(v)$. □

B Relations to Existing Interaction Representations

Interaction indices extend attribution from individual players to coalitions of players and quantify the interaction associated with each coalition $C \subseteq N$. A broad class of such indices aggregates the context-wise discrete effects $\Delta_C v(T)$ over $T \subseteq N \setminus C$. In particular, cardinal interaction indices (CIIs) (Grabisch & Roubens, 1999; Fumagalli et al., 2023) take the form

$$I_C(v) = \sum_{T\subseteq N\setminus C} m_c(t)\Delta_C v(T), \qquad c = |C|, \quad t = |T|, \tag{36}$$

where the weight $m_c(t)$ depends only on the cardinalities of the coalition and its context. In the classical setting, interaction indices satisfying linearity, symmetry, and dummy admit this cardinal form. This class includes STI, Faith-Shap, and n-Shapley (Sundararajan et al., 2020; Tsai et al., 2023; Bordt & von Luxburg, 2023).

This appendix provides the details underlying the explanations in Section 3.3. We first prove Proposition 3.3, showing when coalition attributions recover the ordinary Shapley value under equal member-wise distribution. We then express the interaction representations considered in Section 3.3 in terms of Harsanyi dividends to make explicit how STI, Faith-Shap, and n-Shapley allocate higher-order contributions across the retained interaction orders.

B.1 Efficiency and Shapley Recovery

Let $I^k = \{I^k_C\}_{1\le|C|\le k}$ denote an interaction index of maximum order k satisfying linearity, symmetry, and dummy. We define the corresponding player attribution by

$$\psi^k_i(v) = \sum_{\substack{C\ni i \\ 1\le|C|\le k}} \frac{1}{|C|} I^k_C(v). \tag{37}$$

Proposition 3.3 (Efficiency and Shapley recovery). Let I^k satisfy linearity, symmetry, and dummy. Then, the following are equivalent:

$$\sum_{\substack{C\subseteq N\\ 1\le|C|\le k}} I^k_C(v) = v(N) - v(\emptyset) \quad \text{for every } v, \tag{38}$$

and

$$\psi^k_i(v) = \phi_i(v) \quad \text{for every } v \text{ and } i \in N. \tag{39}$$

Proof. ($\rightarrow$) Suppose that I^k satisfies Equation (38). By construction, ψ^k naturally inherits linearity from I^k. It also inherits symmetry. Swapping two symmetric players induces a one-to-one correspondence between the coalitions containing either player, with equal interaction attributions. If i is a dummy player, the dummy axiom for I^k gives $I^k_C(v) = 0$ for every coalition C containing i, and hence $\psi^k_i(v) = 0$. Finally, for efficiency,

$$\begin{aligned}\sum_{i\in N}\psi^k_i(v) &= \sum_{i\in N}\sum_{\substack{C\ni i\\ 1\le|C|\le k}} \frac{1}{|C|} I^k_C(v)\\ &= \sum_{\substack{C\subseteq N\\ 1\le|C|\le k}} I^k_C(v) = v(N) - v(\emptyset).\end{aligned} \tag{40}$$

Thus, ψ^k satisfies linearity, symmetry, dummy, and efficiency. By the uniqueness characterization of the Shapley value, $\psi^k_i(v) = \phi_i(v)$ for every $i \in N$.

($\leftarrow$) Conversely, suppose that $\psi^k_i(v) = \phi_i(v)$ for every player i. Then

$$\begin{aligned}\sum_{\substack{C\subseteq N\\ 1\le|C|\le k}} I^k_C(v) &= \sum_{i\in N}\psi^k_i(v) = \sum_{i\in N}\phi_i(v)\\ &= v(N) - v(\emptyset),\end{aligned} \tag{41}$$

where the first equality follows from the equal member-wise distribution in Equation (37). Hence, I^k satisfies efficiency. □

B.2 Higher-Order Dividend Allocation

The interaction representations considered in Section 3.3 differ in how dividends above the maximum order k are allocated among coalitions of order at most k. Appendix H of Bordt & von Luxburg (2023) gives a closely related comparison in terms of Shapley-GAM components, which, for a fixed input, correspond to the Möbius components of the associated value function. We restate these allocation rules in Harsanyi-dividend notation.

For a coalition C with $c = |C| \le k$, write

$$I^k_C(v) = \alpha_C + \sum_{\substack{L\supset C\\ |L|>k}} w^k_{c,|L|}\alpha_L, \tag{42}$$

where $\alpha_L = \Delta_L v(\emptyset)$ is the Harsanyi dividend. Thus, $w^k_{c,\ell}$ specifies the share of an order-ℓ dividend assigned to each retained coalition $C \subset L$ of size c. For STI, n-Shapley, and Faith-Shap, the corresponding weights are (Sundararajan et al., 2020; Tsai et al., 2023; Bordt & von Luxburg, 2023)

$$w^{k,\mathrm{STI}}_{c,\ell} = \begin{cases} 0, & c < k,\\ \binom{\ell}{k}^{-1}, & c = k,\end{cases} \tag{43}$$

$$w^{k,n\text{-SV}}_{c,\ell} = \sum_{r=0}^{k-c}\binom{k-c}{r}\frac{B_r}{1+\ell-c-r}, \tag{44}$$

$$w^{k,\mathrm{Faith}}_{c,\ell} = (-1)^{k-c}\frac{c}{k+c}\binom{k}{c}\frac{\binom{\ell-1}{k}}{\binom{\ell+k-1}{k+c}}, \tag{45}$$

where B_r denotes the r-th Bernoulli number.

The difference is therefore in the allocation of the same higher-order dividends. STI preserves the Harsanyi dividends below order k and allocates the higher-order terms only to order-k coalitions. In contrast, n-Shapley and Faith-Shap generally assign part of the higher-order terms to coalitions below order k. These different coalition-level representations can nevertheless recover the same ordinary Shapley value under the member-wise aggregation characterized in Proposition 3.3.

For the representation in Theorem 3.2, the STI allocation has a consequence that is particularly relevant to our analysis. Because contributions from interactions above order k are not redistributed to lower-order coalitions, they remain concentrated in the context-dependent order-k effect $\Delta_C v(T)$. Consequently, variation of $\Delta_C v(T)$ across contexts provides a natural diagnostic of whether the coalition is affected by unresolved higher-order interactions. This motivates the variance-based analysis in Section 4.

C Proofs for Context Variance and Interaction Energy

The results in Section 4.2 connect the mean and context variance of an interaction effect to higher-order interactions involving the same coalition. We prove these results using a conditional representation of the context distribution. We first relate the mean effect to next-order means, then derive the higher-order expression for conditional variance and the bound on full context variance.

C.1 Context Sampling and Conditional Mean–Variance Representations

We represent sampling from p_C by first drawing an inclusion probability and then including each player outside C independently with that probability. The resulting conditional independence lets us express the mean and variance through higher-order interaction effects in the subsequent proofs.

Fix a nonempty coalition $C \subseteq N$, with $c = |C|$ and $n = |N|$. Draw independent $U_i \sim \mathrm{Unif}(0,1)$ for $i \in N$ and set

$$Q_C = \min_{i \in C} U_i.$$

Include a player $j \notin C$ in the context when $U_j < Q_C$. Sorting the U_i gives a uniformly random permutation, so this context consists of the players preceding the first member of C.

Since $\Pr(Q_C > q) = (1-q)^c$, the density of Q_C is

$$b_c(q) = c(1-q)^{c-1}, \qquad 0 < q < 1. \tag{46}$$

The variables outside C are independent of Q_C. Conditional on $Q_C = q$, their inclusion indicators are therefore independent Bernoulli(q) variables. For a fixed $T \subseteq N \setminus C$, with $t = |T|$, integration over Q_C gives

$$\begin{aligned}\Pr\big(\{j \notin C : U_j < Q_C\} = T\big) &= \int_0^1 q^t (1-q)^{n-c-t} b_c(q)\, dq \\ &= \frac{c\, t!(n-t-1)!}{n!} = p_C(T).\end{aligned} \tag{47}$$

Thus the sampled context $T = \{j \notin C : U_j < Q_C\}$ has distribution p_C.

Write ν_q^M for the distribution on subsets of a finite set M that includes each player independently with probability q:

$$\nu_q^M(T) = q^{|T|}(1-q)^{|M|-|T|}, \qquad T \subseteq M.$$

For each nonempty C, the conditional mean from Section 4.2 can then be written as

$$\mu_C(q) = \mathbb{E}_{T \sim \nu_q^{N \setminus C}}[\Delta_C v(T)]. \tag{48}$$

Applying the laws of total expectation and total variance gives

$$\mu_C = \int_0^1 b_c(q)\mu_C(q)\, dq, \tag{49}$$

$$\sigma_C^2 = \mathbb{E}_{Q_C}[\mathrm{Var}(\Delta_C v(T) \mid Q_C)] + \mathrm{Var}_{Q_C}[\mu_C(Q_C)]. \tag{50}$$

C.2 Mean Effects Across Interaction Orders

The mean effect of a coalition can be expressed through its empty-context effect and the means of next-order interactions.

Proposition 4.1 (Mean effects across interaction orders). For every nonempty coalition $C \subsetneq N$, with $c = |C|$,

$$\mu_C = \Delta_C v(\emptyset) + \frac{1}{c+1} \sum_{j \notin C} \mu_{C \cup j}. \tag{51}$$

Proof. We first relate changes in the conditional mean $\mu_C(q)$ to next-order effects, then average this relation over the inclusion probability q. To isolate the change associated with each player, assign each $j \notin C$ its own inclusion probability x_j. With $M = N \setminus C$, define

$$F_C(x) = \sum_{T \subseteq M} \Delta_C v(T) \prod_{j \in T} x_j \prod_{j \in M \setminus T} (1 - x_j), \qquad x \in [0,1]^M. \tag{52}$$

This is the multilinear extension of the set function $T \mapsto \Delta_C v(T)$ (Owen, 1972). It satisfies $F_C(q\mathbf{1}) = \mu_C(q)$ and $F_C(\mathbf{1}_T) = \Delta_C v(T)$.

A standard property of multilinear extensions is that their mixed derivatives are the multilinear extensions of the corresponding discrete derivatives (Grabisch, 2000). Since F_C extends $T \mapsto \Delta_C v(T)$, differentiating with respect to $R \subseteq M$ yields the extension of

$$\Delta_R\big(\Delta_C v\big)(T) = \Delta_{C \cup R} v(T).$$

Therefore,

$$\partial_R F_C(q\mathbf{1}) = \sum_{T \subseteq M \setminus R} q^{|T|} (1-q)^{|M|-|R|-|T|} \Delta_{C \cup R} v(T) = \mu_{C \cup R}(q). \tag{53}$$

Since $F_C(q\mathbf{1}) = \mu_C(q)$, differentiating the restriction of F_C to the diagonal $x = q\mathbf{1}$ gives, by the multivariate chain rule,

$$\mu_C'(q) = \frac{d}{dq} F_C(q\mathbf{1}) = \sum_{j \notin C} \frac{\partial F_C}{\partial x_j}(q\mathbf{1}). \tag{54}$$

Using the preceding identity with $R = \{j\}$ yields

$$\mu_C'(q) = \sum_{j \notin C} \mu_{C \cup j}(q). \tag{55}$$

To recover the unconditional mean, we integrate against $b_c(q) = c(1-q)^{c-1}$. Using Equation (49) and integration by parts,

$$\begin{aligned}
\mu_C &= \int_0^1 c(1-q)^{c-1} \mu_C(q)\, dq \\
&= \mu_C(0) + \int_0^1 (1-q)^c \mu_C'(q)\, dq \\
&= \Delta_C v(\emptyset) + \sum_{j \notin C} \int_0^1 (1-q)^c \mu_{C \cup j}(q)\, dq \\
&= \Delta_C v(\emptyset) + \frac{1}{c+1} \sum_{j \notin C} \mu_{C \cup j}.
\end{aligned}$$

Here, $\mu_C(0) = \Delta_C v(\emptyset)$, and the last equality uses $b_{c+1}(q) = (c+1)(1-q)^c$, the inclusion-probability density associated with $C \cup j$. □

C.3 Higher-Order Structure of Conditional Variance

The conditional variance measures deviations from $\mu_C(q)$. The following theorem expresses it as a weighted sum of squared higher-order conditional means, with a nonnegative contribution from each strict superset of C.

Theorem 4.2 (Higher-order structure of conditional context variance). For any nonempty coalition $C \subsetneq N$ and $q \in (0,1)$,

$$\mathrm{Var}\left(\Delta_C v(T) \mid Q_C = q\right) = \sum_{\emptyset \neq R \subseteq N \setminus C} [q(1-q)]^{|R|} \mu_{C \cup R}(q)^2. \tag{56}$$

Proof sketch. We identify the higher-order components of the deviation from the conditional mean, then compute their variance contributions.

1. Express the context-wise effect as a function of independent Bernoulli context indicators.
2. Expand its deviation from the conditional mean in centered products and identify their coefficients with higher-order conditional means.
3. Use independence and zero means to eliminate cross terms, then sum the remaining variance contributions.

Proof. Fix $q \in (0,1)$ and let $M = N \setminus C$. Let $B = (B_j)_{j \in M}$ have independent $\mathrm{Bernoulli}(q)$ coordinates. By the conditional sampling representation in Section C.1, B has the same distribution as the context indicator vector conditional on $Q_C = q$. Since $F_C(\mathbf{1}_T) = \Delta_C v(T)$,

$$\mathrm{Var}\left(\Delta_C v(T) \mid Q_C = q\right) = \mathrm{Var}(F_C(B)), \qquad \mathbb{E}[F_C(B)] = F_C(q\mathbf{1}) = \mu_C(q). \tag{57}$$

To determine how higher-order interactions contribute to $F_C(B) - \mu_C(q)$, we express this deviation as a sum of products of the centered indicators $B_j - q$ and identify their coefficients with higher-order conditional means.

For $x \in [0,1]^M$, we first expand the polynomial $F_C(x)$ in powers of $x_j - q$. Since F_C is multilinear, substituting $x_j = q + (x_j - q)$ produces only products in which each centered coordinate appears at most once. Thus,

$$F_C(x) = a_\emptyset + \sum_{j \in M} a_{\{j\}}(x_j - q) + \sum_{\{j,k\} \subseteq M} a_{\{j,k\}}(x_j - q)(x_k - q) + \cdots.$$

Here R indexes the coordinates in each product, and the coefficients are taken at the fixed value of q. Evaluating at $x = q\mathbf{1}$ gives $a_\emptyset = F_C(q\mathbf{1}) = \mu_C(q)$.

For nonempty R, differentiating once in each coordinate indexed by R and evaluating at $x = q\mathbf{1}$ isolates a_R. Indeed, terms missing a coordinate in R vanish under differentiation, while terms containing additional coordinates retain a centered factor that vanishes upon evaluation. Hence,

$$a_R = \partial_R F_C(q\mathbf{1}) = \mu_{C \cup R}(q), \tag{58}$$

where the last equality follows from Equation (53). Substituting $x = B$ therefore yields

$$F_C(B) - \mu_C(q) = \sum_{\emptyset \neq R \subseteq M} \mu_{C \cup R}(q) \prod_{j \in R} (B_j - q). \tag{59}$$

Therefore, the coefficients identify the higher-order conditional means that determine deviations from $\mu_C(q)$. Squaring the expansion and taking expectations eliminates all cross terms, since the centered indicators $B_j - q$ are independent and have zero mean. For the squared terms, independence gives

$$\prod_{j \in R} \mathbb{E}[(B_j - q)^2] = \prod_{j \in R} \mathrm{Var}(B_j) = [q(1-q)]^{|R|}.$$

Therefore,

$$\begin{aligned}\mathrm{Var}\left(\Delta_C v(T) \mid Q_C = q\right) &= \mathbb{E}\left[(F_C(B) - \mu_C(q))^2\right] \\ &= \sum_{\emptyset \neq R \subseteq M} \mu_{C\cup R}(q)^2\, \mathbb{E}\left[\left(\prod_{j\in R}(B_j - q)\right)^2\right] \\ &= \sum_{\emptyset \neq R \subseteq M} [q(1-q)]^{|R|} \mu_{C\cup R}(q)^2.\end{aligned} \tag{60}$$

□

C.4 Bounding Context Variance by Next-Order Effects

The conditional variance decomposition identifies higher-order effects at a fixed inclusion probability. We now show that the full context variance requires a correspondingly large aggregate second moment of next-order effects under their context distributions $p_{C\cup j}$.

Theorem 4.3 (Context variance implies next-order interaction energy). For any nonempty coalition $C \subsetneq N$, with $c = |C|$,

$$\sum_{j\notin C} \mathcal{E}_{C\cup j} = \sum_{j\notin C} \mathbb{E}_{T\sim p_{C\cup j}}[\Delta_{C\cup j} v(T)^2] \geq \frac{(c+1)^2}{c(n-c)} \sigma_C^2. \tag{61}$$

Proof sketch. We bound both components of the full context variance through next-order conditional second moments.

1. Use the decomposition from Theorem 4.2 to bound the variance at fixed q.
2. Use the derivative of $\mu_C(q)$ and a beta variance inequality to bound the variance of the conditional mean.
3. Combine these bounds and integrate over Q_C, expressing the resulting weights through $p_{C\cup j}$.

Proof. Let $M = N \setminus C$ and $m = |M| = n - c$. By Equation (50),

$$\sigma_C^2 = \mathbb{E}_{Q_C}[\mathrm{Var}(\Delta_C v(T) \mid Q_C)] + \mathrm{Var}_{Q_C}(\mu_C(Q_C)).$$

To relate both terms to next-order effects, define

$$H_C(q) = \sum_{j\in M} \mathbb{E}_{T\sim \nu_q^{M\setminus j}}[\Delta_{C\cup j} v(T)^2].$$

We first bound the conditional variance at fixed $q \in (0,1)$. For each $j \in M$, applying Theorem 4.2 to $C \cup j$ and adding the squared conditional mean gives

$$\mathbb{E}_{T\sim \nu_q^{M\setminus j}}[\Delta_{C\cup j} v(T)^2] = \sum_{S\subseteq M\setminus j} [q(1-q)]^{|S|} \mu_{C\cup j\cup S}(q)^2,$$

where $S = \emptyset$ accounts for the squared mean.

Set $R = S \cup j$, so that $|S| = |R| - 1$. Multiplying by $q(1-q)$ and summing over j yields

$$\begin{aligned} q(1-q)H_C(q) &= \sum_{\emptyset\neq R\subseteq M} |R|[q(1-q)]^{|R|} \mu_{C\cup R}(q)^2 \\ &\geq \sum_{\emptyset\neq R\subseteq M} [q(1-q)]^{|R|} \mu_{C\cup R}(q)^2 \\ &= \mathrm{Var}(\Delta_C v(T) \mid Q_C = q). \end{aligned} \tag{62}$$

The inequality uses $|R| \geq 1$. For fixed $q \in (0,1)$, equality holds if and only if $\mu_{C\cup R}(q) = 0$ for every $R \subseteq M$ with $|R| \geq 2$. In particular, this condition holds when all effects involving C and at

least two additional players vanish across contexts. In that case, the conditional variance is exactly accounted for by next-order effects.

For the variance of the conditional mean, Equation (55), Cauchy-Schwarz, and Jensen's inequality give

$$\mu'_C(q)^2 = \left(\sum_{j\in M} \mu_{C\cup j}(q)\right)^2 \le m \sum_{j\in M} \mu_{C\cup j}(q)^2 \le m H_C(q).$$

We use a weighted Poincaré inequality, which bounds the variance of a function by a weighted average of its squared derivative. Specializing the beta-distribution inequality (Saumard, 2019) to $Q_C \sim \mathrm{Beta}(1, c)$ and applying it to the polynomial μ_C gives

$$\begin{aligned}\mathrm{Var}_{Q_C}\left(\mu_C(Q_C)\right) &\le \frac{1}{c+1}\mathbb{E}_{Q_C}[Q_C(1-Q_C)\mu'_C(Q_C)^2] \\ &\le \frac{m}{c+1}\mathbb{E}_{Q_C}[Q_C(1-Q_C)H_C(Q_C)].\end{aligned}$$

Combining this with Equation (62) in the total variance formula, we obtain

$$\sigma_C^2 \le \frac{n+1}{c+1}\mathbb{E}_{Q_C}[Q_C(1-Q_C)H_C(Q_C)], \tag{63}$$

using $1 + m/(c+1) = (n+1)/(c+1)$.

It remains to express this expectation under $p_{C\cup j}$. For fixed $j \in M$ and $T \subseteq M \setminus \{j\}$, write $t = |T|$. The weight of $\Delta_{C\cup j}v(T)^2$ in $\mathbb{E}_{Q_C}[Q_C(1-Q_C)H_C(Q_C)]$ is

$$\begin{aligned}\int_0^1 c(1-q)^{c-1}q(1-q)q^t(1-q)^{m-1-t}\,dq &= c\,\mathrm{B}(t+2, n-t) \\ &= \frac{c(t+1)!(n-t-1)!}{(n+1)!} \\ &= \frac{c(t+1)}{(c+1)(n+1)}p_{C\cup j}(T),\end{aligned}$$

where $p_{C\cup j}(T) = (c+1)t!(n-t-1)!/n!$ and B is a beta function. Substituting these weights into Equation (63) gives

$$\begin{aligned}\sigma_C^2 &\le \frac{c}{(c+1)^2}\sum_{j\notin C}\mathbb{E}_{T\sim p_{C\cup j}}\left[(|T|+1)\Delta_{C\cup j}v(T)^2\right] \\ &\le \frac{c(n-c)}{(c+1)^2}\sum_{j\notin C}\mathbb{E}_{T\sim p_{C\cup j}}\left[\Delta_{C\cup j}v(T)^2\right].\end{aligned}$$

The last inequality uses $|T|+1 \le n-c$, since $T \subseteq N \setminus (C\cup\{j\})$. Rearranging proves the theorem.

□

D Estimation of Context-Dependent Interaction Effects

Appendix A.2 shows that, for a uniformly sampled permutation π, the set π_C of players preceding every member of a nonempty coalition C follows the context distribution p_C. Hence, context-dependent effects under p_C can be estimated directly from uniform random permutations.

Let $\pi^{(1)}, \ldots, \pi^{(M)}$ be independent uniform permutations of N and define

$$z_m := \Delta_C v\left(\pi_C^{(m)}\right), \qquad m = 1, \ldots, M.$$

Then $z_1, \ldots, z_M$ are i.i.d. samples of $\Delta_C v(T)$ for $T \sim p_C$. We estimate

$$\hat{\mu}_C = \frac{1}{M}\sum_{m=1}^{M} z_m, \qquad \hat{\sigma}_C^2 = \frac{1}{M-1}\sum_{m=1}^{M}(z_m - \hat{\mu}_C)^2, \qquad \hat{\mathcal{E}}_C = \frac{1}{M}\sum_{m=1}^{M} z_m^2. \tag{64}$$

For $M \geq 2$, these are unbiased estimators of μ_C, σ_C^2, and $\mathcal{E}_C$, respectively. For singleton $C = \{i\}$, $\hat{\mu}_C$ reduces to the standard permutation estimator of the Shapley value. Thus, the same sampled context-wise effects provide estimates of the scalar expectation, its context variance, and interaction energy.

E Experimental Details

E.1 California Housing Details and Additional Comparisons

Experimental setup. We use the eight California Housing features and an 80/20 train-test split. The `XGBRegressor` consists of 100 depth-4 trees and achieves $R^2 = 0.7735$ on the test set. We evaluate 200 test instances using a fixed background set $\mathcal{B}$ of 256 training examples. For an instance x and coalition S, we define the interventional game as

$$v_x(S) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} f(x_S, b_{\bar{S}}),$$

where features in S take their values from x and the remaining features take their values from a background example b. For each instance, we enumerate all 2^8 coalition values. The reported scalar expectations, context variances, and interaction energies are therefore computed exactly for each finite-background game.

Selective-inspection evaluation. For each local game, we rank the 28 pair coalitions and measure the fraction of total next-order interaction energy captured by the top-ranked pairs, where the total is the sum of $\mathcal{E}_{\text{next}}(C)$ over all pairs. A triple extending multiple selected pairs contributes through each corresponding pair. The oracle ranking orders pairs directly by $\mathcal{E}_{\text{next}}(C)$ and gives the maximum coverage for a fixed number of inspected pairs. Coverage is computed separately for each of the 200 local games and then averaged across games. Reported confidence intervals are obtained from 2,000 paired bootstrap resamples of these 200 games, using the same resampled games for all ranking methods within each bootstrap replicate.

Table 1: Pair-to-triple next-order interaction-energy coverage on California Housing. Entries are means over 200 local games, with 95% paired game-bootstrap intervals in brackets. The first metric column reports the energy captured by three selected pairs. The second reports the fraction of pairs required to capture 80% of the energy. Random values are analytic expectations.

Ranking score	Coverage with 3 pairs (%)	Pairs inspected for 80% coverage (%)
Oracle $\mathcal{E}_{\text{next}}$	57.2 [55.4, 59.1]	23.6 [22.7, 24.6]
Context variance σ_C^2	**56.7** [54.8, 58.7]	**23.8** [22.9, 24.8]
Mean $\lvert\mu_C\rvert$	34.5 [32.9, 36.4]	41.2 [39.1, 43.1]
SII	35.1 [33.5, 36.9]	42.2 [40.1, 44.3]
BII	35.0 [33.3, 36.8]	41.8 [39.8, 44.0]
Faith-Shap	35.0 [33.4, 36.8]	42.0 [39.9, 44.2]
Random	10.7	82.1

Comparison with scalar interaction scores. The "Mean" baseline ranks pairs by $|\mu_C|$. At order two, μ_C is the corresponding maximal-order Shapley-Taylor interaction. We also compute exact SII and BII (Grabisch & Roubens, 1999), and order-2 Faith-Shap (Tsai et al., 2023) scores from the same enumerated games and rank pairs by their magnitudes. We do not report order-2 n-Shapley separately because its maximal-order pair terms coincide with SII and therefore induce the same pair ranking (Bordt & von Luxburg, 2023).

Table 1 reports two inspection criteria. With three selected pairs, context variance captures 56.7% of the next-order interaction energy, close to the oracle value of 57.2%, whereas $|\mu_C|$ and the three interaction indices capture about 35%. To capture 80% of the next-order interaction energy, context variance requires inspecting 23.8% of the pairs, compared with about 42% for the scalar interaction scores. Thus, in these games, ranking by context variance identifies pair coalitions whose immediate extensions contain substantial higher-order interaction energy using fewer inspected pairs than ranking by scalar interaction magnitude.

E.2 IMDB Details and Supporting Analyses

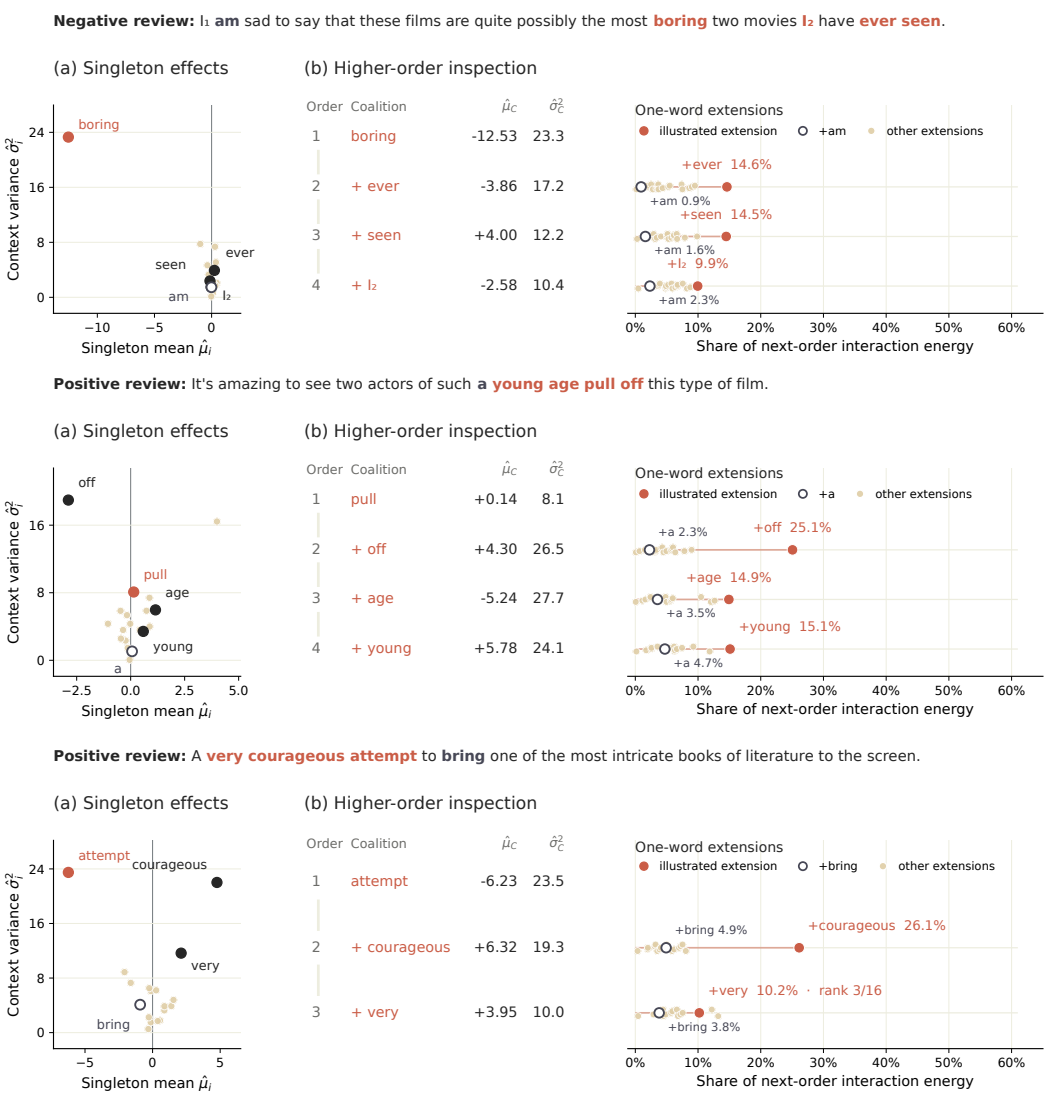


Figure 4: Additional higher-order interaction examples in BERT sentiment explanations. Each row shows (a) singleton Shapley values and context variances and (b) higher-order extensions starting from the highlighted singleton. Coral marks the illustrated extension at each order, and outlined markers show low-energy extensions from the same parent coalition. Percentages indicate each extension's share of next-order interaction energy.

Experimental setup and estimation. We use a BERT sentiment classifier on IMDB reviews and treat words as players. WordPiece tokens corresponding to the same word are grouped into one player, and special tokens are always included. For a coalition S, we define the game value as the logit margin

$$v_x(S) = f_{\text{pos}}(x_S) - f_{\text{neg}}(x_S),$$

where words outside S are replaced by `[MASK]` at their original positions. The sentence-level games contain 16–24 players. We estimate μ_C, σ^2_C, and $\mathcal{E}_C$ using the permutation estimators in Appendix D.

For Figure 3, we evaluate all 19 players in the sentence. We use 512 permutation contexts for effects through order three and 256 contexts for fourth-order effects. For a parent coalition C, the reported

share of an extension $C \cup j$ is

$$\frac{\hat{\mathcal{E}}_{C\cup j}}{\sum_{\ell \notin C} \hat{\mathcal{E}}_{C\cup \ell}},$$

which measures the fraction of estimated next-order interaction energy associated with that extension.

Illustrated higher-order extensions. We evaluate all pairwise extensions of `disappoint.not` $\times$ `disappoint` has the largest interaction-energy share and is shown in Figure 3. We then examine `does` $\times$ `not` $\times$ `disappoint`, which is fixed before evaluating its fourth-order extensions. Among these extensions, `movie` is highlighted in the figure. At each order, `+has` provides a low-energy comparison from the same parent coalition. Together, the highlighted extensions trace the higher-order interaction structure from `disappoint` to "movie does not disappoint."

Additional interaction examples. Figure 4 shows three additional sentence-level examples with 21, 20, and 18 players, respectively. Singleton statistics use 256 permutation contexts, and all higher-order effects shown in the figure use 512 contexts. The sentiment labels indicate the labels of the source reviews.

In the first example, `boring` has both the largest singleton mean magnitude and the largest context variance. Its leading extensions form `boring` $\times$ `ever` $\times$ `seen` $\times$ `I`$_2$ The final extension accounts for 9.9% of the next-order interaction energy, with the remaining energy distributed across several other extensions.

In the second example, `pull` has a near-zero singleton mean, but its highest-energy pairwise extension is `pull` $\times$ `off` (25.1%). The subsequent `age` and `young` extensions account for 14.9% and 15.1% of the corresponding next-order interaction energy.

In the third example, `courageous` $\times$ `attempt` is the highest-energy pairwise extension and changes the coalition mean from -6.23 to $+6.32$. Among the 16 third-order extensions, `very` ranks third in interaction energy.

Together, these examples show several forms of higher-order structure that are not apparent from singleton means alone, including distributed higher-order effects around a strongly attributed word, substantial pairwise interaction behind a near-zero singleton, and a reversal in the sign of the coalition mean after an extension.