Seeing Through Distractions: Stable Attribution via the Core

Sai Ganesh Nagarajan SGNAGARAJAN@IMADA.SDU.DK
University of Southern Denmark, IMADA

Toshinori Yamauchi T.YAMAUCHI@CHIBA-U.JP
Chiba University, Graduate School of Informatics

Hiroshi Kera KERA@CHIBA-U.JP
Chiba University; National Institute of Informatics

Abstract

Shapley-value-based explanations are a standard approach to feature attribution in machine learning. Yet, in many realistic settings, particularly in computer vision, groups of features can act as spurious contextual cues and bias a classifier toward a label. In such cases, Shapley values may systematically overestimate the importance of these groups. We formalize this effect through the notion of a *contextual distractor*. We show that, for a broad family of non-convex cooperative games, the least-core assigns more appropriate attribution to such distractor groups than general semivalue-based methods, including Shapley, Banzhaf, and weighted variants thereof. We derive explicit conditions under which this gap emerges, thereby identifying regimes in which the least-core provides a stable explanation while averaging-based attributions can be misleading. We complement our theory with experiments on an image-classification task, where the assumptions are verified empirically and the observed behavior aligns with our theoretical predictions.

1. Introduction

A central goal in explainable AI is to understand why a classifier makes a particular decision. For a fixed input, this often amounts to identifying a subset of features that best accounts for the prediction. In image classification, for instance, one may seek a small subset of pixels, patches, or segments whose presence suffices to preserve the predicted label under masking. The size constraint is important: without it, the full image would often be a trivial but uninformative explanation. This problem becomes especially challenging in computer vision, where explanations must scale to high-dimensional inputs and, at the same time, remain meaningful.

Game-theoretic methods have proved useful in this setting because they provide principled ways to reason about feature importance. In particular, the Shapley value [30] has become a standard attribution mechanism, due to its simplicity and its appealing axiomatic properties. However, two difficulties remain. The first, which is widely recognized, is computational: exact Shapley values require evaluating the characteristic function on all subsets of the N players. As a result, much of the literature has focused on approximation methods, especially sampling-based schemes [1, 14, 20, 24, 32], as well as more efficient computation under structural assumptions on the characteristic function [6, 11]. These guarantees have been further improved under assumptions such as submodularity [3, 4] and supermodularity [22]. However, such assumptions are unlikely to hold in natural settings, especially in computer vision.

© S.G. Nagarajan, T. Yamauchi & H. Kera.

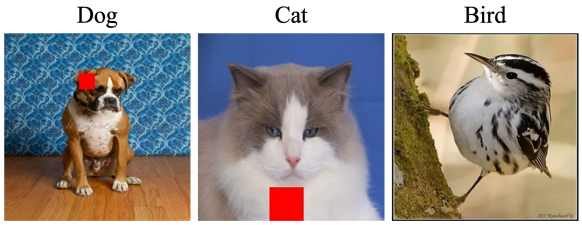

Figure 1: Controlled marker-bias setup. A red marker is added to all dog images and a subset of cat images, while bird images have no marker. This creates a dog-biased contextual cue that can become misleading for cat predictions. We refer to this red patch as a *contextual distractor*. In Section 3, we show that the Shapley value assigns substantially higher attribution to such patches than the least-core.

While the computational challenge of Shapley values is well understood, a second and more conceptual issue that is less studied, is the effect of averaging-based attributions in the presence of spurious contextual cues. In particular, a group of features may appear beneficial on average, even though it is harmful precisely in the contexts that are most informative for the prediction. In computer vision, for example, background patches, textures, or correlated artifacts can bias the classifier toward a label without capturing the actual object of interest.[1] See Figure 1 for an illustration.

This phenomenon is closely related to well-known shortcut examples in computer vision, where classifiers rely on background snow [26], water or habitat cues [5, 23], or correlated context such as a person with a skateboard [31]. These cases suggest that a feature group may appear useful on average, yet still be harmful in the contexts that matter most. This is precisely the kind of behavior that Shapley values can miss through contextual averaging. It therefore motivates moving beyond averaging-based attributions toward stability-based notions such as the *least-core* [25], which is the perspective we develop in this work. Please see Appendix A.1 for a discussion on related work.

Our Contributions. The main contributions of this paper are as follows.

1. We identify and formalize a shortcut-like effect through the notion of a *contextual distractor* (Section 2.1), namely a group of features that is harmful on important contexts but helpful on less informative ones.

2. We prove, in a broad class of non-convex games, conditions under which the least-core assigns strictly smaller attribution to a distractor group than any semivalue (Theorem 3), and in Appendix A.6 we specialize this comparison to common semivalues including Shapley, weighted Banzhaf, and Beta-Shapley.

3. We validate our assumptions on the controlled marker-bias setup in Figure 1, show empirically that Shapley over-attributes the distractor group relative to the least-core, and in Appendix A.7.1 use the least-core value to quantify the stability of several explanation methods.

1. Furthermore, the characteristic function that is induced from such effects is unlikely to satisfy regularities such as submodularity/supermodularity.

2. Explanations in Classification

In the section, we detail the basic formalism of how explanations in classification induce a characteristic game.[2] Shapley values have been widely used in image classification to visualize which pixels contribute most to a model's output, for example to a classifier's predicted label.

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the image and label spaces, respectively, and suppose $\mathcal{Y} = \{1, \ldots, K\}$. An image classifier is a map

$$f : \mathcal{X} \to \mathbb{R}^K,$$

where $f(x)$ is the vector of logits for image $x \in \mathcal{X}$. The predicted label is given by the largest logit.

To associate a cooperative game with a fixed image x, let N denote the set of pixels (or features) of x. For a subset $S \subseteq N$, let x_S denote the masked image that retains only the pixels in S and replaces the remaining pixels by a fixed baseline. The value of a coalition S is then defined through the classifier output on x_S. For notational consistency, let $v_f : 2^N \to \mathbb{R}$ denote the value function induced by f, and unless stated otherwise, for an image x with label y, define

$$v_f(S) := [f(x_S)]_y,$$

that is, the logit corresponding to class y. In particular, $v_f(S)$ may be negative. When the classifier is clear from context, we drop the subscript and simply write v. The game induced by f is then denoted by $\mathcal{G} = (N, v)$.

2.1. Contextual Distractors

In this section, we highlight a regime in which core-based notions can be especially useful for computer vision explanations. While Shapley values provide a natural average-case attribution, they need not capture whether a feature or group is actually *stably supportive* of the prediction. In particular, for non-monotone and non-convex games, Shapley values may assign nontrivial credit to a player that is in fact harmful in the contexts that matter most.

To illustrate the phenomenon, consider the game on three players $\{A, B, C\}$ with

$$v(\{A,B,C\}) = 60, \qquad v(\{A,B\}) = 45, \qquad v(\{A,C\}) = 95, \qquad v(\{B,C\}) = 45,$$

$$v(\{A\}) = 10, \qquad v(\{B\}) = 15, \qquad v(\{C\}) = 10, \qquad v(\emptyset) = 0.$$

This game is non-monotone, and one can verify that it is neither supermodular nor submodular. The Shapley vector is

$$\phi^{\text{SHAP}} = (27.5,\ 5,\ 27.5).$$

The least-core value is

$$\varepsilon^* = 25.$$

If we choose, from the least-core polytope, the point closest to the Shapley vector, we obtain

$$x^* = (35,\ -10,\ 35).$$

The main discrepancy is now concentrated on player B: the Shapley value assigns B a positive attribution of about 5, whereas the least-core assigns it a negative attribution of -10. By contrast,

2. Please refer to Appendix A.2, for all the game-theoretic preliminaries regarding semivalues and the least-core.

the attributions to A and C remain comparatively stable. In fact, any feasible least-core allocation x must satisfy

$$x(B) \leq -10,$$

The reason is that B behaves like what we call a *contextual distractor*. Indeed, the coalition $\{A, C\}$ already forms a highly informative context, with value

$$v(\{A, C\}) = 95,$$

yet adding B reduces the value to

$$v(\{A, B, C\}) = 60.$$

Thus, B is harmful precisely when inserted into an already high-value context. At the same time, B appears helpful in several smaller contexts, for example, $v(\{B\}) = 15$ and $v(\{A, B\}) = v(\{B, C\}) = 45$. This explains why an averaging-based method such as Shapley still assigns it positive value. In a vision setting, such a group may correspond to a patch, segment, or collection of pixels that acts as a misleading contextual cue.

Definition 1 (γ-Contextual Distractor) *Let $\gamma > 0$. A group $D \subset N$ is called a γ-contextual distractor if there exists a context $T \subseteq N \setminus D$ such that*

$$\Delta_D(T) := v(T \cup D) - v(T) \leq -\gamma.$$

Definition 2 (High-value Contexts) *For a threshold $\tau \in \mathbb{R}$, define*

$$H_\tau := \{T \subseteq N \setminus D : \ v(T) \geq \tau\}.$$

Given these two basic definitions [3], we will quantify the differences produced by semivalues and the least core on this group D that acts as contextual distractor. Note that we work under no major structural assumptions on v, such as additivity, monotonicity, super/sub-modularity etc. So in general, the game is still non-convex.

2.2. Semivalue–least-core gap

We will make the following assumptions that will help us analyze the bounds arising from the least-core and the Shapley values, in the presence of a group D that acts as a *contextual distractor* for the set of *high-value contexts*.

(A1) **Lipschitzness with Single Players.** There exists $\delta > 0$ such that for every player $i \in N$ and every set $S \subseteq N \setminus \{i\}$,

$$|v(S \cup \{i\}) - v(S)| \leq \delta.$$

(A2) **Harmful High-value Contexts.** There exists $\gamma > 0$ and $\beta > 0$ such that for every $T \subseteq N \setminus D$,

$$\Delta_D(T) \leq -\gamma \qquad \text{whenever } v(T) \geq \tau,$$

and

$$\Delta_D(T) \geq \beta \qquad \text{whenever } v(T) < \tau.$$

That is, D is γ-harmful on high-value contexts and β-helpful otherwise.

3. Any subset $S \subseteq N$ can be high value, but we are interested in contexts that are subsets of $N \setminus D$, to analyze how the group D interacts with the remaining subsets.

(A3) **Monotonicity of Context Value Probabilities.** Let $M := |N \setminus D|$. For each $k \in \{0, \dots, M\}$, let T be drawn uniformly from the family of all size k-subsets of $N \setminus D$, and define

$$q_k(\tau) := \Pr\big(v(T) \geq \tau \,\big|\, |T| = k\big) = \frac{1}{\binom{M}{k}} \sum_{\substack{T \subseteq N \setminus D \\ |T| = k}} \mathbf{1}\{v(T) \geq \tau\}.$$

We assume that the sequence $k \mapsto q_k(\tau)$ is nondecreasing, i.e.,

$$q_0(\tau) \leq q_1(\tau) \leq \cdots \leq q_M(\tau).$$

and that there exists $m^* \in \{0, 1, \dots, M\}$ such that $q_{m^*}(\tau) > 0$.

We stress that these assumptions are on the structural properties of the game induced by v, and are independent of the choice of explanation method and the least-core computation. Assumption (A1) is a natural Lipschitz-type regularity condition: in large-scale settings such as computer vision, the addition or removal of a single pixel or patch should not drastically alter the coalition value. Assumption (A2) is *designed* to encode the behavior of a contextual distractor: the group D is harmful when added to already informative contexts, but helpful on weak ones, reflecting shortcut-like behavior. Finally, Assumption (A3) captures the intuition that larger contexts are more likely to contain semantically meaningful signal, and hence are more likely to be high-value and at some layer m^*, there exists a high-value context.

Below we state our main theorem regarding how this gap between the semivalue attribution and core can be characterized with a sufficient condition. We request the readers to refer to Appendix A.6, on how this applies to commonly occuring semivalues, such as Shapley, weighted Banzhaf and Beta-Shapley.

Theorem 3 (Semivalue–least-core gap) *Given a game $\mathcal{G} = (N, v)$, let ε^* be the least-core value of $\mathcal{G}$, as given by the optimal objective value of (5). Let x^* be any allocation that is feasible for ε^*-core (from (5)) and let $\phi_\alpha^{\text{semi}}$ be the semivalue attribution with weights $\{\alpha_k\}_{k=0}^M$. Under Assumptions (A1)–(A3), define*[4]

$$\tilde{\beta} := \frac{\beta}{\gamma}, \qquad \tilde{\delta} := \frac{\delta |D|}{\gamma}, \qquad \tilde{\varepsilon}^* := \frac{\varepsilon^*}{\gamma},$$

If

$$\Pr_{\mu_\alpha}\big(v(T) \geq \tau\big) < \frac{1 + \tilde{\beta} - \tilde{\varepsilon}^*}{\tilde{\beta} + \tilde{\delta}} \;\; \textit{and} \;\; 1 + \tilde{\beta} - \tilde{\varepsilon}^* > 0,$$

then

$$\phi_\alpha^{\text{semi}}(D) - x^*(D) > 0.$$

Considering the lack of space, we defer the proof to Appendix A.5 and discuss how to interpretation our theorem in Appendix A.6.1.

3. Experiments

We conduct a controlled marker-bias experiment to test whether the theoretical Shapley–least-core gap appears in image classification.

4. This is done to make the quantities scale-invariant under the scaling of v by some constant $c > 0$.

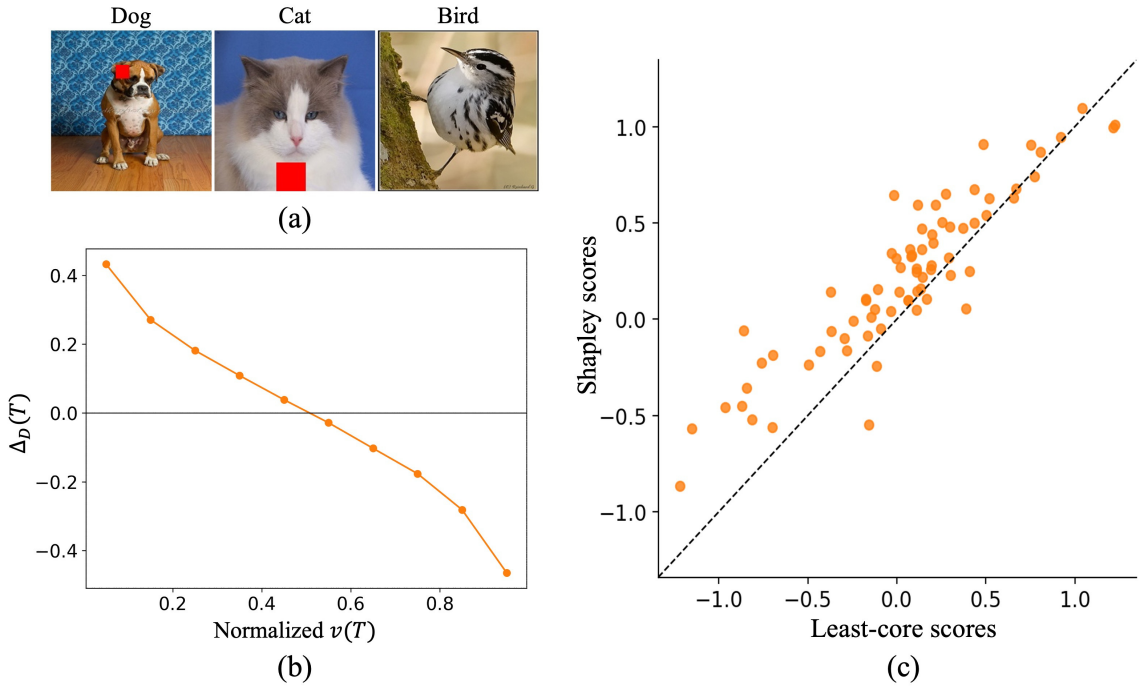

Figure 2: Analysis of the controlled marker-bias experiment. (a) Example images with synthetic markers. (b) For 100 cat-class images, the x-axis shows the normalized cat-class logit $v(T)$ of masked images, where coalitions T are sampled from non-marker regions. The y-axis shows the logit change caused by adding the marker region D, i.e., $v(T \cup D) - v(T)$. (c) Comparison of Shapley and least-core scores for the marker region over the same 100 cat-class images. Each point corresponds to one image, and the dashed line indicates equal attribution by the two methods. Refer to Appendix 3 for the details on how the experiments were conducted.

4. Conclusion

We end with a perspective of Hart and Mas-Colell [16], who note the historical movement from more complex cooperative solution concepts, such as the stable set and the core, toward simpler ones such as the Shapley value:

A simpler solution may be easier to study and apply, which increases its usefulness. However, it is important to look at different solution concepts, based on different postulates, because they illuminate the problem from different angles.

This viewpoint captures our message well: the Shapley value offers simplicity, but the least-core provides a complementary and, in our setting, more robust perspective on explanation.

5. Camera Ready Edits Summary

We list the main changes made to the camera-ready version. Most of the edits, especially the additional statements, were made in the Appendix due to space considerations.

1. Added a sentence in Pg. 5, just before Theorem 4, about our Assumptions, especially Assumption (A2).
2. Modified the statement of Theorem 4 to clarify that it applies to any least-core allocation and does not depend on a particular choice.
3. Added more information about practically selecting a point from the least-core polytope in Appendix A.4.
4. Updated our discussions on the Assumption and the interpretation of our Theorem in A.6.1 (Page 22 and 23).
5. Updated the experiments section in A.7.1, to clarify that the least-core stability score serves as benchmark for other methods and that our goal was to "quantify" stability properties of well-known methods using the core polytope.
6. Moved what was previously Remark 3 and a few sentences before that about the contextual distractor group D, into Appendix A.6.1 (Intepretation and Discussion section), to make space in the camera-ready version.

References

[1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

[2] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International conference on learning representations*, 2022.

[3] Eric Balkanski and Yaron Singer. Mechanisms for fair attribution. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 529–546, 2015.

[4] Eric Balkanski, Umar Syed, and Sergei Vassilvitskii. Statistical cost sharing. *Advances in Neural Information Processing Systems*, 30, 2017.

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

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

[7] Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. *Computational aspects of cooperative game theory*. Morgan & Claypool Publishers, 2011.

[8] 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.

[9] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018.

[10] Pradeep Dubey, Abraham Neyman, and Robert James Weber. Value theory without efficiency. *Mathematics of Operations Research*, 6(1):122–128, 1981.

[11] Shaheen S Fatima, Michael Wooldridge, and Nicholas R Jennings. A linear approximation method for the shapley value. *Artificial Intelligence*, 172(14):1673–1699, 2008.

[12] Ramón Flores, Elisenda Molina, and Juan Tejada. The shapley value as a tool for evaluating groups: Axiomatization and applications. In *Handbook of the Shapley Value*, pages 255–279. Chapman and Hall/CRC, 2019.

[13] Ian Gemp, Marc Lanctot, Luke Marris, Yiran Mao, Edgar Duéñez-Guzmán, Sarah Perrin, Andras Gyorgy, Romuald Elie, Georgios Piliouras, Michael Kaisers, et al. Approximating the core via iterative coalition sampling. *arXiv preprint arXiv:2402.03928*, 2024.

[14] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.

[15] Donald B Gillies. Solutions to general non-zero-sum games. *Contributions to the Theory of Games*, 4(40):47–85, 1959.

[16] Sergiu Hart, Andreu Mas-Colell, et al. The potential of the shapley value. *the Shapley value*, pages 127–137, 1988.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

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

[19] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pages 2907–2916. PMLR, 2020.

[20] Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. Fastshap: Real-time shapley value estimation. *ICLR 2022*, 2022.

[21] Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. *arXiv preprint arXiv:2110.14049*, 2021.

[22] David Liben-Nowell, Alexa Sharp, Tom Wexler, and Kevin Woods. Computing shapley value in supermodular coalitional games. In *International Computing and Combinatorics Conference*, pages 568–579. Springer, 2012.

[23] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

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

[25] Michael Maschler, Bezalel Peleg, and Lloyd S Shapley. Geometric properties of the kernel, nucleolus, and related solution concepts. *Mathematics of operations research*, 4(4):303–338, 1979.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[27] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[28] Lloyd S Shapley. Markets as cooperative games. Technical report, 1955.

[29] Lloyd S. Shapley. Cores of convex games. *International Journal of Game Theory*, 1(1):11–26, 1971. doi: 10.1007/BF01753431.

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

[31] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11070–11078, 2020.

[32] Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18, 2010.

[33] Kosuke Sumiyasu, Kazuhiko Kawamoto, and Hiroshi Kera. Identifying important group of pixels using interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6017–6026, 2024.

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

[35] Jiachen T Wang and Ruoxi Jia. Data banzhaf: A robust data valuation framework for machine learning. In *International conference on artificial intelligence and statistics*, pages 6388–6421. PMLR, 2023.

[36] Tom Yan and Ariel D Procaccia. If you like shapley then you'll love the core. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5751–5759, 2021.

Appendix A. Technical appendices and supplementary material

A.1. Related work

Our work is related to three strands of literature. First, Shapley-value-based explanations and, more broadly, semivalue methods such as Data Shapley, Beta Shapley, Data Banzhaf, and weighted Banzhaf have become standard tools for attribution in machine learning [21, 24, 34, 35]. Unlike this line of work, which remains averaging-based, we compare semivalue methods against a stable attribution method, which is the least-core.

Second, a growing literature studies limitations of Shapley-style explanations, including their dependence on the choice of value function or feature distribution, and the possibility of misleading feature-importance scores [1, 18, 19, 34]. Our work complements these critiques by identifying a specific mechanism behind such failures in vision-like settings, namely the presence of contextual distractors.

Third, our work is related to the literature on shortcut learning, spurious correlations, and explanation-based bias detection in classifiers [2]. This work documents that models can rely on spurious cues and studies whether post hoc explanations can reveal such behavior. Our contribution is about formalizing this phenomenon through coalitional games and the notion of a contextual distractor. Furthermore, we show both theoretically and empirically when stability-based explanations can differ sharply from averaging-based ones.

Recent work has also modified Shapley-style attribution considering interactions [8, 33], but these approaches remain within the averaging paradigm are modified Shapley values, whereas our focus is on the least-core as an alternative principle.

Finally, recent work has highlighted the least-core as a viable alternative to Shapley-style explanations and has developed computational methods for approximating it in large games [13, 36]. Our contribution is complementary: rather than focusing primarily on computation, we identify a specific reason to prefer the least-core in computer vision, namely the presence of contextual distractors, and we show how this leads to a separation between the least-core and Shapley/semivalue methods in a broad non-convex regime.

A.2. Game Theory Preliminaries

In this section, we introduce some game theory preliminaries, including the concept of the least-core and semivalues. Consider a transferable-utility game $\mathcal{G} = (N, v)$, where N denotes th set of players and $v : 2^N \to \mathbb{R}$ denotes the characteristic function or value function. A payoff vector is any vector $x \in \mathbb{R}^{|N|}$, where x_i denotes the payoff assigned to player $i \in N$. For any coalition $S \subseteq N$, we write

$$x(S) := \sum_{i \in S} x_i.$$

When $x(N) = v(N)$, the payoff vector is efficient. See [7] (Chapter 2) for more details.

The goal of an explanation method is precisely to distribute this total value among the features. In this view, each subset $S \subseteq N$ defines a coalition, and its worth is given by a scalar coalition value $v(S)$. Different explanation methods correspond to different solution concepts for performing this allocation/attribution.

A.3. Semivalues

Semivalues, introduced by Dubey, Neyman, and Weber [10], form a general class of value allocations that includes the well-known Shapley and Banzhaf values as special cases.

Theorem 4 (Representation of semivalues [10]) *A value function $\phi_\alpha^{\text{semi}}$ is a semivalue if and only if there exist coefficients*

$$\alpha_0, \alpha_1, \ldots, \alpha_{N-1} \in \mathbb{R} \qquad \textit{with} \qquad \sum_{k=0}^{N-1} \alpha_k = 1$$

such that, for every player $i \in N$,

$$\phi_\alpha^{\text{semi}}(i; v) = \sum_{k=0}^{N-1} \alpha_k \frac{1}{\binom{N-1}{k}} \sum_{\substack{S \subseteq N \setminus \{i\} \\ |S|=k}} \big(v(S \cup \{i\}) - v(S)\big).$$

We highlight the special case of Shapley values here:

Shapley value. If

$$\alpha_k = \frac{1}{N} \qquad \text{for all } k = 0, \ldots, N-1,$$

then

$$\phi^{\text{Shap}}(i; v) = \frac{1}{N} \sum_{k=0}^{N-1} \frac{1}{\binom{N-1}{k}} \sum_{\substack{S \subseteq N \setminus \{i\} \\ |S|=k}} \big(v(S \cup \{i\}) - v(S)\big), \tag{1}$$

which is exactly the Shapley value.

Group Semivalue. To define a semivalue attribution to a group of players D, we can view this group as a single player [12, 30] and we will consider subsets S of $N \setminus D$ and we let $M := |N \setminus D|$. Then, we can write the "group" semivalue as:

$$\phi_\alpha^{\text{semi}}(D; v) = \sum_{k=0}^{M} \alpha_k \frac{1}{\binom{M}{k}} \sum_{\substack{S \subseteq N \setminus D \\ |S|=k}} \big(v(S \cup D) - v(S)\big). \tag{2}$$

Where,

$$\alpha_0, \alpha_1, \ldots, \alpha_M \in \mathbb{R} \qquad \text{with} \qquad \sum_{k=0}^{M} \alpha_k = 1$$

Semivalues form a standard generalization of the Shapley value: they retain symmetry, dummy-player, and linearity-type properties, but need not satisfy efficiency. In particular, the Shapley value is the unique efficient semivalue, while the Banzhaf value is obtained by dropping efficiency and averaging uniformly over all coalitions.

Computing Semivalues. Let $\phi_\alpha^{\text{semi}}(.; v)$ denote the general group semivalue on subsets of $N \setminus D$ as in (2). We will drop the v to simplify notation and refer to it as $\phi_\alpha^{\text{semi}}(\cdot)$ Now the corresponding distribution over the subsets is represented by the probabilities $\{\alpha_k\}_{k=0}^M$.The key property of semivalues is that these weights depend only on the context size, so that $\alpha_k = \alpha(k; M, N)$.

This induces a distribution over subsets $T \subseteq N \setminus D$ of the form:

$$\mu_\alpha(T) = \frac{\alpha_{|T|}}{\binom{M}{|T|}}, \qquad M := |N \setminus D|,$$

that is, one first selects a context size $k \in \{0, \ldots, M\}$ according to the weights $\{\alpha_k\}_{k=0}^M$, and then samples T uniformly among all k-subsets of $N \setminus D$.

Thus, the group semivalue can be written as:

$$\phi_\alpha^{\text{semi}}(D) = \mathbb{E}_{T \sim \mu_\alpha}[\Delta_D(T)] \tag{3}$$

This viewpoint naturally suggests estimation by sampling contexts from the semivalue-induced distribution and averaging the corresponding marginal contributions. However, rigorous approximation guarantees are generally known only under additional structural assumptions on the game, such as versions of supermodularity or submodularity, as well as in certain other restricted settings [4, 22].

A.4. Least Core

The core provides a payoff vector x by seeking a *stable* allocation. Let $x(S) := \sum_{i \in S} x_i$ denote the total payoff assigned to a coalition $S \subseteq N$. In particular, when all players are considered together, we obtain the *grand coalition* N, and one typically imposes efficiency:

$$x(N) = v(N).$$

The core asks whether this grand-coalition allocation is stable against deviations by smaller coalitions. Specifically, if there exists a subset $S \subseteq N$ such that

$$x(S) < v(S),$$

then the players in S can improve upon the proposed allocation by leaving the grand coalition and forming their own coalition. Indeed, since coalition S can generate total value $v(S)$ on its own, there exists an alternative payoff vector for the members of S under which every player in S is strictly better off. For example, under transferable utility, one such deviation is obtained by redistributing the excess value equally:

$$y_i = x_i + \frac{v(S) - x(S)}{|S|}, \qquad i \in S,$$

so that $y(S) = v(S)$ and $y_i > x_i$ for every $i \in S$. Thus, a payoff vector lies in the core precisely when no coalition has such a profitable deviation. This brings us to the definition of the core attributed to Shapley and Gillies [15, 28].

Definition 5 *Given a game $\mathcal{G} = (N, v)$, a core of a game is the set of outcomes such that $x(S) \geq v(S)$ for every subset $S \subseteq N$.*

A natural question is whether the core is nonempty. In general, the answer is no: there exist games with an empty core; see Example 2.22 in [7]. There are two standard ways to address this issue.

One approach is to impose additional structure on the game. In particular, if the game is *supermodular* (or *convex*), meaning that for all $S, T \subseteq N$, it holds that $v(S \cup T) + v(S \cap T) \geq v(S) + v(T)$, then the core is nonempty, and moreover the Shapley value lies in the core [29]. Thus, in this structured setting, the Shapley value is not only axiomatically appealing, but also stable. In general non-convex games, however, no such guarantee is available. A second approach, which applies to arbitrary games, is to relax the notion of stability and consider the ε-core instead, which was introduced by [25]. Now Definition 5 changes as:

Definition 6 *Given a game $\mathcal{G} = (N, v)$ and a value $\varepsilon > 0$, the ε-core of a game is the set of outcomes such that $x(S) \geq v(S) - \varepsilon$, for every subset $S \subseteq N$.*

With Definition 6, we can see that there is some large enough ε, for which the core is non-empty. And this brings us to the definition of a least core of a game.

Definition 7 *Given a game $\mathcal{G} = (N, v)$, the least core of the game is defined as:*

$$\varepsilon^*(\mathcal{G}) := \inf\{\varepsilon : \varepsilon\text{-core of } \mathcal{G} \text{ is non-empty}\} \tag{4}$$

Computing the least-core. One can formulate the least-core as a linear program.

$$\min\ \varepsilon \quad \text{s.t.} \quad \sum_{i \in N} x_i = v(N), \qquad x(S) \geq v(S) - \varepsilon \ \ \forall S \subseteq N. \tag{5}$$

Suppose ε^* is the optimal objective value of this program, we refer to it as the *least-core value*. We also say that x^* is a *feasible* allocation for ε^*-core, if (x^*, ε^*) are feasible for the above LP. Let us denote the least-core polytope by $\mathrm{LC}(\varepsilon^*)$. Since $\mathrm{LC}(\varepsilon^*)$ is generally set-valued, one can obtain a canonical point-valued attribution by applying a selector to this polytope.

One natural choice is an analytic-center selector. Let the slack of coalition S at allocation x be

$$s_S(x) := x(S) - v(S) + \varepsilon^*.$$

Ignoring inequalities that are redundant or forced to be tight on the affine hull of $\mathrm{LC}(\varepsilon^*)$, the analytic center of the least-core is defined as

$$\sigma_{\mathrm{AC}}(v) \in \arg \min_{x \in \mathrm{relint}(\mathrm{LC}(\varepsilon^*))} -\sum_{S \subsetneq N} \log\big(s_S(x)\big).$$

Equivalently, this selector chooses the point in the relative interior of the least-core that maximizes the product of the coalition slacks:

$$\sigma_{\mathrm{AC}}(v) \in \arg \max_{x \in \mathrm{relint}(\mathrm{LC}(\varepsilon^*))} \prod_{S \subsetneq N} s_S(x).$$

Thus, the analytic-center selector favors allocations that are centrally located inside the least-core polytope, in the sense that no coalition constraint is made unnecessarily close to binding.

Another computationally convenient choice is a Euclidean projection selector. Let

$$r(v) \in \mathbb{R}^N$$

be any reference allocation satisfying $r(N) = v(N)$. This reference allocation may be chosen in many ways; for example, it may be the uniform allocation, a domain-specific baseline allocation, an empirical attribution vector, or the Shapley value when the Shapley value is available. The Euclidean least-core selector associated with r is

$$\sigma_r(v) := \arg \min_{x \in \mathrm{LC}(\varepsilon^*)} \frac{1}{2}\|x - r(v)\|_2^2.$$

Since the objective is strictly convex and $\mathrm{LC}(\varepsilon^*)$ is a convex polytope, this optimization problem has a unique solution whenever the least-core is nonempty. In particular, even if the reference allocation $r(v)$ itself does not belong to the least-core, $\sigma_r(v)$ returns the closest least-core allocation to $r(v)$ in Euclidean distance.

A classical alternative is the nucleolus [7], which also refines the least-core to a unique allocation by lexicographically minimizing coalition excesses. The nucleolus has a strong cooperative-game-theoretic pedigree, but it is often less practical computationally, since it requires solving a sequence of increasingly constrained optimization problems rather than a single convex program.

In practice, one can use approximate procedures that search for highly violated coalitions. Recent works such as [13, 36] pursue this direction through sampling-based and gradient-based schemes for approximately solving the least-core problem.

A.5. Proof of Theorem 3

Proof From the sufficient condition, there exists some $m^* \in \{0, \dots, M\}$ such that $q_{m^*}(\tau) > 0$. We may therefore w.l.o.g consider the m^* to be the smallest such m where $q_m(\tau) > 0$. Thus,

$$q_{m^*}(\tau) > 0 \qquad \text{and} \qquad q_k(\tau) = 0 \quad \forall k < m^*.$$

First, we look at upper bounding the attribution that a feasible ε^*-core allocation may assign to the group D. Now given that $q_{m^*}(\tau) > 0$, we have:

$$S \subseteq N \setminus D, \qquad |S| = m^*,$$

such that

$$v(S) \geq \tau$$

Now by Assumption (A3), we have that $q_M(\tau) > 0$. Therefore,

$$v(N \setminus D) \geq \tau.$$

Now let x^* be any point in the ε^*-core. Then

$$x^*(N) = v(N) \qquad \text{and} \qquad x^*(S) \geq v(S) - \varepsilon^* \quad \text{for all } S \subseteq N.$$

Applying this with $S = N \setminus D$ gives

$$x^*(N \setminus D) \geq v(N \setminus D) - \varepsilon^*.$$

Therefore

$$x^*(D) = x^*(N) - x^*(N \setminus D) \leq v(N) - \big(v(N \setminus D) - \varepsilon^*\big) = \Delta_D(N \setminus D) + \varepsilon^*.$$

Since $v(N \setminus D) \geq \tau$, Assumption (A2) yields

$$\Delta_D(N \setminus D) \leq -\gamma,$$

and hence

$$x^*(D) \leq -\gamma + \varepsilon^*. \tag{6}$$

Now, we will get a lower bound on the semivalue attribution for D.

By Assumption (A1), adding the elements of D one at a time and telescoping yields

$$|\Delta_D(T)| \leq \delta|D| \qquad \text{for every } T \subseteq N \setminus D.$$

In particular,

$$\Delta_D(T) \geq -\delta|D|.$$

Using the semivalue interpretation from (3) and performing a decomposition according to H_τ,

$$\phi_\alpha^{\text{semi}}(D) = \mathbb{E}\big[\Delta_D(T)\mathbf{1}_{H_\tau}\big] + \mathbb{E}\big[\Delta_D(T)\mathbf{1}_{H_\tau^c}\big].$$

On $H_\tau^c = \{T : \ v(T) < \tau\}$, Assumption (A2) gives $\Delta_D(T) \geq \beta$, while on H_τ we use the Lipschitz lower bound $\Delta_D(T) \geq -\delta|D|$. Hence

$$\phi_\alpha^{\text{semi}}(D) \geq \beta - (\beta + \delta|D|) \Pr_{\mu_\alpha}(H_\tau).$$

Thus, combining with the least-core upper bound, we get:

$$\phi_\alpha^{\text{semi}}(D) - x^*(D) \geq \gamma + \beta - \varepsilon^* - (\beta + \delta|D|) \Pr_{\mu_\alpha}\big(v(T) \geq \tau\big)$$

Therefore,

$$\phi_\alpha^{\text{semi}}(D) - x^*(D) > 0$$

holds when:

$$\Pr_{\mu_\alpha}\big(v(T) \geq \tau\big) < \frac{\gamma + \beta - \varepsilon^*}{\beta + \delta|D|}.$$

Scaling by γ gives the final form:

$$\Pr_{\mu_\alpha}\big(v(T) \geq \tau\big) < \frac{1 + \tilde{\beta} - \tilde{\varepsilon}^*}{\tilde{\beta} + \tilde{\delta}}.$$

Of course, $\gamma + \beta > \varepsilon^*$, ensures that the right-hand side remains positive and that the condition is meaningful.

Now, define,

$$A_{m^*}^+ := \sum_{k=m^*}^{M} \alpha_k,$$

which is precisely the probability that the semivalue selects a context of size at least m^*.

If $A_{m^*}^+ = 0$, then, $\Pr_{\mu_\alpha}(H_\tau) = \sum_{k=m^*}^{M} q_k(\tau)\alpha_k = 0$.

Hence the gap simplifies to

$$\phi_\alpha^{\text{semi}}(D) - x^*(D) \geq \gamma + \beta - \varepsilon^*.$$

Therefore, the condition $\gamma + \beta > \varepsilon^*$ alone suffices to ensure the gap is positive.

■

A.6. Additional Corollaries and Interpretation

Our main theorem, describes the gap between general semivalue distributions and the core. In this section, we expand on how the conditions look like for specific commonly occuring semivalues.

Claim A.1 *Let* $m^* := \min\{m \in \{0, \ldots, M\} : q_m(\tau) > 0\}$. *For any semivalue characterized by* $\{\alpha_k\}_{k=0}^M$, *if*

$$A^+_{m^*} := \sum_{k=m^*}^{M} \alpha_k > 0,$$

then:

$$\Pr_{\mu_\alpha}(v(T) \geq \tau) = \Pr_{\mu_\alpha}(v(T) \geq \tau \big| |T| \geq m^*) A^+_{m^*}$$

Proof

First we note that:

$$\Pr_{\mu_\alpha}(v(T) \geq \tau) = \sum_{k=0}^{M} q_k(\tau)\alpha_k = \sum_{k=m^*}^{M} q_k(\tau)\alpha_k$$

The last line is from the definition of m^*. When, $A^+_{m^*} > 0$, we may rewrite the above equation as follows:

$$\begin{aligned} \Pr_{\mu_\alpha}(v(T) \geq \tau) &= \sum_{k=m^*}^{M} q_k(\tau)\alpha_k \\ &= \frac{\sum_{k=m^*}^{M} q_k(\tau)\alpha_k}{A^+_{m^*}} A^+_{m^*} \\ &= \Pr_{\mu_\alpha}\big(v(T) \geq \tau \,\big|\, |T| \geq m^*\big) A^+_{m^*}. \end{aligned} \tag{7}$$

■

Claim A.2 *Let* $m^* := \min\{m \in \{0, \ldots, M\} : q_m(\tau) > 0\}$, *and let* T *be drawn according to the Shapley distribution* μ_{Shap} *on subsets of* $N \setminus D$. *Then*

$$\Pr_{\mu_{\text{Shap}}}\big(v(T) \geq \tau\big) = \frac{1}{M+1} \sum_{k=m^*}^{M} q_k(\tau).$$

Proof First by definition of m^*, one has $q_{m^*}(\tau) > 0$ and $q_k(\tau) = 0$ for all $k < m^*$

$$\sum_{k=m^*}^{M} q_k(\tau) = \sum_{k=0}^{M} q_k(\tau)$$

Now, one can use the conditional distributions $q_k(\tau)$ to represent $\Pr_{\mu_{\text{Shap}}}\big(v(T) \geq \tau\big)$ as follows:

$$\begin{aligned}\sum_{k=0}^{M} q_k(\tau) &= \sum_{k=0}^{M} \frac{1}{\Pr_{\mu_{\text{Shap}}}(|T| = k)} q_k(\tau) \Pr_{\mu_{\text{Shap}}}(|T| = k) \\ &= (M+1) \sum_{k=0}^{M} q_k(\tau) \Pr_{\mu_{\text{Shap}}}(|T| = k) \\ &= (M+1) \Pr_{\mu_{\text{Shap}}}\big(v(T) \geq \tau\big)\end{aligned} \tag{8}$$

In the first line, we can multiply and divide by $\Pr_{\mu_{\text{Shap}}}(|T| = k)$, since, under uniform sampling of the context size, these probabilities are always positive and exactly equal to $\frac{1}{M+1}$. This gives us the second line and the final line simply marginalizes over the context size. Combining the last line with our previous equation, we have the claim. ■

Claim A.2 helps to obtain a sufficient condition, always in terms of the Shapley distribution, even for general semivalues, if and when one is able to decouple the $\Pr_{\mu_\alpha}(v(T) \geq \tau)$ cleanly into $q_k(\tau)$ and α_k terms. This is possible by considering some loose upper bounds, but may not always gives us the tightest possible condition.

Corollary 8 (Shapley specialization) *Under the hypotheses of Theorem 3, suppose that the semivalue attribution method is the Shapley value, i.e.,*

$$\alpha_k = \frac{1}{M+1}, \qquad k = 0, \ldots, M.$$

Let $m^* \in \{0, \ldots, M\}$ be such that $q_{m^*}(\tau) > 0$, and let x^* be any allocation feasible for the ε^*-core. Then a sufficient condition for

$$\phi^{\text{Shap}}(D) - x^*(D) > 0$$

is

$$\Pr_{\mu_{\text{Shap}}}\big(v(T) \geq \tau \,\big|\, |T| \geq m^*\big) < \frac{(M+1)(1+\tilde{\beta}-\tilde{\varepsilon}^*)}{(M-m^*+1)(\tilde{\beta}+\tilde{\delta})}.$$

Proof From Theorem 3, we have that, the following condition leads to the gap between the semivalue and the least-core:

$$\Pr_{\mu_{\text{Shap}}}\big(v(T) \geq \tau\big) < \frac{1+\tilde{\beta}-\tilde{\varepsilon}^*}{\tilde{\beta}+\tilde{\delta}} \quad \text{and} \quad 1+\tilde{\beta}-\tilde{\varepsilon}^* > 0,$$

Now the Shapley values are described by $\alpha_k = 1/(M+1)$, for all $k \in \{0, 1, \ldots, M\}$, (see (1)).

Using Claim A.1 and substituting the α_k values appropriately, we get:

$$\Pr_{\mu_{\text{Shap}}}\big(v(T) \geq \tau \,\big|\, |T| \geq m^*\big) < \frac{(M+1)(1+\tilde{\beta}-\tilde{\varepsilon}^*)}{(M-m^*+1)(\tilde{\beta}+\tilde{\delta})}.$$

■

Since the Shapley distribution, uniformly samples the context size, the condition greatly simplifies. Note that if the RHS of the condition is greater than or equal to 1, then the gap is always positive.

Now, we look at the Bahnzaf and a general weighted Bahnzaf measures.

Corollary 9 (Weighted Banzhaf specialization) *Under the hypotheses of Theorem 3, suppose that the semivalue attribution method is the p-weighted Banzhaf value, i.e.,*

$$\alpha_k(p) = \binom{M}{k} p^k (1-p)^{M-k}, \qquad k = 0, \dots, M,$$

for some $0 < p < 1$. Let

$$A^+_{m^*}(p) := \sum_{k=m^*}^{M} \alpha_k(p) = \Pr\big(\mathrm{Bin}(M,p) \geq m^*\big).$$

Then the following hold.

(a) ***Monotonicity in*** p***.*** *It holds that:*

$$Q(p) := \Pr_{\mu_{\mathrm{Banz}(p)}}\big(v(T) \geq \tau\big) = \sum_{k=m^*}^{M} \alpha_k(p)\, q_k(\tau)$$

is nondecreasing in p. Consequently, the lower bound on the semivalue–least-core gap obtained by Theorem 3,

$$\phi^{\mathrm{Banz}(p)}(D) - x^*(D) \geq \gamma\Big[(1 + \tilde{\beta} - \tilde{\varepsilon}^*) - (\tilde{\beta} + \tilde{\delta})\, Q(p)\Big],$$

is nonincreasing in p.

(b) ***Simplification for small*** p***.*** *If*

$$p \leq \frac{1}{M+1},$$

then the sequence $k \mapsto \alpha_k(p)$ is nonincreasing. Hence a sufficient condition for

$$\phi^{\mathrm{Banz}(p)}(D) - x^*(D) > 0$$

is

$$\Pr_{\mu_{\mathrm{Sh}}}\big(v(T) \geq \tau\big) < \frac{M - m^* + 1}{M+1}\, \frac{1 + \tilde{\beta} - \tilde{\varepsilon}^*}{\tilde{\beta} + \tilde{\delta}} \qquad \textit{and} \qquad 1 + \tilde{\beta} - \tilde{\varepsilon}^* > 0.$$

Since $k \mapsto q_k(\tau)$ is nondecreasing by Assumption (A3), the reverse form of Chebyshev's sum inequality yields

$$\sum_{k=m^*}^{M} \alpha_k(p)\, q_k(\tau) \leq A^+_{m^*}(p)\, \frac{1}{M - m^* + 1} \sum_{k=m^*}^{M} q_k(\tau).$$

$$\frac{1}{M - m^* + 1} \sum_{k=m^*}^{M} q_k(\tau) < \frac{1 + \tilde{\beta} - \tilde{\varepsilon}^*}{A^+_{m^*}(p)\,(\tilde{\beta} + \tilde{\delta})} \qquad \text{and} \qquad 1 + \tilde{\beta} - \tilde{\varepsilon}^* > 0.$$

In particular, since $A^+_{m^*}(p) \leq 1$, the stronger but simpler sufficient condition

$$\frac{1}{M - m^* + 1} \sum_{k=m^*}^{M} q_k(\tau) < \frac{1 + \tilde{\beta} - \tilde{\varepsilon}^*}{\tilde{\beta} + \tilde{\delta}} \qquad \text{and} \qquad 1 + \tilde{\beta} - \tilde{\varepsilon}^* > 0$$

also suffices. By Claim A.2, this may be rewritten as

Proof Let

$$Q(p) := \Pr_{\mu_{\mathrm{Banz}(p)}}\big(v(T) \geq \tau\big) = \sum_{k=m^*}^{M} \alpha_k(p)\, q_k(\tau),$$

where

$$\alpha_k(p) = \binom{M}{k} p^k (1-p)^{M-k}, \qquad k = 0, \ldots, M.$$

By Theorem 3, a sufficient condition for

$$\phi^{\mathrm{Banz}(p)}(D) - x^*(D) > 0$$

is

$$Q(p) < \frac{1 + \tilde{\beta} - \tilde{\varepsilon}^*}{\tilde{\beta} + \tilde{\delta}} \qquad \text{and} \qquad 1 + \tilde{\beta} - \tilde{\varepsilon}^* > 0.$$

(a) We first prove that $Q(p)$ is nondecreasing in p. Differentiating,

$$Q(p) = \sum_{k=0}^{M} \binom{M}{k} p^k (1-p)^{M-k} q_k(\tau),$$

and using the standard binomial identity, we obtain

$$Q'(p) = M \sum_{k=0}^{M-1} \big(q_{k+1}(\tau) - q_k(\tau)\big) \binom{M-1}{k} p^k (1-p)^{M-1-k}.$$

By Assumption (A3), the sequence $k \mapsto q_k(\tau)$ is nondecreasing, and hence

$$q_{k+1}(\tau) - q_k(\tau) \geq 0 \qquad \text{for all } k = 0, \ldots, M-1.$$

Therefore $Q'(p) \geq 0$ for all $p \in (0,1)$, which shows that $Q(p)$ is nondecreasing in p. It follows that the lower bound on the gap,

$$\phi^{\mathrm{Banz}(p)}(D) - x^*(D) \geq \gamma\Big[(1 + \tilde{\beta} - \tilde{\varepsilon}^*) - (\tilde{\beta} + \tilde{\delta})\, Q(p)\Big],$$

is nonincreasing in p. Thus increasing p makes the sufficient condition harder to satisfy and reduces the guaranteed gap between the weighted Banzhaf attribution and the least-core allocation.

(b) Now suppose that

$$p \leq \frac{1}{M+1}.$$

We claim that the sequence $k \mapsto \alpha_k(p)$ is nonincreasing. Indeed,

$$\frac{\alpha_{k+1}(p)}{\alpha_k(p)} = \frac{M-k}{k+1} \cdot \frac{p}{1-p}, \qquad k = 0, \ldots, M-1.$$

Since $\frac{M-k}{k+1} \le M$ and $p/(1-p) \le 1/M$ under $p \le 1/(M+1)$, it follows that

$$\frac{\alpha_{k+1}(p)}{\alpha_k(p)} \le 1,$$

and hence $k \mapsto \alpha_k(p)$ is nonincreasing.

Since $k \mapsto q_k(\tau)$ is nondecreasing by Assumption (A3), the reverse form of Chebyshev's sum inequality for oppositely ordered sequences gives

$$\sum_{k=m^*}^{M} \alpha_k(p)\, q_k(\tau) \le \frac{1}{M-m^*+1} \left(\sum_{k=m^*}^{M} \alpha_k(p) \right) \left(\sum_{k=m^*}^{M} q_k(\tau) \right).$$

Recalling that

$$A^+_{m^*}(p) := \sum_{k=m^*}^{M} \alpha_k(p),$$

we obtain

$$Q(p) \le A^+_{m^*}(p)\, \frac{1}{M-m^*+1} \sum_{k=m^*}^{M} q_k(\tau).$$

Therefore, a sufficient condition for

$$\phi^{\mathrm{Banz}(p)}(D) - x^*(D) > 0$$

is

$$\frac{1}{M-m^*+1} \sum_{k=m^*}^{M} q_k(\tau) < \frac{1+\tilde{\beta}-\tilde{\varepsilon}^*}{A^+_{m^*}(p)(\tilde{\beta}+\tilde{\delta})} \qquad \text{and} \qquad 1+\tilde{\beta}-\tilde{\varepsilon}^* > 0.$$

Finally, since $A^+_{m^*}(p) \le 1$, the stronger but simpler sufficient condition

$$\frac{1}{M-m^*+1} \sum_{k=m^*}^{M} q_k(\tau) < \frac{1+\tilde{\beta}-\tilde{\varepsilon}^*}{\tilde{\beta}+\tilde{\delta}} \qquad \text{and} \qquad 1+\tilde{\beta}-\tilde{\varepsilon}^* > 0$$

also suffices.

By Claim A.2,

$$\Pr_{\mu_{\mathrm{Shap}}}\big(v(T) \ge \tau\big) = \frac{1}{M+1} \sum_{k=m^*}^{M} q_k(\tau),$$

and hence we get the following condition.

$$\Pr_{\mu_{\mathrm{Shap}}}\big(v(T) \ge \tau\big) < \frac{M-m^*+1}{M+1}\, \frac{1+\tilde{\beta}-\tilde{\varepsilon}^*}{\tilde{\beta}+\tilde{\delta}} \qquad \text{and} \qquad 1+\tilde{\beta}-\tilde{\varepsilon}^* > 0.$$

■

Corollary 10 (Monotonicity for the Beta-Shapley family) *Under the hypotheses of Theorem 3, suppose that the semivalue attribution method is the Beta-Shapley semivalue with parameters $a, b > 0$, i.e.,*

$$\alpha_k^{(a,b)} = \binom{M}{k} \frac{\mathrm{Beta}(k+b,\ M-k+a)}{\mathrm{Beta}(b,a)}, \qquad k = 0, \ldots, M.$$

Define

$$Q(a,b) := \Pr_{\mu_{a,b}}\big(v(T) \geq \tau\big) = \sum_{k=0}^{M} \alpha_k^{(a,b)}\, q_k(\tau).$$

If $k \mapsto q_k(\tau)$ is nondecreasing, then:

(a) for fixed b, the quantity $Q(a,b)$ is nonincreasing in a;

(b) for fixed a, the quantity $Q(a,b)$ is nondecreasing in b.

Consequently, the lower bound on the Beta-Shapley–least-core gap obtained by Theorem 3,

$$\phi^{\mathrm{Beta}(a,b)}(D) - x^*(D) \geq \gamma\Big[(1 + \tilde{\beta} - \tilde{\varepsilon}^*) - (\tilde{\beta} + \tilde{\delta})\, Q(a,b)\Big],$$

is nondecreasing in a and nonincreasing in b. In particular, increasing a enlarges the guaranteed gap, while increasing b reduces it.

Proof A standard beta-binomial representation of the Beta-Shapley weights is as follows: if

$$\Theta_{a,b} \sim \mathrm{Beta}(b,a), \qquad K_{a,b} \mid \Theta_{a,b} = \theta \sim \mathrm{Bin}(M,\theta),$$

then

$$\Pr(K_{a,b} = k) = \alpha_k^{(a,b)}, \qquad k = 0, \ldots, M.$$

Hence

$$Q(a,b) = \sum_{k=0}^{M} \alpha_k^{(a,b)}\, q_k(\tau) = \mathbb{E}\big[q_{K_{a,b}}(\tau)\big].$$

Now, for $\theta \in [0,1]$, define

$$g(\theta) := \mathbb{E}\big[q_{K(\theta)}(\tau)\big], \qquad K(\theta) \sim \mathrm{Bin}(M,\theta).$$

We first claim that g is nondecreasing in θ. Indeed, let $U_1, \ldots, U_M$ be i.i.d. $\mathrm{Unif}[0,1]$, and define

$$K(\theta) := \sum_{i=1}^{M} \mathbf{1}\{U_i \leq \theta\}.$$

If $0 \leq \theta_1 \leq \theta_2 \leq 1$, then $K(\theta_1) \leq K(\theta_2)$ almost surely. Since $k \mapsto q_k(\tau)$ is nondecreasing by Assumption (A3), it follows that

$$q_{K(\theta_1)}(\tau) \leq q_{K(\theta_2)}(\tau) \qquad \text{almost surely,}$$

and hence

$$g(\theta_1) \leq g(\theta_2).$$

Thus g is nondecreasing.

Next, fix a, and let $b_2 > b_1 > 0$. The density of $\Theta_{a,b} \sim \mathrm{Beta}(b, a)$ is proportional to

$$\theta^{b-1}(1-\theta)^{a-1}, \qquad 0 < \theta < 1.$$

Therefore

$$\frac{f_{a,b_2}(\theta)}{f_{a,b_1}(\theta)} = c\,\theta^{b_2-b_1}$$

for some constant $c > 0$, and this ratio is increasing in θ. Hence Θ_{a,b_2} is stochastically larger than Θ_{a,b_1}. Since g is nondecreasing,

$$Q(a, b_2) = \mathbb{E}[g(\Theta_{a,b_2})] \geq \mathbb{E}[g(\Theta_{a,b_1})] = Q(a, b_1).$$

Thus $Q(a, b)$ is nondecreasing in b.

Similarly, fix b, and let $a_2 > a_1 > 0$. Then

$$\frac{f_{a_2,b}(\theta)}{f_{a_1,b}(\theta)} = c\,(1-\theta)^{a_2-a_1},$$

which is decreasing in θ. Hence $\Theta_{a_2,b}$ is stochastically smaller than $\Theta_{a_1,b}$. Since g is nondecreasing,

$$Q(a_2, b) = \mathbb{E}[g(\Theta_{a_2,b})] \leq \mathbb{E}[g(\Theta_{a_1,b})] = Q(a_1, b).$$

Thus $Q(a, b)$ is nonincreasing in a.

Finally, Theorem 3 gives

$$\phi^{\mathrm{Beta}(a,b)}(D) - x^*(D) \geq \gamma\Big[(1 + \tilde{\beta} - \tilde{\varepsilon}^*) - (\tilde{\beta} + \tilde{\delta})\,Q(a, b)\Big].$$

Since $Q(a, b)$ is nonincreasing in a and nondecreasing in b, the claimed monotonicity of the lower bound follows immediately. ■

A.6.1. INTERPRETATION AND DISCUSSION OF OUR THEOREM

Below we discuss the key aspects in our main Theorem.

On Contextual Distractors: The group D should be interpreted as being sufficiently small so that the complement $N \setminus D$ still retains semantically meaningful signal. In particular, D need not be highly informative on its own, rather, its importance arises through its interaction with other contexts, where it can significantly alter the coalition value. We study a hypothetical contextual distractor D that may influence many contexts $T \subseteq N \setminus D$. Our analysis is not concerned with identifying or computing such a set D. Rather, it asks the following counterfactual question: if such a group D exists, even without being known explicitly, how would its importance be attributed by different solution concepts, such as the Shapley value and the least-core?

Constants in the Theorem: Firstly, we emphasize that the quantities $\tilde{\beta}$, $\tilde{\delta}$, and $\tilde{\varepsilon}^*$ are entirely dependent on the underlying game $\mathcal{G}$ and on the contextual distractor setup. In contrast, the coefficients $\{\alpha_k\}_{k=0}^{M}$ encode the choice of semivalue and therefore reflect the explanation method itself. The main theorem is useful precisely because it links these two sides: it relates structural, game-dependent properties of the contextual distractor regime to method-dependent quantities that determine how the semivalue averages over contexts of different sizes. Also, our theorem applies to *any* choice of feasible allocation from the least-core polytope. If one is concerned with choosing a specific allocation, see one possible approach discussed in Section A.4.

Now, the condition $\beta > \varepsilon^* - \gamma$ compares the helpfulness of D, quantified by β, with its harmfulness on high-value contexts, quantified by $-\gamma$, while incorporating the least-core stability cost ε^* (i.e., this is the cost to maintain stable coalitions). As $\varepsilon^* \downarrow 0$, the effect of this stability term becomes negligible, and the gap between the two attribution methods become more pronounced since D is beneficial outside the high-value regime ($\beta > 0$) but harmful within it ($-\gamma < 0$).

Once the first condition is satisfied, the second condition controls the probability of encountering high-value contexts, conditional on sampling from sufficiently large contexts. This can be seen by a slightly different but an equivalent form of our sufficient condition (see Claim A.1), which applies when $A^+_{m^*} > 0$, and is given by:

$$\Pr_{\mu_\alpha}(v(T) \geq \tau \big| |T| \geq m^*) < \frac{1 + \tilde{\beta} - \tilde{\varepsilon}^*}{A^+_{m^*}\big(\tilde{\beta} + \tilde{\delta}\big)}.$$

Intuitively, if the semivalue distribution places little or no probability mass on these tail layers, then it effectively ignores the large contexts where the most informative signal may reside. In that case, the semivalue can miss precisely those important contexts that drive the core-based bound, leading to a large discrepancy between the two attribution methods. Conversely, if the semivalue assigns substantial mass to the tail, then it is more likely to sample these high-value contexts, and the resulting gap between the semivalue and the core is correspondingly reduced. Note that, Assumption (A2) does not itself imply the conclusion of the theorem. Rather, it captures the structural behavior of the distraction phenomenon. If Assumption (A2) alone were enough, then one would obtain a gap unconditionally. Instead, our theorem only provides a sufficient condition, and the existence of the gap still depends on how the semivalue weights contexts and on the least-core slack ε^*.

Finally, we stress that this sufficient condition is easy to check simply by sampling according to the semivalue distribution and estimating the probabilities.

Additionally, the aforemnetioned corollaries help clarify how the choice of semivalue affects performance in the contextual distractor regime. In particular, they show that semivalues whose parameters place greater weight on larger context sizes are, in this setting, better aligned with the underlying structure of the problem and can therefore be more suitable than methods that concentrate on smaller contexts.

A.7. Additional Details for Experiments

We conduct a controlled marker-bias experiment to test whether the theoretical Shapley–least-core gap appears in image classification. We construct a three-class dog/cat/bird task and train a ResNet-50 [17] classifier. As shown in Fig. 2(a), a red marker region D is added to 100% of dog images

and 30% of cat images, while bird images have no marker. This makes D strongly associated with dog, so that it can act as a misleading contextual cue for cat images.

We first verify that D behaves as a contextual distractor for cat predictions. Each image is divided into a grid of patches, which define the player set N. The marker region $D \subseteq N$ is represented by the set of patches overlapping the red marker. We then sample coalitions $T \subseteq N \setminus D$ from the non-marker patches and compute $\Delta_D(T) = v(T \cup D) - v(T)$. As shown in Fig. 2(b), $\Delta_D(T)$ decreases as the context becomes stronger and changes from positive to negative. Thus, the marker helps cat prediction in weak contexts but suppresses it when strong cat evidence is already present. Figure 2(c) shows that Shapley assigns larger values to D than the least-core for most images. This supports our theoretical prediction that Shapley can over-credit a contextual distractor by averaging over contexts, whereas least-core penalizes features that destabilize high-value coalitions.

We compare the attribution assigned to the marker region and these are computed in the following manner. Images (224×224) are divided into an 8×8 grid, yielding 64 patch players. Shapley values are estimated by Monte Carlo permutation sampling using 100 random permutations per image. For each permutation, we sequentially add the 64 patches and record the change in the model output, resulting in 100×64 sampled output changes per image. For least-core, we avoid enumerating all coalitions and instead solve the linear program on sampled constraints [36]. We sample 6,400 coalitions by independently including each player with probability $p = 0.5$, and add the empty and grand coalitions. Optimization is implemented with CVXPY. Neural-network evaluations are performed with PyTorch on CUDA, while the CVXPY optimization is typically run on CPU.

Below we use the least-core polytope to define a notion of "explanation quality" for several well-known explanation methods, and evaluate their *stability score* through the least-core value.

A.7.1. EPSILON EVALUATION FOR EXPLANATION METHODS

We evaluate explanation methods using a ResNet-50 classifier [17] on 100 ImageNet test images. Each image is resized to 224×224 and divided into an 8×8 grid, yielding 64 patch players.

For a fixed attribution vector x, we define the epsilon value as the worst coalition deficit:

$$\varepsilon(x) = \max_{S \subseteq N} \left(v(S) - x(S) \right),$$

where $x(S) = \sum_{i \in S} x_i$. In practice, this maximum is approximated over 4,096 sampled coalitions.

For Grad-CAM [27] and Grad-CAM++ [9], the explanation methods do not directly produce an epsilon value. We therefore convert each heat map into a patch-level allocation x by averaging the heat map within each patch. Since these heat maps do not necessarily satisfy the efficiency constraint, we rescale x so that

$$\sum_i x_i = v(N).$$

We then evaluate $\varepsilon(x)$ by sampled coalitions.

For Monte Carlo Shapley, we compute $\varepsilon(x)$ as a post-hoc evaluation of the resulting attribution vector. Shapley values are estimated by Monte Carlo permutation sampling using 64 random patch orderings per image $64 \times 64 = 4,096$ sampled output changes per image).

For least-core, we use the same 64 patch players and solve the least-core optimization over 4,096 sampled coalition constraints. We directly report the optimized ε returned by the least-core solver.

Thus, the least-core epsilon is not a post-hoc evaluation of a generated heat map, but the objective value obtained during the least-core optimization.

Lower epsilon indicates that the explanation better satisfies the sampled core constraints and is therefore more stable from a coalitional perspective. Table 1 summarizes the results. Here, the least-core is added as the benchmark value of ε^* that can be ideally achieved.

Table 1: Epsilon values on 100 ImageNet test images. Lower is better.

	Grad-CAM	Grad-CAM++	Shapley-MC	Least-Core (Benchmark)
ε (mean $\pm$ std)	0.6990 ± 0.1445	0.6899 ± 0.1397	0.6851 ± 0.1637	$\mathbf{0.5822} \pm \mathbf{0.1389}$