# Local Additive Feature Attribution: A Mathematical Taxonomy and Reporting Checklist

Rebecca Afriyie Sarpong and Daniel Commey



***Abstract*—Feature-attribution methods are central to explainable artificial intelligence. Their assumptions are expressed in several mathematical languages: cooperative-game values, path integrals, gradient operators, perturbation distributions, and backpropagation rules. This survey proposes a common framework for local additive feature attribution. It organizes Shapley, path-based, gradient/backpropagation, perturbation, and CAM-style methods around five specification choices: value function, reference, path, perturbation distribution, and conservation rule. It then compares these methods through an axiom-by-method matrix and links common failure modes, including baseline sensitivity, off-manifold perturbations, sanity-check failures, adversarial manipulation, and method disagreement, to the assumptions that produce them. Finally, the survey proposes a ten-item reporting checklist for studies that use local additive attributions. The central message is that attribution results are meaningful only relative to the mathematical assumptions under which they are defined, and that those assumptions should be reported.**

***Index Terms*—Explainable AI, local feature attribution, additive feature attribution, Shapley values, Integrated Gradients, saliency, LRP, DeepLIFT, Grad-CAM, axiomatic methods, faithfulness, reporting checklist.**

## I. Introduction

FEATURE attribution has become one of the dominant interfaces between complex predictive models and their human users. Faced with a model that maps a $d$-dimensional input $x \in \mathbb{R}^d$ to a prediction $f(x) \in \mathbb{R}$, a user asks: *which features mattered, and by how much?* A feature-attribution method answers with a vector $\boldsymbol{\phi}(f, x) \in \mathbb{R}^d$ that distributes credit (or blame) for the prediction among the input features. Over the last decade, dozens of such methods have been proposed, including Integrated Gradients [1], SHAP [2], LIME [3], Grad-CAM [4], LRP [5], DeepLIFT [6], and many others, each motivated by a different intuition about what an explanation should be.

The rapid growth of attribution methods has not been matched by comparable clarity about their assumptions. Two methods applied to the same prediction frequently produce explanations that agree only on coarse features and disagree, sometimes sharply, on their relative ordering. Krishna et al. [7] documented this phenomenon at scale across six attribution methods and four tasks, the median Spearman rank correlation between explanations of the same prediction was below 0.5. They termed it the *disagreement problem*. Bilodeau et al. [8] established a complementary negative result: no feature-attribution method can simultaneously satisfy a small set of reasonable desiderata, so disagreement among methods is mathematically unavoidable. Earlier critiques along similar lines include [9], [10]. Empirical evaluations on the same benchmarks reach incompatible conclusions about which method is most "faithful" [11], [12]. And it is now well documented that adversarial manipulations can produce arbitrary attribution maps with negligible changes to the prediction [13]–[15]. These pathologies often trace to *implicit mathematical choices* that different methods make and that users rarely see.

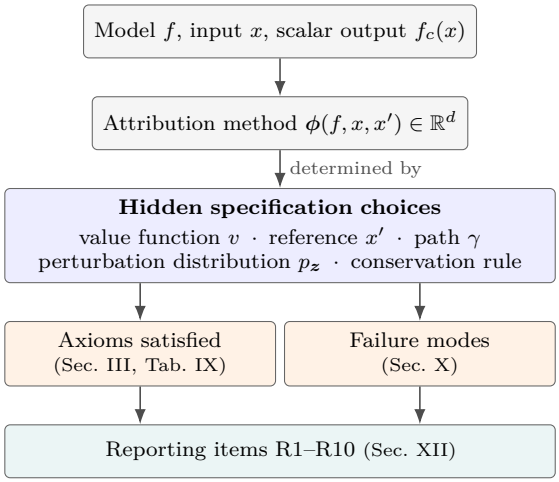


Fig. 1. Attribution methods as specifications of hidden mathematical choices. A local additive attribution is determined by five specification choices (value function, reference, path, perturbation distribution, conservation rule); these choices fix which axioms the method satisfies and which failure modes it is exposed to, and the proposed reporting checklist asks studies to state them.

This survey uses those mathematical choices as its organizing principle and compares feature-attribution methods through their underlying mathematical objects: the *value function* $v$ that defines "feature presence", the *baseline* or reference $x'$ against which contributions are measured, the *path* $\gamma$ along which integration occurs, the *perturbation distribution* that defines local linearity, and the *conservation rule* that distributes a quantity through a network. Many important disagreements and failure modes can be traced to differences in one or more of these objects. Figure 1 summarizes how this framing organizes the paper.

### A. Why Axioms?

A second organizing principle of this survey is the use of *axioms* as the primary tool of comparison. Axioms

TABLE I
FIVE MATHEMATICAL CHOICES USED AS THE ORGANIZING FRAME FOR THE SURVEY.

| Object | Question answered | Examples affected |
|---|---|---|
| Value function | What does feature absence mean? | SHAP, KernelSHAP, TreeSHAP |
| Reference | Compared to what? | IG, DeepLIFT, GradientSHAP |
| Path | Along what trajectory? | IG, Guided IG, Blur IG |
| Perturbation distribution | Which neighbourhood is local? | LIME, occlusion, RISE |
| Conservation rule | What quantity is propagated? | LRP, DeepLIFT |

are precise properties an attribution method may or may not satisfy: completeness (the attributions sum to the prediction-minus-baseline), implementation invariance (functionally equivalent models receive identical attributions), sensitivity (a feature that changes the prediction must receive nonzero attribution), symmetry, dummy, and others. We adopt the axiomatic perspective for three reasons:

1) **Axioms are model-agnostic.** A claim like "Integrated Gradients satisfies completeness" is a theorem about the method, not a hypothesis about a particular dataset or architecture. This makes axioms the right invariants for a survey: they remain true across benchmarks [1], [16].
2) **Axioms expose disagreement at its source.** When two methods produce different explanations, the question "which is correct?" is usually ill-posed; it amounts to asking which set of axioms one prefers [17], [18]. The function-approximation perspective of Han et al. [19] and the impossibility result of Bilodeau et al. [8] make the trade-off explicit: choosing a method is choosing which axioms to retain. Making the axioms explicit converts an unresolvable empirical dispute into a modeling decision.
3) **Axioms enable equivalence and reduction theorems.** Several apparently distinct methods can be shown to coincide under appropriate axiomatizations. KernelSHAP, exact Shapley, and certain forms of weighted linear regression compute the same quantity [2], [20]. DeepLIFT with the Rescale rule converges to Integrated Gradients in the limit of small input increments [21]. These equivalences are invisible from a heatmap-comparison perspective; they appear only through the axioms.

### B. Attribution vs. Interpretability, Explanation, and Causality

The terms *interpretability*, *explanation*, and *attribution* are often used interchangeably in the XAI literature, but the underlying objects are mathematically distinct. *Interpretability* is a property of a model, defined here as the extent to which a human can predict, audit, or modify its behaviour from its structure alone [22], [23]. *Explanation* is a broader category that includes example-based, counterfactual, concept-based, and rule-based justifications. *Attribution* is the narrow problem of decomposing a single prediction into per-feature contributions, typically as a real-valued vector. This survey concerns the third object: *local*, *additive*, *post-hoc* feature attribution for differentiable and tree-structured models.

We further distinguish attribution from *causal* feature analysis. A causal feature effect asks how the prediction would change under an external intervention on a feature, in the sense of Pearl's do-calculus. Most attribution methods, including all gradient-based methods, compute associational quantities relative to a chosen reference distribution. They become causal only under assumptions about the data-generating process that are rarely stated and even more rarely satisfied [24], [25]. Conflating the two is a major source of misinterpretation in high-stakes domains, and we return to it in Section X.

### C. Scope

**Included.** Local, post-hoc feature attribution for predictive models. This covers Shapley-value methods, path-integral methods, gradient and backpropagation methods, perturbation-based and surrogate methods, and CAM-style visual attribution insofar as it is derived from gradients. We treat attribution as a mathematical object and devote substantial space to axioms, value functions, baselines, paths, and conservation rules. We include the evaluation theory of faithfulness, infidelity, sanity checks, and ROAR-style retraining that has emerged in parallel with the methods themselves.

**Excluded or briefly treated.** Inherently interpretable models (linear models, decision lists, rule sets), global rule extraction, counterfactual explanations, example-based explanations, mechanistic interpretability, concept-bottleneck models, and visualization tools without attribution semantics. We comment on the boundary with concept methods (TCAV [26], network dissection [27]) where useful, but treat them as a complementary research programme outside the attribution methods surveyed here.

### D. Position Relative to Prior Surveys

Several broad XAI surveys preceded this one, including the foundational surveys of Guidotti et al. [28], Adadi and Berrada [29], Gilpin et al. [30], the book-length treatment of Molnar [31], the global-interpretation survey of Saleem et al. [32], the gradient-method technical review of Wang et al. [33], the additive feature-attribution review of Cremades et al. [34], the Shapley-specific survey of Li et al. [35], and the systematic evaluation review of Nauta et al. [36]. Table II compares this paper against those references along eight dimensions. The present survey differs from this literature in two respects:

1) Existing surveys have treated several of these families in depth, but the cross-family axiom structure remains fragmented. This article jointly cross-references Shapley,

path-based, gradient/backpropagation, and perturbation methods through a single *axiom-by-method matrix*. Existing surveys are either family-specific (Shapley [35], gradient [33], additive [34]) or breadth-prioritised [28]–[30].

2) The assumption sensitivity of attribution methods has not yet been consolidated into a reporting checklist for papers that use local additive attributions. This paper proposes such a checklist in Section XII and maps it to the analytic corpus in Appendix B.

*a) Comparison criteria for Table II.:* We assigned ✓ when the surveyed reference dedicates at least one section, subsection, or comparable structural unit to the column dimension; ∘ when the dimension is treated in fewer than two paragraphs, in an appendix, or only as part of a broader framing; and blank when the dimension is not addressed. The table is a scope comparison, not a bibliometric ranking. Ambiguous cases are scored in favour of the prior survey, and the interpretation of each mark is recorded in the supplementary scoring sheet.

### E. Survey Methodology

This paper is a structured narrative survey and taxonomy-driven review. The search aimed to assemble the methodological lineage needed to compare local attribution operators mathematically; estimating the size of the XAI literature or achieving exhaustive bibliometric coverage was out of scope. The analytic corpus contains 105 unique cited works spanning method-defining papers, axiomatic characterizations, evaluation studies, failure-mode analyses, boundary cases, and prior surveys. Papers were selected for their role in the taxonomy: defining an attribution operator, stating an axiom or reduction, evaluating attribution behaviour, documenting a failure mode, or clarifying the boundary between attribution and adjacent forms of explanation.

*a) Databases.:* Google Scholar, Semantic Scholar (S2-API), the ACL Anthology, and the proceedings indices of NeurIPS, ICML, ICLR, AAAI, IJCAI, ECCV, ICCV, CVPR, ACL, EMNLP, NAACL, and KDD.

*b) Time range.:* Searches were conducted between January and April 2025. The planned publication window ran from 1953 (Shapley's $n$-person-games paper) through the end of 2024. We also included the directly relevant 2025 axiomatic characterization of Integrated Gradients by Lundstrom and Razaviyayn [16], identified during final verification.

*c) Query strings.:* We ran the following queries verbatim on each database; the conjunctions were taken as Boolean AND, the disjunctions as Boolean OR.

- `("feature attribution" OR "Shapley value" OR "Integrated Gradients" OR LRP OR DeepLIFT OR Grad-CAM OR SHAP OR LIME) AND (axiom OR axiomatic OR completeness OR consistency)`
- `("Aumann-Shapley" OR Banzhaf OR "Owen value" OR "quantitative input influence" OR "game theory") AND ("feature attribution" OR "individual prediction" OR explanation)`
- `("attribution" OR "saliency") AND (faithfulness OR infidelity OR "sanity check" OR ROAR OR insertion-deletion)`
- `("attribution" OR "explanation") AND (adversarial OR fragile OR manipulation OR Lipschitz)`

*d) Selection process.:* Candidate records were deduplicated by title, DOI, and arXiv identifier where available. Papers were retained when they satisfied at least one of the following roles: (i) introduced a method included in the taxonomy, (ii) supplied an axiomatic characterization or reduction used in the comparison matrix, (iii) proposed an evaluation metric or benchmark used in Section IX, (iv) documented a failure mode used in Section X, or (v) provided a prior survey against which the present article is positioned. Historical sources were retained when they are needed for the mathematical genealogy of Shapley values, cooperative-game attribution, or early saliency methods.

*e) Boundary and exclusion criteria.:* The primary corpus is restricted to local additive attribution for predictive models. Global-interpretability methods, counterfactual and example-based explanations, concept-based explanations, and mechanistic interpretability are treated only when they clarify a boundary case for the taxonomy. Duplicate versions of the same work were represented by the most complete or most widely cited version.

*f) Corpus construction.:* The corpus is organized by analytic role. This choice is appropriate for a taxonomy whose unit of comparison is the attribution operator and its mathematical assumptions. Reproducibility is supported through the query strings, inclusion roles, boundary criteria, and the final analytic corpus underlying the tables. The corpus emphasizes English-language sources from major machine-learning, computer-vision, natural-language-processing, and AI venues, together with foundational mathematical sources required for the Shapley and axiomatic lineage.

### F. Contributions and Roadmap

The paper contributes three concrete artifacts: a taxonomy, an axiom matrix, and a reporting checklist.

1) **A unified mathematical taxonomy (Sections II–VII).** Section II fixes a common notation for $f$, $x$, $x'$, the coalition algebra over $N = \{1, \dots, d\}$, value functions $v$, masks $\boldsymbol{z}$, and paths $\gamma(\alpha)$. Section III catalogues the axioms under which different methods are characterized. Sections IV–VII survey the four method families in this common frame.
2) **An axiom-by-method matrix and failure-mode formalization (Sections VIII–X).** Section VIII presents the main comparison table of the paper, the axiom-by-method matrix (Table IX), together with the complexity landscape and several known reductions between methods (KernelSHAP and exact Shapley in expectation [2], [20]; DeepLIFT and Integrated

TABLE II
POSITION OF THIS SURVEY AGAINST PRIOR XAI/ATTRIBUTION SURVEYS. ✓= COVERED SUBSTANTIVELY; ○ = BRIEF OR PARTIAL; BLANK = NOT COVERED. "SHAP." SHAPLEY METHODS; "IG" PATH METHODS; "CAM" GRAD-CAM FAMILY; "AXIOM MATRIX" AN EXPLICIT AXIOM-BY-METHOD CROSS REFERENCE; "VF TAX." VALUE-FUNCTION TAXONOMY; "PATH/BASELINE TAX." BASELINE AND PATH TAXONOMY; "EVAL/FAILURE" FORMAL EVALUATION AND FAILURE-MODE TREATMENT; "CHECKLIST" REPORTING CHECKLIST AS A DELIVERABLE. THE FINAL ROW RECORDS THE INTENDED SCOPE OF THE PRESENT WORK, NOT AN INDEPENDENT QUALITY RANKING.

| Survey | Shap. | IG | CAM | Axiom matrix | VF tax. | Path/baseline tax. | Eval/Failure | Checklist |
|---|---|---|---|---|---|---|---|---|
| Guidotti et al. 2018 [28] | ○ | ○ | | | | | ○ | |
| Adadi & Berrada 2018 [29] | ○ | ○ | ○ | | | | ○ | |
| Gilpin et al. 2018 [30] | ○ | ○ | ○ | | | | | |
| Molnar 2022 [31] | ✓ | ✓ | ○ | | ○ | ○ | ○ | |
| Linardatos et al. 2021 [37] | ○ | ○ | ○ | | | | ○ | |
| Saleem et al. 2022 [32] | ○ | | | | | | ○ | |
| Nauta et al. 2023 [36] | ○ | ○ | ○ | | | | ✓ | |
| Li et al. 2024 [35] | ✓ | | | | ○ | | ○ | |
| Wang et al. 2024 [33] | | ○ | ○ | | | ○ | ○ | |
| Cremades et al. 2024 [34] | ✓ | ✓ | | | ○ | ○ | ○ | |
| This work | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

TABLE III
REVIEW CORPUS SUMMARY. ROLE COUNTS ARE NON-EXCLUSIVE BECAUSE ONE CITED PAPER MAY INTRODUCE A METHOD, STATE AN AXIOM, AND PROVIDE AN EVALUATION RESULT.

| Role in analytic corpus | Included papers |
|---|---|
| Method-introducing papers | 56 |
| Axiomatic characterization papers | 26 |
| Evaluation / benchmark papers | 20 |
| Failure-mode papers | 15 |
| Prior surveys | 10 |
| Boundary-case papers | 15 |
| Total unique cited papers in analytic corpus | 105 |

Gradients in the small-increment limit [21]; Grad-CAM and class-conditional gradient projections [4]; LRP-$\epsilon$ and gradient-×-input on bias-free ReLU networks [21]). Section X recasts the most-cited critiques of attribution methods [9], [10], [14], [15], [18] as *mathematical* problems tied to explicit modelling choices.

3) **A proposed reporting checklist (Section XII).** Feature-attribution methods encode modelling choices. Section XII expresses this point as a ten-item checklist for papers that report or rely on attribution results.

The remainder of the paper is organized as follows. Section II fixes notation. Section III states the axioms. Sections IV–VII survey the four method families. Section VIII presents the unifying comparison framework. Section IX reviews evaluation theory. Section X analyzes failure modes. Section XI surveys applications across model families. Section XII presents the proposed reporting checklist. Section XIII states the scope boundaries of the survey. Section XIV closes with open problems and a research agenda.

### G. Central Claim

The argument of the paper is summarized by the following claim:

> *There is no assumption-free feature-attribution method. Every local additive attribution method defines feature importance through choices about value functions, references, paths, perturbation distributions, or conservation rules. Trustworthy use therefore requires reporting a heatmap or ranking together with the assumptions under which the attribution was computed and interpreted.*

The checklist supports reproducible reporting of the assumptions required to compute and interpret an attribution. The remainder of the paper develops this claim formally and shows how common failure cases arise when these assumptions are left implicit.

## II. Problem Formulation and Notation

This section fixes a single notation used throughout the paper. Every method in Sections IV–VII is described in the symbols introduced here, making differences between methods visible as differences in specific mathematical objects across papers with otherwise distinct notation.

### A. Model and Input Space

We consider a predictive model

$$f\colon \mathcal{X} \to \mathbb{R},$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is the input space and $d$ is the number of input features. For multi-class problems, $f$ is the scalar logit or pre-softmax score associated with a target class $c$; when the dependence on $c$ matters, we write $f_c$. We make no structural assumption on $f$: it may be a deep network, a tree ensemble, or any other function from $\mathcal{X}$ to $\mathbb{R}$. Where differentiability is required, we will say so explicitly.

The data distribution on $\mathcal{X}$ is denoted $p_X$, with $X \sim p_X$ a random input. An individual input is written $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, with the $i$-th coordinate $x_i$.

### B. Baselines and References

Every additive attribution method requires, implicitly or explicitly, a *reference point* against which the prediction $f(x)$ is compared. We denote a single reference point as

$x' \in \mathbb{R}^d$ and a reference distribution as $\mathcal{D}_{x'}$. Common choices include

- a fixed point such as the all-zeros vector $x' = \mathbf{0}$, used in many gradient implementations;
- the population mean $x' = \mathbb{E}[X]$;
- a sample from the training distribution, $x' \sim p_X$, giving an *expected baseline* [38];
- a structured reference (e.g., a blurred image, a paraphrase of a sentence, an isoelectric protein sequence) appropriate to the input modality.

We use $\Delta f(x; x') \coloneqq f(x) - f(x')$ for the prediction-minus-baseline difference, which most additive methods target as the quantity to decompose.

### C. Coalitions and Value Functions

Let $N = \{1, 2, \ldots, d\}$ index the features. A *coalition* is a subset $S \subseteq N$ of features deemed "present"; its complement $\bar{S} = N \setminus S$ is the set of "absent" features. The number of features in $S$ is $|S|$, and $|N| = d$.

A *value function* is a set function

$$v \colon 2^N \to \mathbb{R}, \qquad S \mapsto v(S),$$

intended to represent "the prediction when only features in $S$ are observed". In Shapley-based attribution, the value function is often the most consequential modelling choice, and the ambiguity in defining it is the source of much downstream disagreement. Three canonical choices appear repeatedly:

**Marginal (interventional) value.**

Replace absent features with their reference values:

$$v^{\text{int}}_{x'}(S) = \mathbb{E}_{x' \sim \mathcal{D}_{x'}}\left[ f\big(x_S, x'_{\bar{S}}\big) \right], \tag{1}$$

where $(x_S, x'_{\bar{S}})$ denotes the vector with feature $i$ set to $x_i$ if $i \in S$ and to $x'_i$ otherwise. This corresponds to a hard intervention.

**Conditional value.**

Condition on observed features:

$$v^{\text{cond}}(S) = \mathbb{E}\big[ f(X) \,\big|\, X_S = x_S \big]. \tag{2}$$

This restricts to the data manifold but requires estimating $d$-dimensional conditional expectations [25], [39].

**Single-reference value.**

The deterministic limit of the interventional value with a fixed baseline:

$$v^{\text{ref}}_{x'}(S) = f\big(x_S, x'_{\bar{S}}\big). \tag{3}$$

The interventional and conditional value functions agree only when features are mutually independent under $p_X$. In all other cases they yield different Shapley values [17], [24], [40].

### D. Paths and Perturbations

Path-based methods integrate gradients along a continuous curve in input space. A *path* from a baseline $x'$ to an input $x$ is a differentiable map

$$\gamma \colon [0, 1] \to \mathbb{R}^d, \qquad \gamma(0) = x', \quad \gamma(1) = x.$$

The canonical choice is the straight line $\gamma(\alpha) = x' + \alpha(x - x')$, used by Integrated Gradients. Alternatives include adaptive paths [41], paths through image scale space [42], and paths defined by a generative model [43].

Perturbation-based methods do not integrate but instead evaluate $f$ at masked inputs. A *mask* is a vector $\boldsymbol{z} \in [0, 1]^d$ that interpolates between the input and a reference; we use $\boldsymbol{z} \odot x + (1 - \boldsymbol{z}) \odot x'$ to denote a masked input. When $\boldsymbol{z} \in \{0, 1\}^d$ the mask is binary, and we identify $\boldsymbol{z}$ with the coalition $S = \{i : \boldsymbol{z}_i = 1\}$.

### E. Attribution Vectors and Explanation Surrogates

An *attribution vector* for the prediction $f(x)$ relative to a baseline $x'$ is a vector

$$\boldsymbol{\phi}(f, x, x') \in \mathbb{R}^d$$

whose $i$-th entry $\phi_i$ is the signed contribution of feature $i$. Most methods we discuss are *additive*, meaning they aim to satisfy the *local accuracy* (or completeness, or efficiency) property:

$$\sum_{i=1}^{d} \phi_i \;=\; f(x) - f(x') \;=\; \Delta f(x; x'). \tag{4}$$

Equation (4) is the closest the field has to a universal target; we will return to it as Axiom 5 in Section III.

Lundberg and Lee [2] formalized the class of *additive feature attribution methods* as those that admit an explanation surrogate

$$g(\boldsymbol{z}) = \phi_0 + \sum_{i=1}^{d} \phi_i \, \boldsymbol{z}_i, \qquad \boldsymbol{z} \in \{0, 1\}^d, \tag{5}$$

where $\phi_0 = f(x')$ is the offset. Setting $\boldsymbol{z} = \mathbf{1}$ recovers (4). This linear-in-mask surrogate is the common language of LIME, SHAP, DeepLIFT, LRP, and Integrated Gradients; their differences lie in *how* the coefficients $\phi_i$ are computed, not in the form of $g$.

### F. A Running Example: Disagreement Without Error

A two-feature interaction already shows why attribution methods can disagree without either method being erroneous. Let

$$f(x_1, x_2) = x_1 x_2, \qquad x = (1, 1).$$

With the zero baseline $x'_0 = (0, 0)$, straight-line IG follows $\gamma(\alpha) = (\alpha, \alpha)$ and gives

$$\text{IG}_1(x; x'_0) = \text{IG}_2(x; x'_0) = \int_0^1 \alpha \, d\alpha = \frac{1}{2}.$$

The corresponding single-reference Shapley value with $v^{\text{ref}}_{x'_0}$ also assigns $(1/2, 1/2)$, because neither feature has value without the other.

Now change only the reference point to $x'_1 = (1, 0)$. The same model and same prediction give

$$\text{IG}(x; x'_1) = (0, 1),$$

TABLE IV
RUNNING EXAMPLE FOR $f(x_1, x_2) = x_1 x_2$ AT $x = (1, 1)$. DIFFERENT SPECIFICATIONS ANSWER DIFFERENT QUESTIONS.

| Specification | Attribution | Interpretation |
|---|---|---|
| IG, $x' = (0, 0)$ | $(1/2, 1/2)$ | interaction split evenly |
| IG, $x' = (1, 0)$ | $(0, 1)$ | credit for changing $x_2$ only |
| Single-reference Shapley, $x' = (0, 0)$ | $(1/2, 1/2)$ | coalition interaction split evenly |
| Off-manifold perturbation under $X_2 = X_1$ | specification-dependent | masked inputs leave $\mathrm{supp}(p_X)$ |

TABLE V
NOTATION USED THROUGHOUT THE SURVEY.

| Symbol | Meaning |
|---|---|
| $f$ | Predictive model $\mathcal{X} \to \mathbb{R}$ |
| $x, x_i$ | Input vector and $i$-th feature |
| $d$ | Number of features |
| $N$ | Feature index set $\{1, \dots, d\}$ |
| $x'$ | Baseline / reference input |
| $\mathcal{D}_{x'}$ | Reference distribution |
| $\Delta f(x; x')$ | $f(x) - f(x')$ |
| $S \subseteq N$ | Coalition of present features |
| $v(S)$ | Value function on coalitions |
| $\gamma(\alpha)$ | Path from baseline to input |
| $\boldsymbol{z}$ | Perturbation mask in $[0, 1]^d$ |
| $\boldsymbol{\phi} = (\phi_i)$ | Attribution vector in $\mathbb{R}^d$ |
| $g$ | Additive explanation surrogate |
| $p_X$ | Data distribution |
| $\nabla f$ | Gradient of model w.r.t. input |

because the path changes only the second coordinate and decomposes $f(1, 1) - f(1, 0) = 1$. Both answers are valid relative to their baselines; they answer different comparison questions. If, in addition, the data distribution has $X_2 = X_1$, then the coalition inputs $(1, 0)$ and $(0, 1)$ used by interventional Shapley or occlusion are off the data manifold. A conditional or manifold-aware value function therefore changes the interpretation again, not because the model changed but because the meaning of "feature absence" changed.

### G. Notation Summary

Table V summarizes the notation used throughout the rest of the paper.

## III. Axiomatic Foundations

This section catalogues the axioms used to characterize feature attribution methods. The axioms originate in two distinct traditions: the *cooperative-game* tradition derived from Shapley's 1953 paper [44], and the *path-method* tradition introduced by Sundararajan, Taly, and Yan in their 2017 paper on Integrated Gradients [1]. A central mathematical observation of this survey, made explicit in Section VIII, is that these two axiom systems are not independent: many path-method axioms have direct game-theoretic analogues, and some methods (notably DeepLIFT and SHAP variants) can be characterized in either language.

### A. Shapley Axioms

Let $v\colon 2^N \to \mathbb{R}$ be a value function with $v(\emptyset) = 0$. A *value* $\phi$ assigns to each player $i \in N$ a real number $\phi_i(v)$.

**Axiom 1** (Efficiency / Completeness)**.** $\sum_{i=1}^{d} \phi_i(v) = v(N)$.

**Axiom 2** (Symmetry)**.** If players $i, j$ are interchangeable in $v$, meaning $v(S \cup \{i\}) = v(S \cup \{j\})$ for every $S \subseteq N \setminus \{i, j\}$, then $\phi_i(v) = \phi_j(v)$.

**Axiom 3** (Dummy / Null Player)**.** If $v(S \cup \{i\}) = v(S)$ for every $S \subseteq N \setminus \{i\}$, then $\phi_i(v) = 0$.

**Axiom 4** (Additivity / Linearity)**.** For any two value functions $v_1, v_2$ and any $\alpha, \beta \in \mathbb{R}$, $\phi_i(\alpha\, v_1 + \beta\, v_2) = \alpha\, \phi_i(v_1) + \beta\, \phi_i(v_2)$.

**Result 1** (Shapley, 1953 [44])**.** *Assumptions:* a finite player set $N = \{1, \dots, d\}$; a real-valued set function $v\colon 2^N \to \mathbb{R}$ with $v(\emptyset) = 0$. Under these assumptions, there is a unique value $\phi$ on $N$ satisfying Axioms 1–4 for all such $v$, given by

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(d - |S| - 1)!}{d!} \big[v(S \cup \{i\}) - v(S)\big]. \quad (6)$$

The quantity in brackets, $v(S \cup \{i\}) - v(S)$, is the *marginal contribution* of feature $i$ to coalition $S$. The weight $|S|!\,(d - |S| - 1)!/d!$ is the probability that, under a uniformly random ordering of the $d$ players, exactly the members of $S$ precede $i$; the Shapley value is thus the expected marginal contribution of $i$ over random arrival orders.

The uniqueness half of Result 1 follows from a basis argument that is worth recording because it recurs in later characterizations. The *unanimity games* $u_T(S) = \mathbf{1}[T \subseteq S]$, for nonempty $T \subseteq N$, form a basis of the $(2^d - 1)$-dimensional space of value functions with $v(\emptyset) = 0$. On a unanimity game, the dummy axiom forces $\phi_i(u_T) = 0$ for $i \notin T$, and symmetry with efficiency forces $\phi_i(u_T) = 1/|T|$ for $i \in T$. Linearity then determines $\phi$ on every $v = \sum_T c_T u_T$, and evaluating the resulting expression coalition by coalition yields (6). The full argument is in the original paper [44].

### B. Additive Feature Attribution

Recall the additive surrogate of (5). Lundberg and Lee [2] showed that a single triple of axioms forces any method expressible as a linear-in-mask surrogate to coincide with Shapley values.

**Axiom 5** (Local Accuracy)**.** $g(\mathbf{1}) = f(x)$, i.e. $\sum_i \phi_i + \phi_0 = f(x)$.

**Axiom 6** (Missingness)**.** For any $i$ with $x_i = x'_i$, $\phi_i = 0$.

**Axiom 7** (Consistency [2], [45])**.** If $f'$ is a model such that the marginal contribution of feature $i$ in $f'$ is at least

its marginal contribution in $f$ for every coalition, then $\phi_i(f', x) \geq \phi_i(f, x)$.

**Result 2** (Lundberg and Lee, 2017 [2], Thm. 1)**.** *Assumptions:* the attribution takes the additive surrogate form (5) for a binary mask $\boldsymbol{z} \in \{0,1\}^d$; a value function $v_{f,x}$ over coalitions of features is fixed. Under these assumptions, the unique solution satisfying Axioms 5–7 coincides with the Shapley value $\phi_i = \phi_i^{\text{Shap}}(v_{f,x})$. The proof is Theorem 1 of [2].

This is the formal sense in which SHAP *unifies* the additive attribution family. The unification, however, leaves open the choice of $v_{f,x}$, and as we saw in Section II that choice is itself consequential.

### C. Alternative Cooperative-Game Values

The finite-player Shapley value is the dominant attribution index in XAI, but it is not the only cooperative-game value relevant to explanation. The Banzhaf value [46] averages marginal contributions under a different coalition weighting scheme, placing equal weight on coalitions, while the Shapley value averages player arrival orders. Owen's multilinear extension [47] connects finite games to polynomial extensions on the unit cube and is one route by which sampling and interaction calculations can be studied analytically. For continuous populations and cost-sharing problems, the Aumann-Shapley value [48] replaces finite coalitions with pathwise marginal rates. Specialized to a differentiable cost function $f$ on $\mathbb{R}^d$ with reference point $x'$, the Aumann-Shapley charge to coordinate $i$ is the diagonal-path integral

$$\phi_i^{\text{AS}}(x; x') = (x_i - x_i') \int_0^1 \frac{\partial f}{\partial x_i}\big(x' + \alpha(x - x')\big)\, d\alpha, \quad (7)$$

which is exactly the Integrated Gradients operator of Section V. This connection is substantive: it identifies IG as a fixed-path, feature-level cost-sharing rule, with the baseline and the straight-line path replacing the cooperative game's population model.

### D. Path-Method Axioms

The axiom set of Sundararajan et al. [1] targets attribution methods defined for differentiable models.

**Axiom 8** (Sensitivity-(a))**.** If the input $x$ and baseline $x'$ differ in exactly one feature and the predictions differ ($f(x) \neq f(x')$), then that feature must receive a nonzero attribution.

**Axiom 9** (Sensitivity-(b))**.** If $f$ does not mathematically depend on feature $i$, then $\phi_i = 0$. (This is the differentiable analogue of dummy.)

**Axiom 10** (Implementation Invariance)**.** If two networks $f_1$ and $f_2$ compute the same function ($f_1(x) = f_2(x)$ for all $x$), their attributions are identical.

**Axiom 11** (Completeness)**.** $\sum_i \phi_i = f(x) - f(x')$.

**Axiom 12** (Linearity)**.** $\phi$ is linear in $f$.

**Axiom 13** (Symmetry-Preserving)**.** If $x$ and $x'$ are symmetric in two features $i, j$ ($x_i = x_j$ and $x_i' = x_j'$) and $f$ is symmetric in those features, then $\phi_i = \phi_j$.

**Result 3** (Sundararajan et al., 2017 [1]; further axiomatic characterizations by Lundstrom and Razaviyayn, 2025 [16])**.** *Assumptions:* $f$ is differentiable along the straight-line path $\gamma(\alpha) = x' + \alpha(x - x')$; the gradient is integrable on $\alpha \in [0, 1]$. Under these assumptions and within the attribution classes considered in the cited papers, Integrated Gradients along the straight-line path is singled out by Axioms 8–10 together with completeness, linearity, and symmetry-preservation. Lundstrom and Razaviyayn [16] give three additional independent axiomatic characterizations under varied axiom subsets.

That several different axiom combinations identify the same straight-line operator strengthens the case for IG as a canonical member of the path family, but it does not make the operator assumption-free: each characterization fixes the attribution class and the path in advance.

### E. Conservation and Backpropagation Axioms

Layer-wise relevance propagation (LRP) [5] and DeepLIFT [6] are characterized not by a coalition or a path but by a *conservation rule* that distributes a quantity through the network's computation graph.

**Axiom 14** (Layer-wise Conservation [5])**.** For every layer $\ell$, the sum of relevances assigned to its inputs equals the sum of relevances received from its outputs:

$$\sum_{i \in \text{in}(\ell)} R_i^{(\ell)} = \sum_{j \in \text{out}(\ell)} R_j^{(\ell+1)}. \quad (8)$$

At the output layer, $R^{(L)} = f(x)$. At the input layer, $\sum_i R_i^{(0)} = f(x)$, so layer-wise conservation implies completeness.

**Axiom 15** (Summation-to-Delta [6])**.** DeepLIFT attributions $C_{\Delta x_i \Delta y}$ satisfy $\sum_i C_{\Delta x_i \Delta y} = \Delta y$, where $\Delta y = f(x) - f(x')$.

Although LRP and DeepLIFT are sometimes presented as alternatives to Integrated Gradients, all three satisfy a completeness axiom of the form $\sum_i \phi_i = \Delta y$. They differ in *how* the distribution is performed, not in *what* is preserved.

### F. Robustness and Stability Axioms

Beyond the classical axioms, several authors have proposed properties that capture an attribution method's robustness to small perturbations.

**Axiom 16** (Lipschitz Stability [49])**.** There exists $L < \infty$ such that $\|\boldsymbol{\phi}(x_1) - \boldsymbol{\phi}(x_2)\| \leq L \|x_1 - x_2\|$ for all $x_1, x_2$.

**Axiom 17** (Continuity)**.** $\boldsymbol{\phi}$ is continuous in $x$ (a strictly weaker condition than Axiom 16).

These properties are not satisfied by raw gradient saliency in practice [13], [14], and their failure is one of the principal motivations for SmoothGrad and similar averaging methods (Section VI).

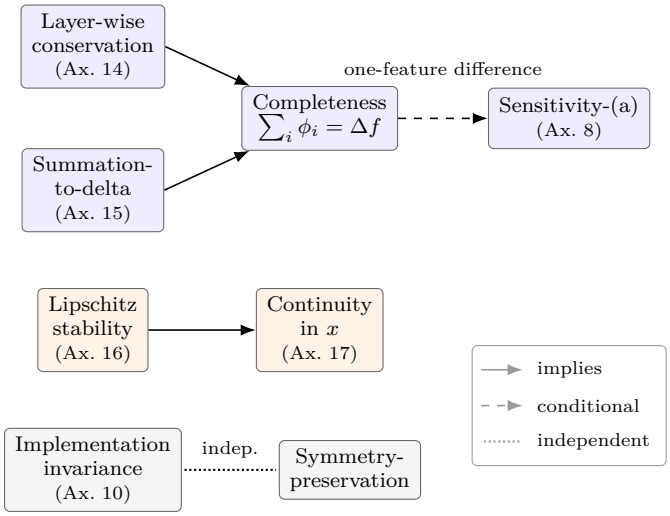


Fig. 2. Inter-axiom relationships. Layer-wise conservation (Axiom 14) and summation-to-delta (Axiom 15) each imply completeness. Completeness implies Sensitivity-(a) only in the Sensitivity-(a) setup, where $x$ and $x'$ differ in exactly one feature. Lipschitz stability is strictly stronger than continuity. Implementation invariance and symmetry-preservation are logically independent (Remark 1); the converses of all displayed implications fail in general.

### G. Inter-axiom Relationships

Many axioms imply or are implied by others. We collect the most useful relationships here for reference; they will recur in Section VIII.

*Remark* 1 (Inter-axiom implications). The following implications follow from the definitions of the axioms; we list them as observations.

(i) Completeness ⇒ Sensitivity-(a), under the Sensitivity-(a) setup: if $x$ and $x'$ differ in exactly one feature and $f(x) \neq f(x')$, the sum constraint forces that changed feature to receive nonzero attribution. Without the one-feature-difference condition, completeness implies only that at least one feature has nonzero attribution.

(ii) Layer-wise Conservation ⇒ Completeness: summing the per-layer conservation from output to input yields $\sum_i R_i^{(0)} = f(x)$.

(iii) Lipschitz Stability ⇒ Continuity in $x$ by definition.

(iv) Implementation Invariance and symmetry-preservation are logically independent; neither implies the other in general.

The converses generally fail: a method may be continuous without being Lipschitz, may satisfy completeness without being implementation invariant (e.g., a discrete-gradient method on a ReLU network), and may satisfy Sensitivity-(a) without satisfying completeness (e.g., raw saliency).

The path forward in the next four sections is to take each major method family and identify, with reference to this axiom catalogue, exactly which axioms it does and does not satisfy. The resulting matrix, collected in Section VIII, is a central comparison table of this survey.

## IV. Shapley and Cooperative-Game Attribution

The Shapley value [44], originally developed to distribute the joint payoff of a cooperative game among its players, is the most extensively studied attribution method. Its appeal is axiomatic: it is the unique solution to a small set of intuitive constraints (Result 1). Its difficulty is computational: exact evaluation has cost exponential in the number of features. The methods surveyed in this section can be understood as different points on the trade-off curve between fidelity to the axiomatic Shapley value and the cost of approximating it.

### A. Exact Shapley Values

For a value function $v$ on $2^N$, the Shapley value $\phi_i$ is given by (6). Equivalently,

$$\phi_i(v) = \frac{1}{d!} \sum_{\pi \in \Pi(N)} \big[ v(\mathrm{Pre}_i^\pi \cup \{i\}) - v(\mathrm{Pre}_i^\pi) \big], \qquad (9)$$

where $\Pi(N)$ is the set of permutations of $N$ and $\mathrm{Pre}_i^\pi$ is the set of features that precede $i$ in $\pi$. The two forms agree by a counting argument: a fixed coalition $S \subseteq N \setminus \{i\}$ arises as $\mathrm{Pre}_i^\pi$ for exactly $|S|!\,(d - |S| - 1)!$ of the $d!$ permutations (order the members of $S$, place $i$, order the rest), so grouping the sum in (9) by the value of $\mathrm{Pre}_i^\pi$ recovers the weights of (6). The form (9) is more convenient for sampling-based approximation: drawing permutations uniformly at random gives an unbiased Monte Carlo estimator whose variance decays at the standard $O(1/M)$ rate in the number of sampled permutations $M$.

The principal cost of exact Shapley is the $2^d$ evaluations of $v$, plus the work of evaluating $v$ itself, which under either the interventional or conditional definitions ((1)–(2)) requires in turn a model call or an expectation. For $d > 20$ exact evaluation is generally infeasible, and the methods below trade an approximation error for computational tractability.

### B. Pre-SHAP Game-Theoretic Feature Contributions

Modern SHAP terminology can obscure an older line of work that already treated local explanation as cooperative-game credit allocation. Štrumbelj and Kononenko [50] formulated individual classification explanations as feature-value contributions whose sum equals the change from an expected output to the model's prediction, and proposed a sampling approximation to avoid enumerating all feature subsets. This places black-box individual explanation in the Shapley lineage before the later SHAP unification.

Quantitative Input Influence (QII) [51] developed a related but distinct transparency framework: it asks how much an input or group of inputs influences an output under specified interventions or perturbation distributions. QII is important in this survey because it separates the *influence query* from the *estimator*. That separation anticipates the value-function distinction that now dominates Shapley explanations: a numerical attribution is meaningful only after the intervention, conditioning, or perturbation semantics have been fixed.

### C. KernelSHAP

KernelSHAP [2] reformulates Shapley value estimation as weighted linear regression. Sampling masks $\boldsymbol{z} \sim p_{\text{Shap}}(\boldsymbol{z})$ with the *Shapley kernel*

$$\pi_{\text{Shap}}(\boldsymbol{z}) = \frac{d-1}{\binom{d}{|\boldsymbol{z}|}\,|\boldsymbol{z}|(d-|\boldsymbol{z}|)}, \tag{10}$$

KernelSHAP fits the additive surrogate $g(\boldsymbol{z})$ of (5) by weighted least squares with weights $\pi_{\text{Shap}}$:

$$\hat{\boldsymbol{\phi}} = \underset{\phi_0,\ldots,\phi_d}{\arg\min} \sum_{\boldsymbol{z}\in\{0,1\}^d} \pi_{\text{Shap}}(\boldsymbol{z})\Big(v(\boldsymbol{z}) - \phi_0 - \sum_{i=1}^{d}\phi_i \boldsymbol{z}_i\Big)^2, \tag{11}$$

where $v(\boldsymbol{z})$ is the model evaluated on the masked input under the chosen value function. The kernel weight (10) is infinite at $|\boldsymbol{z}| \in \{0, d\}$; those two masks are handled as the hard constraints $\phi_0 = v(\mathbf{0})$ and $\sum_i \phi_i + \phi_0 = v(\mathbf{1})$, the latter being completeness. Lundberg and Lee proved that the solution of (11) over the full mask distribution coincides exactly with the Shapley values of (6). In practice, $M \ll 2^d$ masks are sampled, yielding an estimator with bias and variance that have since been characterized explicitly by Covert and Lee [20] and refined by the unbiased estimator of [52].

KernelSHAP is *model-agnostic*: it requires only black-box access to $f$, making it applicable to gradient-free models. Its correctness, however, depends on the value function used in the regression's prediction targets. By default KernelSHAP marginalizes over a fixed reference (the interventional value function), and as a result it inherits all of the assumptions about feature independence discussed in Section II.

### D. TreeSHAP

For tree ensembles, the structural recursion of the model can be exploited to compute Shapley values exactly in polynomial time. Lundberg et al. [45] introduced TreeSHAP, which runs in $O(TLD^2)$ time on a tree ensemble with $T$ trees, $L$ leaves, and depth $D$. The algorithm maintains, at each node, a polynomial in "coalition mass" that encodes the contribution of each subtree to all possible coalitions; this polynomial is propagated through the tree and combined at the leaves.

TreeSHAP supports two value functions: *path-dependent* (the empirical conditional expectation along training-time tree splits) and *interventional* (the marginal value over a reference dataset). The two yield different attributions whenever features are correlated, a discrepancy now widely recognized in the practical SHAP literature [17], [24]. TreeSHAP also satisfies the consistency axiom (Axiom 7), a property that fails for naive gain-based feature importance.

### E. DeepSHAP and GradientSHAP

DeepSHAP [2] extends DeepLIFT-Rescale rules to approximate Shapley values for deep networks. The key insight is that DeepLIFT's per-neuron multipliers can be interpreted as expectations over a baseline distribution, and aggregated by linear combination to yield a Shapley-style attribution. The approximation is exact for linear models and for the composition of linear and ReLU layers with a single fixed baseline; for more complex architectures it is an empirical heuristic.

GradientSHAP averages Integrated Gradients computations over a distribution of baselines drawn from the data [2], [38]. Specifically,

$$\phi_i^{\text{GradSHAP}} = \mathbb{E}_{x'\sim p_X,\,\alpha\sim U[0,1]}\Bigg[(x_i - x_i') \cdot \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i}\Bigg]. \tag{12}$$

The expectation over $x'$ is the link to Shapley reasoning, as discussed in detail by Erion et al. [38]; the expectation over $\alpha$ is the standard Integrated Gradients path integral. The combination yields a sampling-based estimator of the expected Shapley value with respect to the data distribution.

### F. SAGE (boundary case): Global Shapley Effects

**Scope note.** SAGE produces a *global* feature-importance measure, one level above the local-attribution scope of this survey. We include it because it shares the Shapley axiomatic foundation with KernelSHAP and TreeSHAP and because its loss-based value function illustrates how the same axiomatic machinery extends beyond per-prediction attribution. We mark SAGE as a boundary case in the axiom matrix (Table IX) and treat its remaining global-explanation siblings as out of scope (Section XIII).

The methods above attribute a single prediction. SAGE (Shapley Additive Global ExplanationS) [53] extends the framework to a global feature-importance measure by replacing the per-input value function with a loss-based one:

$$v^{\text{SAGE}}(S) = -\mathbb{E}_{(X,Y)}\big[\ell(f(X_S, X'_{\bar{S}}), Y)\big] + \text{const.} \tag{13}$$

Here $X'_{\bar{S}}$ marginalizes over the missing features. Shapley values of $v^{\text{SAGE}}$ measure how much of the model's predictive performance is attributable to each feature globally. SAGE inherits the same independence concerns as KernelSHAP but resolves them at the dataset level; KernelSHAP resolves them per prediction. It is tempting to treat any global Shapley score as a feature-selection criterion, but this is a separate modelling decision. Fryer, Strümke, and Nguyen [54] show through counterexamples that the classical Shapley axioms do not by themselves guarantee suitability for subset selection; the game formulation must match the statistical objective of the selected feature set.

### G. Shapley-Taylor Interaction Indices

Standard Shapley values attribute the prediction to single features. In many applications, including protein–protein interactions, drug combinations, and NLP feature interactions, joint effects are central. Dhamdhere et al. [55]

introduced the Shapley-Taylor interaction index, generalizing Shapley to subsets:

$$\phi_T^{\mathrm{ST},k}(v) = \sum_{S \subseteq N \setminus T} w^k_{|S|,d}\, \Delta_T v(S), \tag{14}$$

where $T \subseteq N$ with $|T| \le k$, $\Delta_T$ is the $|T|$-th discrete derivative, and $w^k$ is a generalization of the Shapley weight. The index is uniquely characterized by an axiom set that extends the classical Shapley axioms (Theorems 3 and 4 of [55]), and reduces to Shapley values when $k = 1$ and $|T| = 1$.

Janizek, Sturmfels, and Lee [56] give a continuous analogue via second-order path integrals (*Integrated Hessians*), which we cover in Section V; the discrete and continuous interaction frames are related by Theorem 5 of [56]. Interaction discovery can also be approached directly from learned model structure. Tsang et al. [57] detect statistical interactions from neural-network weights, making the interaction object explicit without deriving it solely from local feature perturbations. At the model-element level, Neuron Shapley [58] applies Shapley valuation to neurons or filters as the players. Both works are boundary cases for this survey's local input-attribution scope, but they sharpen an important point: the "players" in a cooperative explanation game need not be raw input coordinates.

### H. Interventional vs. Conditional Value Functions

In Shapley-based attribution, the value function is often the most consequential modelling choice. The choice between $v^{\mathrm{int}}$ ((1)) and $v^{\mathrm{cond}}$ ((2)) determines the answers to questions that look identical at the level of an explanation but differ at the level of inference.

**Interventional Shapley** answers the question: "if I forced feature $i$ to take its baseline value, how would the prediction change?" This is the natural quantity for debugging, mechanism discovery, and causal reasoning under the implicit assumption that the features can be intervened on independently.

**Conditional Shapley** answers: "conditional on observing $x_i$, how does the expected prediction change?" This is the natural quantity for predictive importance under the observed data distribution.

The two coincide only when features are mutually independent. In the presence of correlated features, which is the usual case, they differ, and the gap can be large. Janzing, Minorics, and Blöbaum [24] argued that the conditional value function conflates association with causation and recommends the interventional value for almost all use cases. Aas, Jullum, and Løland [39] took the opposite position for risk-management applications, arguing that an interventional value function evaluates the model at points the data manifold never visits, producing implausibly large attributions for highly correlated features. Frye et al. [25] proposed manifold-aware Shapley values that compute the interventional quantity but only over on-manifold coalition completions.

Sundararajan and Najmi [17] catalogued no fewer than four distinct "Shapley values for model explanation": the conditional expectation, conditional expectation w.r.t. the model, the baseline expectation, and the random baseline expectation. They showed that they disagree on simple, low-dimensional examples. Their recommendation, which we endorse, is that any paper presenting Shapley attributions should explicitly name which of these is being computed.

### I. Estimation: Variance, Bias, and Recent Algorithms

Sampling-based Shapley estimators introduce both bias and variance. Chen et al. [52] surveyed the estimator landscape, distinguishing

1) *Permutation sampling*, which gives an unbiased Monte Carlo estimator of (9);
2) *KernelSHAP-style* weighted regression, which is more sample-efficient on well-conditioned value functions;
3) *Unbiased KernelSHAP* [20], which corrects the bias introduced by paired sampling.

For tree models, TreeSHAP is exact; for differentiable models, GradientSHAP is approximate but inexpensive; for arbitrary models, KernelSHAP is the default. The choice is again a value-function question: each estimator is optimal for a slightly different definition of the underlying Shapley value, as Figure 3 of Chen et al. makes explicit.

### J. Summary

Table VI summarizes the Shapley-family methods discussed above. The table makes the central point of this section: the methods agree on what they aim to compute, namely a Shapley decomposition of the prediction, but differ in their value function, their approximation algorithm, and the assumptions under which they are exact.

## V. Path-Based Attribution

Path-based methods compute attributions by integrating the gradient of $f$ along a continuous path in input space from a baseline $x'$ to the target input $x$. They are the continuous analogue of cooperative-game methods: where Shapley values average marginal contributions over discrete coalitions, path methods average gradients over a continuous trajectory. The canonical instance, Integrated Gradients, was introduced by Sundararajan, Taly, and Yan [1] and remains the most widely used member of the family.

### A. Integrated Gradients

Let $f$ be differentiable. Integrated Gradients (IG) attributes to feature $i$ the quantity

$$\mathrm{IG}_i(x; x') = (x_i - x'_i) \int_0^1 \frac{\partial f\big(x' + \alpha(x - x')\big)}{\partial x_i}\, d\alpha. \tag{15}$$

The integral is taken along the straight-line path $\gamma(\alpha) = x' + \alpha(x - x')$, evaluated by Riemann sums in practice (typically $K \in [20, 300]$ steps). Equation (15) satisfies completeness exactly:

$$\sum_{i=1}^{d} \mathrm{IG}_i(x; x') = f(x) - f(x'). \tag{16}$$

TABLE VI
SHAPLEY-VALUE VARIANTS FOR FEATURE ATTRIBUTION. THE "VALUE FUNCTION" COLUMN REFERS TO THE DEFINITIONS IN (1)–(3). "EXACT" MEANS THE METHOD RECOVERS THE SHAPLEY VALUE WITHOUT SAMPLING, UNDER THE STATED VALUE FUNCTION.

| Method | Value function | Estimator | Complexity | Feature dependence |
|---|---|---|---|---|
| Exact Shapley | any | exhaustive | $O(2^d)$ | handled by choice of $v$ |
| KernelSHAP | interventional (default) | weighted regression | $O(Md)$ per sample | ignored |
| Unbiased KernelSHAP [20] | interventional | paired sampling | $O(Md)$ | ignored |
| TreeSHAP [45] | path-dep. or interventional | exact recursion | $O(TLD^2)$ | path-dependent option |
| DeepSHAP [2] | multiplier-based | DeepLIFT propagation | $O(\text{net})$ | approximation |
| GradientSHAP [38] | expected baseline | path + baseline avg. | $O(KM)$ | via baseline distribution |
| SAGE [53] | loss-based, global | permutation sampling | $O(Md^2)$ | via marginalization |
| Shapley-Taylor [55] | generalizes interventional | enumeration up to order $k$ | $O\left(\binom{d}{k}\right)$ | ignored at order $k$ |

Writing $g(\alpha) = f(\gamma(\alpha))$, the chain rule gives $g'(\alpha) = \sum_i \partial_i f(\gamma(\alpha))\,(x_i - x'_i)$ because $\gamma'_i(\alpha) = x_i - x'_i$ on the straight line, and so

$$f(x) - f(x') = g(1) - g(0) = \int_0^1 g'(\alpha)\,d\alpha = \sum_{i=1}^{d} \mathrm{IG}_i(x; x'), \tag{17}$$

the middle equality being the fundamental theorem of calculus and the last exchanging the finite sum with the integral.

**Result 4** (Sundararajan et al. [1])**.** *Assumptions:* $f$ is differentiable on the straight-line path between $x'$ and $x$; the gradient $\partial f/\partial x_i$ is integrable on $[0, 1]$. Under these assumptions, Integrated Gradients along the straight-line path satisfies sensitivity-(a), sensitivity-(b), implementation invariance, completeness, linearity, and symmetry-preservation (Axioms 8–10 plus completeness and symmetry). Failures of differentiability (e.g., ReLU kink points) are handled by the standard interpretation of partial derivatives as subgradients almost everywhere along the path.

Implementation invariance follows from the fact that $\partial f/\partial x_i$ depends only on the input-output behaviour of $f$, not on the network's parametric form. Completeness follows from the chain rule. Sensitivity-(a) follows because the integrand is nonzero whenever $f(x) \neq f(x')$ along $\gamma$. These properties are robust under any monotone reparameterization of the path, but not under arbitrary changes to $\gamma$: implementation invariance, in particular, fails for paths that depend on the model's parametrization, an observation made precise by Lundstrom and Razaviyayn [16].

### B. Baselines and Path Sensitivity

The IG attribution depends on the baseline $x'$. This dependence is mathematically necessary to define $\Delta f$ and is substantively important. The choice of $x'$ encodes what counts as the "absence" of a feature.

Sturmfels, Lundberg, and Lee [59] catalogued the most common baselines and their failure modes:

- The *zero baseline* $x' = \mathbf{0}$ is the default in many implementations but is often pathological. In image models, black pixels are themselves features, and Sensitivity-(a) is then satisfied in a misleading way: every dark pixel in $x$ receives an attribution of magnitude $|x_i \cdot \partial f/\partial x_i|$ that conflates "important to the model" with "not equal to zero".
- The *mean baseline* $x' = \mathbb{E}[X]$ has the merit of being on or near the data manifold but can still be highly atypical. For ImageNet, the mean is a uniform grey image.
- The *expected baseline*, due to Erion et al. [38], replaces a fixed $x'$ with an expectation over the training distribution:

$$\mathrm{EG}_i(x) = \mathbb{E}_{x' \sim p_X,\, \alpha \sim U[0,1]}\left[(x_i - x'_i) \cdot \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i}\right]. \tag{18}$$

  This was already encountered as (12); in the path-integral framing it is the natural extension of IG to a distribution of baselines.
- *Adversarial* or *informative* baselines are chosen so that $f(x')$ is itself meaningful (e.g., a baseline of the target class for a counterfactual attribution).

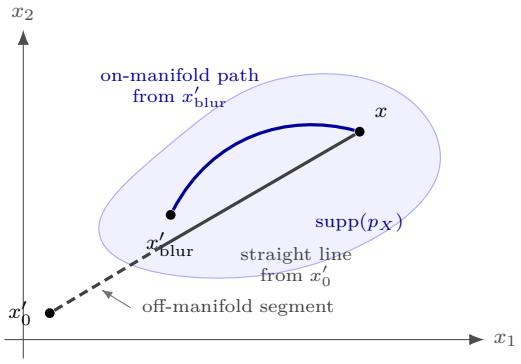


Fig. 3. Baseline and path dependence for path-based attribution in a two-feature input space. The shaded region is the support of the data distribution. The straight-line path from a zero baseline $x'_0$ spends its early segment (dashed) outside supp($p_X$), where $f$ extrapolates and gradients are unconstrained by training data. A path from an in-distribution baseline $x'_{\text{blur}}$ that stays inside the support queries $f$ only where it was trained. Both decompose the same total, $\sum_i \phi_i = f(x) - f(x')$ for their respective $x'$, but the per-feature split depends on the baseline and the path.

### C. Adaptive and Region-Based Paths

The straight-line path is mathematically convenient but may be poorly matched to the data manifold. It traverses regions of $\mathcal{X}$ where the model may behave erratically, a well-known manifestation of gradient saturation in deep networks means $\partial f/\partial x_i \approx 0$ along long stretches of the path [6], [21], washing out attributions for important features.

**Guided Integrated Gradients** (Guided IG) [41] replaces the straight-line path with an adaptive trajectory that steers around low-gradient regions. At each step, the path follows the direction of steepest absolute partial derivative, restricted to features that have not yet been "saturated". Guided IG retains completeness (it is still a path integral) but loses the symmetry-preservation property: a path that depends on the gradient landscape is no longer permutation-equivariant.

**XRAI** [60] attributes credit to *image regions*, with individual pixels grouped by a segmentation procedure; each region is then assigned the sum of IG attributions of its constituent pixels, divided by region area to give an importance score. XRAI is best understood as an aggregation layer over IG; it does not change the underlying attribution but improves the human readability of image-domain explanations.

**Blur Integrated Gradients** [42] replaces the path in input space with a path in *scale space*: $\gamma(\alpha)$ is a sequence of progressively less blurred versions of $x$, with the heavily blurred input as the baseline. This avoids the pathologies of zero-baseline IG for images while keeping the path integral interpretable as a decomposition of $\Delta f$. The trade-off is that the resulting attribution is no longer additive in the original pixel coordinates; it lives in scale space.

### D. Higher-Order Path Methods: Integrated Hessians

To attribute joint effects of feature pairs, Janizek, Sturmfels, and Lee [56] introduced Integrated Hessians (IH). Where IG integrates $\partial f/\partial x_i$ along a single path from $x'$ to $x$, IH integrates the second mixed partial $\partial^2 f/\partial x_i \partial x_j$ along a *nested* pair of paths and yields an interaction attribution $\Gamma_{ij}$ satisfying

$$\sum_{i,j} \Gamma_{ij}(x; x') = f(x) - f(x'). \tag{19}$$

**Result 5** (Janizek et al. [56], Thm. 1)**.** *Assumptions:* $f$ is twice-differentiable along the nested path; the mixed partials $\partial^2 f/\partial x_i \partial x_j$ exist and are integrable. Under these assumptions, Integrated Hessians is the unique attribution of second-order feature interactions consistent with the IG axioms when applied to the gradient-of-IG operator, as stated in the cited paper.

IH connects directly to the Shapley-Taylor interaction index [55] (Section IV): the discrete Shapley-Taylor index of order 2 is a coalition-sampling approximation of the IH continuous integral.

### E. Conductance: Internal Path Attribution

The path integral (15) attributes credit to *input* features. Conductance [61] extends the same construction to internal neurons:

$$\mathrm{Cond}_h(x; x') = (x - x')^\top \int_0^1 \frac{\partial f}{\partial h}\bigg|_{\gamma(\alpha)} \frac{\partial h}{\partial x}\bigg|_{\gamma(\alpha)} d\alpha, \tag{20}$$

where $h$ is the activation of a chosen hidden neuron. By construction, conductances of all neurons in a layer sum to $\Delta f$, giving a layer-wise decomposition consistent with the LRP conservation axiom (Axiom 14) without imposing layer-specific rules.

### F. Path Dependence: When Does the Path Matter?

A central theoretical question is: how sensitive are IG and its variants to the choice of path? Every method in this section is an instance of the *generalized path attribution*

$$\mathrm{IG}_i^\gamma(x; x') = \int_0^1 \frac{\partial f}{\partial x_i}\big(\gamma(\alpha)\big)\, \gamma_i'(\alpha)\, d\alpha, \tag{21}$$

for some differentiable $\gamma$ with $\gamma(0) = x'$ and $\gamma(1) = x$; the straight line recovers (15) and the Aumann-Shapley value (7). Summing (21) over $i$ and applying the chain rule as in (17) shows that every path-integral method of this form satisfies completeness, provided $f$ is differentiable along $\gamma$ with integrable gradient and the path has the stated endpoints: the total $\int_0^1 \nabla f(\gamma(\alpha))^\top \gamma'(\alpha)\, d\alpha = f(x) - f(x')$ depends only on the endpoints. The *per-feature* terms of (21), however, are not endpoint-determined: the decomposition that IG, EG, Guided IG, and Blur IG return is path-dependent.

Two paths between the same $x'$ and $x$ that disagree on the per-feature decomposition correspond to two different ways of attributing the same total credit. This is not necessarily an error; it reflects the fact that attribution is underdetermined without additional assumptions. Under the assumptions and attribution class used by Sundararajan et al. [1], the straight-line path is singled out by symmetry-preservation and related axioms, but other choices are defensible under different axiom sets [16].

### G. Summary

Table VII summarizes the path-based methods. The common thread is that each method makes one of two changes to IG (15): either it changes the baseline (EG, GradientSHAP), or it changes the path (Guided IG, Blur IG, XRAI), or both (IH adds an additional integration dimension). The choice of path or baseline determines which axioms hold, which features are highlighted, and how robust the attribution is to gradient saturation.

## VI. Gradient and Backpropagation Attribution

Gradient and backpropagation methods derive an attribution from a single backward pass through the model. They are the cheapest family of attribution methods, typically one or two backward passes, and historically were

TABLE VII
PATH-BASED ATTRIBUTION METHODS. ALL PRESERVE COMPLETENESS; THEY DIFFER IN THE CHOICE OF PATH AND BASELINE, WITH CORRESPONDING EFFECTS ON THE REMAINING AXIOMS.

| Method | Path | Baseline | Completeness | Interactions | Suited to |
|---|---|---|---|---|---|
| Integrated Gradients [1] | straight line | fixed $x'$ | yes | 1st order | tabular, image |
| Expected Gradients [38] | straight line | $\mathbb{E}_{x' \sim p_X}$ | yes | 1st order | tabular, image |
| Guided IG [41] | adaptive | fixed $x'$ | yes | 1st order | image (noisy IG) |
| Blur IG [42] | scale space | blurred input | yes | 1st order | image |
| XRAI [60] | inherits IG path | inherits IG baseline | yes (per region) | 1st order | image (segmented) |
| Integrated Hessians [56] | nested paths | fixed $x'$ | yes (sum over pairs) | 2nd order | feature interactions |
| Conductance [61] | straight line in $x$ | fixed $x'$ | yes (sum over neurons) | 1st order | internal neurons |

the first to be applied to deep networks. They are also the family with the most diverse axiomatic profiles: some satisfy completeness, some do not; some are implementation-invariant, some are not; some are continuous in $x$, and a notable subset is so discontinuous as to be operationally unstable [13], [14].

### A. Raw Saliency and Gradient × Input

The earliest gradient-based attribution is the *saliency map* [62], [63]:

$$\phi_i^{\text{sal}}(x) = \left| \frac{\partial f(x)}{\partial x_i} \right| \quad \text{or} \quad \phi_i^{\text{sal}}(x) = \frac{\partial f(x)}{\partial x_i}. \tag{22}$$

Saliency is the first-order Taylor coefficient of $f$ at $x$. It satisfies sensitivity-(b) and implementation invariance but *not* completeness or sensitivity-(a): a feature can change the prediction substantially yet receive a gradient near zero whenever $f$ is locally flat (gradient saturation [6]).

The *gradient* × *input* variant,

$$\phi_i^{\text{gxi}}(x) = x_i \cdot \frac{\partial f(x)}{\partial x_i}, \tag{23}$$

is the first-order term of the Taylor expansion of $f$ around the origin. It is one of the cheapest sensible attributions, and is the $K = 1$ approximation of Integrated Gradients with $x' = \mathbf{0}$. It satisfies completeness only when $f$ is linear.

### B. SmoothGrad and Noise Averaging

SmoothGrad [64] smooths the saliency map by averaging over noisy versions of the input:

$$\phi_i^{\text{SG}}(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \left[ \left| \frac{\partial f(x+\epsilon)}{\partial x_i} \right| \right]. \tag{24}$$

The averaging effectively replaces $f$ with the Gaussian-smoothed model $\tilde{f}_\sigma = f * \mathcal{N}(0, \sigma^2 I)$, i.e. $\tilde{f}_\sigma(x) = \mathbb{E}_\epsilon[f(x+\epsilon)]$. For $f$ locally integrable with gradient of at most polynomial growth, differentiation and expectation interchange, so

$$\nabla \tilde{f}_\sigma(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \big[ \nabla f(x+\epsilon) \big] : \tag{25}$$

the *signed* SmoothGrad average is an exact gradient of a smoothed model, and inherits whatever axioms hold for gradients of $\tilde{f}_\sigma$. The absolute-value form in (24) breaks this identity because $\mathbb{E}[|\cdot|] \neq |\mathbb{E}[\cdot]|$, so absolute SmoothGrad is not the gradient of any smoothed model. SmoothGrad substantially reduces the high-frequency artefacts visible in raw saliency maps for image classifiers, at the cost of a hyperparameter $\sigma$ that controls the trade-off between locality and stability.

**VarGrad** [9] computes the variance rather than the expectation: $\phi_i^{\text{VG}}(x) = \text{Var}_\epsilon[\partial f(x+\epsilon)/\partial x_i]$. VarGrad highlights features for which the gradient is locally *unstable*, often providing complementary information to SmoothGrad.

Both methods can be viewed as Monte Carlo estimators of the gradient of a noise-smoothed model, and as such inherit the axiomatic profile of that smoothed model. They satisfy implementation invariance but typically not completeness, since the smoothed gradient does not integrate to $f(x)$ in general.

### C. Guided Backpropagation and Deconvolution

Guided backpropagation [65] and deconvolution networks [66] modify the backward pass through ReLU units. Standard backpropagation uses positive forward activations; these methods additionally suppress gradients of opposite sign.

**Result 6** (Empirical finding of Adebayo et al. [9])**.** On standard image classifiers, guided backpropagation produces attribution maps that are visually largely independent of the model's parameters: replacing the weights with random values yields visually similar attributions. We report this as an empirical result from the cited paper, not as a formal theorem.

Result 6 (the model-randomization sanity check) is one of the strongest empirical critiques of guided backpropagation: an attribution that is largely insensitive to model parameters cannot, by itself, support strong claims about model-specific reasoning. We return to the sanity-check framework in Section IX.

### D. DeepLIFT

DeepLIFT [6] replaces gradients with *multipliers* computed against a reference activation. For each neuron $h$ with reference value $h_{\text{ref}}$ and difference $\Delta h = h - h_{\text{ref}}$, the DeepLIFT contribution $C_{\Delta x \Delta h}$ propagates through the network analogously to the chain rule:

$$C_{\Delta x_i \Delta y} = \sum_h m_{x_i h} \cdot C_{\Delta h \Delta y}, \tag{26}$$

where $m_{x_i h} = C_{\Delta x_i \Delta h} / \Delta x_i$ is a multiplier. DeepLIFT satisfies the summation-to-delta axiom (Axiom 15) by construction. Two distinct rules, *Rescale* (for monotone nonlinearities) and *RevealCancel* (for pairs of opposing effects), determine the multiplier at each layer.

**Result 7** (Ancona et al. [21], §3)**.** *Assumptions:* the network is composed only of linear layers and monotone elementwise nonlinearities (e.g., ReLU, sigmoid, tanh); there are no skip connections, attention layers, concatenations, or non-monotone operators; a single fixed baseline $x'$ is used. Under these assumptions, DeepLIFT-Rescale and Integrated Gradients coincide in the small-input-increment limit. When the assumptions fail (Add, GroupNorm, attention, concatenation), the equivalence breaks and the methods can yield different attributions.

Result 7 establishes a key bridge: under restricted architectures and propagation rules, DeepLIFT-Rescale and Integrated Gradients can be viewed as closely related approximations of the same path-based quantity. Outside those assumptions, they should be treated as distinct attribution operators. They differ in computational profile (one forward+backward pass for DeepLIFT vs. $K$ for IG) and in their handling of non-monotone nonlinearities (the RevealCancel rule addresses a case IG does not).

### E. Layer-Wise Relevance Propagation

LRP [5] propagates a quantity called *relevance* backward through the network, subject to the conservation axiom (Axiom 14). The relevance of an output neuron is initialized to $f(x)$; at each layer, it is decomposed into contributions from the preceding layer using a propagation rule. The most common rules are:

**$z$-rule**
: $R_i = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j$, where $z_{ij} = x_i w_{ij}$ is the pre-activation contribution. Numerically unstable near zero.

**$\epsilon$-rule**
: adds a small constant $\epsilon$ to the denominator to suppress numerically unstable terms.

**$\gamma$-rule**
: amplifies positive contributions with multiplier $\gamma$, used for the upper layers of deep classifiers.

The *LRP composition* strategy of Montavon et al. [67] applies different rules to different layers, producing attribution maps that are empirically more faithful to the model's behaviour than any single rule alone.

LRP is implementation-invariant only for the standard rules applied to specific architecture families; bespoke rules can break the property. It satisfies conservation by construction, hence completeness. It does not in general satisfy Sensitivity-(b): an input feature that is formally a dummy can still receive nonzero relevance under certain rules.

### F. FullGrad

FullGrad [68] extends gradient-based attribution to incorporate *bias* terms, which earlier methods ignored. The key observation is that for a network with biases, the gradient $\nabla_x f$ alone does not satisfy completeness; the bias contributions must be added explicitly. FullGrad computes

$$\phi_i^{\text{FG}}(x) = x_i \frac{\partial f}{\partial x_i} + \sum_\ell f_\ell^{\text{bias}}(x), \tag{27}$$

where $f_\ell^{\text{bias}}$ is the contribution from biases at layer $\ell$. FullGrad satisfies completeness exactly: $\sum_i \phi_i^{\text{FG}} = f(x)$. This target includes the bias contribution at $x' = \mathbf{0}$ and therefore differs from $\Delta f$.

### G. The CAM Family: From Class Activation to Shapley-CAM

Class Activation Mapping (CAM) [69] was originally developed for global-average-pooled CNNs as a coarse spatial map of which regions activate a given class. Grad-CAM [4] generalized CAM to arbitrary CNN architectures by using gradients of the class score with respect to the final convolutional feature map:

$$L_{c,(i,j)}^{\text{GC}} = \text{ReLU}\left( \sum_k \alpha_k^c \, A_{(i,j)}^k \right), \quad \alpha_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial f_c}{\partial A_{(i,j)}^k}. \tag{28}$$

Here $A^k$ is the $k$-th feature map of the target convolutional layer and $Z$ is the number of spatial positions. The weights $\alpha_k^c$ average the class-score gradient over space, and the final map is a ReLU-rectified sum of the feature maps.

**Grad-CAM++** [70] replaces the average pooling with a weighted average that emphasizes positive contributions:

$$\begin{aligned} \alpha_k^{c,++} &= \sum_{i,j} \frac{g_{kij}^{(2)}}{2 g_{kij}^{(2)} + \sum_{a,b} A_{(a,b)}^k g_{kij}^{(3)}}, \\ g_{kij}^{(m)} &= \frac{\partial^m f_c}{\partial (A_{(i,j)}^k)^m}. \end{aligned} \tag{29}$$

The motivation is to handle multiple instances of the target class.

**Score-CAM** [71] eliminates gradients entirely by computing weights through perturbation: each feature map is upsampled into a soft mask, applied to the input, and the resulting class score determines the weight. This avoids gradient saturation but adds $O(K)$ forward passes per attribution, where $K$ is the number of channels.

**Ablation-CAM** [72] computes weights by explicitly zeroing one channel at a time: $\alpha_k^c = (f_c - f_c^{\setminus k})/f_c$, where $f_c^{\setminus k}$ is the prediction with channel $k$ ablated.

**LayerCAM** [73] aggregates Grad-CAM-style maps across multiple layers, including layers before the final convolution, giving finer spatial resolution.

**Eigen-CAM** [74] uses the principal components of the feature map matrix as weights, removing the class-conditional gradient entirely.

**HiResCAM** [75] corrects a known failure of Grad-CAM: the ReLU-and-average step in (28) can produce attribution maps that do not faithfully reflect the model's prediction. HiResCAM eliminates the spatial pooling and applies the gradient directly, restoring faithfulness at the cost of some spatial smoothness.

**Shap-CAM** [76] reframes the CAM weighting as a Shapley value over channels, providing an axiomatic justification for Score-CAM's perturbation-based weighting.

What unites the CAM family is a *spatial projection* of attribution onto the last convolutional layer. In their usual form, CAM variants are not pixel-level complete decompositions of $f(x) - f(x')$; they are intermediate-layer attributions that are then upsampled to image resolution.

### H. Transformer Attribution: Attention Is Not Enough

For transformer architectures, the most-cited attribution candidate is the attention map itself. The mathematical critique of [77], [78] showed that attention weights are not generally consistent with model behaviour: alternative attention distributions can produce identical or near-identical predictions, so attention is not uniquely identified by the model's output.

Wiegreffe and Pinter [79] treat attention as one plausible explanatory hypothesis without claiming that it is unique. They propose testing whether the attention pattern is plausible under the model.

Two attribution methods have since become standard for transformers:

**Attention rollout** [80] aggregates attention weights across layers by treating each layer's attention matrix as a transition probability and computing the rollout $\tilde{A}^L = \prod_{\ell=1}^{L}(A^{(\ell)} + I)/2$. The product captures the iterated effect of attention across the depth of the model.

**Transformer relevance propagation** [81] extends LRP to transformers, distributing the class score backward through both attention and feed-forward components. It satisfies a conservation axiom at each layer, restoring the property that attention alone does not.

### I. Summary

Gradient and backpropagation methods range from raw saliency (one forward+backward pass, no axioms beyond sensitivity-(b)) to LRP and DeepLIFT (multiple rules, conservation, implementation invariance under the right conditions). They are united by their differentiability assumption and their computational efficiency, but divided on which axioms they satisfy and on how they behave under known sanity checks. Table VIII summarizes the family.

## VII. Perturbation and Occlusion Attribution

Perturbation methods sidestep the question of value functions, paths, and gradients by directly measuring how the prediction changes when parts of the input are modified. They are conceptually simple ("occlude a feature and see what happens") and fully model-agnostic, requiring only black-box access to $f$. Their mathematical content lies in how perturbations are chosen, how predictions on perturbed inputs are aggregated, and how their attributions should be interpreted given the off-manifold inputs they typically produce.

### A. Occlusion Sensitivity and Prediction Difference

The simplest perturbation method is *occlusion*, introduced for CNNs by Zeiler and Fergus [66]: slide a fixed patch across the input image, record the change in the prediction at each position, and use the resulting map as an attribution. Formally, let $\mathcal{P}_p(x)$ denote $x$ with a patch of features $p \subseteq N$ replaced by a reference value:

$$\phi_p^{\mathrm{occ}}(x) = f(x) - f(\mathcal{P}_p(x)). \tag{30}$$

Occlusion is a single-coalition Shapley estimate: it computes the marginal contribution of the patch $p$ to the full feature set $N$, ignoring all other coalitions. It is unbiased when the features outside $p$ are independent of those inside, and biased otherwise.

The *Prediction Difference Analysis* of Zintgraf et al. [82] replaces the deterministic patch with a conditional expectation: features are "removed" by marginalizing over their conditional distribution given the rest of the input. This addresses the off-manifold issue at the cost of requiring a conditional density model.

### B. LIME: Local Linear Surrogates

LIME [3] fits a local linear surrogate to $f$ in a neighbourhood of $x$. Given a similarity kernel $\pi_x(\boldsymbol{z})$ and a perturbation distribution over masks, LIME solves

$$g^* = \arg\min_{g \in \mathcal{G}} \sum_{\boldsymbol{z}} \pi_x(\boldsymbol{z}) \big(f(\mathcal{P}_{\boldsymbol{z}}(x)) - g(\boldsymbol{z})\big)^2 + \Omega(g), \tag{31}$$

where $\mathcal{G}$ is a family of interpretable models (typically sparse linear) and $\Omega$ is a complexity penalty. The coefficients of $g^*$ are the attributions.

LIME is a general additive surrogate method: any choice of mask distribution and kernel yields a method in the class (5). KernelSHAP is the special case in which the kernel is $\pi_{\mathrm{Shap}}$ of (10) and the regression is unregularized; in that case the surrogate coefficients coincide with the Shapley values [2]. Different kernels yield different attributions, and the LIME kernel (exponential of inverse cosine distance, in the original paper) does *not* satisfy the Shapley axioms.

MAPLE [83] is a supervised-neighbourhood variant of the same local-surrogate idea. Instead of sampling an unsupervised neighbourhood around $x$, MAPLE uses tree-ensemble structure to weight training points and then fits a local linear model under that induced neighbourhood. This gives MAPLE a clearer statistical object than a

TABLE VIII
GRADIENT AND BACKPROPAGATION ATTRIBUTION METHODS. "COST" IS IN BACKWARD PASSES; "SANITY" IS THE MODEL-RANDOMIZATION SANITY CHECK OF [9]: PASS MEANS THE ATTRIBUTION CHANGES SUBSTANTIALLY WHEN MODEL WEIGHTS ARE RANDOMIZED.

| Method | Derivative | Completeness | Impl. Inv. | Sanity | Noise smoothing | Cost |
|---|---|---|---|---|---|---|
| Saliency [62] | 1st | no | yes | pass | no | 1 |
| Gradient × Input | 1st | only if linear | yes | pass | no | 1 |
| SmoothGrad [64] | 1st (smoothed) | no | yes | pass | yes | $N$ |
| VarGrad [9] | 1st (variance) | no | yes | pass | yes (variance) | $N$ |
| Guided BP [65] | 1st (rule-modified) | no | no | *fail* | no | 1 |
| Deconvnet [66] | 1st (rule-modified) | no | no | *fail* | no | 1 |
| DeepLIFT [6] | multipliers | yes | yes (Rescale) | pass | no | 1 |
| LRP [5] | rule-based | yes (conservation) | yes (standard rules) | pass | no | 1 |
| FullGrad [68] | 1st + bias | yes (vs. 0 baseline) | yes | pass | no | 1 |
| Grad-CAM [4] | 1st of last conv | no (region only) | layer-dep. | partial | no | 1 |
| Grad-CAM++ [70] | 1st–3rd of last conv | no | layer-dep. | partial | no | 1 |
| Score-CAM [71] | none (forward only) | no | layer-dep. | pass | no | $K$ fwd |
| HiResCAM [75] | 1st (no pooling) | yes (per layer) | layer-dep. | pass | no | 1 |

generic proximity kernel: the local explanation is tied to the predictive neighbourhood learned by random forests [84] or boosted trees [85]. It also illustrates a general lesson for local surrogates: locality is not a purely geometric choice, but a modelling assumption about which perturbations should stand in for nearby counterfactuals.

### C. Anchors: Rule-Based Local Explanations

**Scope note.** Anchors lie at the boundary of the local-additive scope of this survey: their output is a rule predicate, not a per-feature attribution vector, so the linear-in-mask surrogate of (5) does not apply directly. We include Anchors in this section because the underlying *perturbation distribution* is closely related to that of LIME and because Anchors is widely used as a baseline against attribution methods; we mark it as a boundary case in the axiom matrix (Table IX).

Anchors [86] step away from the additive surrogate form entirely and produce a set of rules $A$ such that, conditional on the rules holding, $f(x') = f(x)$ with high probability:

$$\mathbb{P}\big(f(x') = f(x) \,\big|\, A(x') = 1\big) \geq 1 - \delta, \tag{32}$$

for some tolerance $\delta$ and a coverage condition on the rule. The rules are themselves the explanation; they have no real-valued attribution per feature. Anchors and LIME together span the local-surrogate design space: a local linear model versus a local rule.

### D. Meaningful Perturbations and Extremal Masks

Fong and Vedaldi [87] introduced *meaningful perturbations*, which optimize a smooth mask $\boldsymbol{z} \in [0,1]^d$ that maximally suppresses the prediction subject to a sparsity constraint:

$$\min_{\boldsymbol{z}} \; f(\mathcal{P}_{\boldsymbol{z}}(x)) + \lambda \|\boldsymbol{z}\|_1 + \mu \,\mathrm{TV}(\boldsymbol{z}), \tag{33}$$

where TV is a total-variation regularizer. The optimal mask identifies the regions of $x$ whose removal most disrupts the prediction, yielding a localized form of attribution.

The follow-up Extremal Perturbations [88] replaces the unconstrained $\ell_1$ objective with a hard area constraint: "find the smallest mask of area $a$ that maximizes/minimizes the prediction". This avoids sensitivity to the Lagrange multiplier $\lambda$ and yields more interpretable masks.

### E. RISE and Random-Mask Methods

RISE [89] averages over a large number of random masks:

$$\phi_i^{\mathrm{RISE}}(x) = \frac{1}{N} \sum_{n=1}^{N} f(\mathcal{P}_{\boldsymbol{z}^{(n)}}(x))\, \boldsymbol{z}_i^{(n)} \,/\, p, \tag{34}$$

where $p$ is the probability that feature $i$ is unmasked. RISE is *embarrassingly parallel* and requires no gradient access, making it appealing for true black-box settings. It is a Monte Carlo estimator of the marginal expectation of $f$ with respect to the mask distribution, normalized to give a per-feature score.

### F. Counterfactual Generation: FIDO-CA (boundary case)

**Scope note.** FIDO-CA produces counterfactual generations and does not yield per-feature attribution scores; we include it here because the underlying optimisation formulation (mask + generative completion) lies on the same axis as the perturbation-distribution choice made by LIME and RISE, and because its on-manifold mask is a constructive answer to the off-manifold failure mode of mask-based attribution. Strictly speaking, the contribution sits in the counterfactual-explanation literature catalogued in Section XIII.

Chang et al. [43] introduced FIDO-CA, which generates counterfactual perturbations using a generative model. Where RISE and meaningful perturbations replace masked features with constants or blurred values, FIDO-CA fills them with samples from a conditional generator $G(\boldsymbol{z}, x)$:

$$\phi_i^{\mathrm{FIDO}}(x) = \arg\min_{\boldsymbol{z}} \; -\log f_c\big(G(\boldsymbol{z}, x)\big) + \lambda \|\boldsymbol{z}\|_1. \tag{35}$$

The motivation is that the resulting perturbed inputs are on-manifold, addressing one of the principal failure modes of mask-based methods (Section X).

### G. Real-Time Saliency: Learned Mask Predictors

Dabkowski and Gal [90] trained a separate network $g_\psi$ to predict the optimal mask of (33):

$$g_\psi(x) \approx \arg\max_{\boldsymbol{z}} \; f_c(\mathcal{P}_{\boldsymbol{z}}(x)) - \lambda\|\boldsymbol{z}\|_1. \tag{36}$$

The learned mask predictor gives near-instant attributions at inference time but introduces an additional model whose own validity must be verified.

### H. Information-Bottleneck Attribution

Schulz et al. [91] cast mask-based attribution as restricting the information flow through the network. They add Gaussian noise to intermediate activations and optimize the noise variance per spatial location to maximize the suppression of information about the input while preserving the prediction. The resulting per-location noise scale is the attribution. The framing exposes a duality between mask-based attribution and information bottlenecks that is independently interesting; the cost is one optimization per attribution.

### I. Insertion / Deletion Evaluation

Petsiuk et al. [89] also introduced the *insertion / deletion* curves that have since become a standard evaluation tool (Section IX). Features are added or removed in order of attribution magnitude; the area under the resulting prediction-versus-step curve quantifies how well the attribution identifies features that matter to the model. These curves are not an attribution method per se but a meta-evaluation built on any underlying attribution.

### J. Summary

Perturbation methods are the natural complement to gradient methods. They make no differentiability assumption and require no backpropagation, but pay for that generality in two ways: many forward passes per attribution, and a strong dependence on the perturbation distribution. The choice of how to "remove" a feature (by zeroing, blurring, sampling from $p_X$, sampling from $p(X \mid x_{\bar{S}})$, or generating with $G$) determines whether the resulting attribution is interventional, conditional, or on-manifold, and is a direct counterpart of the value-function choice in Shapley methods (Section IV). The same axes therefore organize both families.

## VIII. Mathematical Comparison Framework

This section presents the main comparison table of the survey. Having surveyed Shapley, path-based, gradient/backpropagation, and perturbation methods, we now place them in a single mathematical frame. We compare them along five axes: axioms satisfied, value function choice, baseline distribution, computational complexity, and known equivalences/reductions. The centerpiece is the axiom-by-method matrix (Table IX) recording which axioms each method satisfies. The matrix illustrates that many of the field's central disagreements, examined as failure modes in Section X, arise from incompatible axiom subsets, not from implementation defects.

### A. Axes of Comparison

We compare methods along the following axes:

**Axioms satisfied.**
Catalogued in Section III; collected in Table IX.

**Value function choice.**
Interventional, conditional, or single-reference (Section II).

**Baseline / path / perturbation.**
The specific reference, path, or perturbation distribution used.

**Computational complexity.**
In flops, forward/backward passes, or samples, as a function of $d$ and any approximation parameter.

**Differentiability assumption.**
Required for gradient/path methods, not for Shapley or perturbation methods.

**Causal assumptions.**
Whether the method's output admits a causal interpretation, and under what conditions.

**Stability / scalability.**
Sensitivity to small input changes; behaviour as $d \to \infty$.

### B. Axiom Satisfaction Matrix

Table IX cross-references methods (rows) against axioms (columns). A ✓ means the method satisfies the axiom unconditionally under the assumptions of its original paper; an ∘ marks a conditional claim (specific architecture, sampling limit, propagation rule, or method variant); a blank means the axiom is not satisfied in general; *N/A* marks axioms that are inapplicable to the method's output type. The matrix records *mathematical* axioms only. Empirical evidence about sanity-check behaviour is implementation- and architecture-dependent and lies outside the axiomatic claims. It is reported separately in Table X, following the critique that sanity checks are properties of an implementation tested on a dataset, not of a method formula [9], [11].

### C. Complexity Landscape

Table XI collects the asymptotic computational costs. The methods span seven orders of magnitude: from a single backward pass for raw saliency to $O(2^d)$ for exact Shapley.

### D. Equivalence and Reduction Theorems

Several methods that appear distinct in the literature are mathematically equivalent or related by explicit reductions; the differences in notation between the original papers tend to hide this.

**Result 8** (KernelSHAP and exact Shapley, Lundberg & Lee [2], Thm. 2)**.** *Assumptions:* the interventional value function (1) is used; the weighted least-squares regression in (5) is solved exactly on the full mask distribution (no sampling). Under these assumptions, the KernelSHAP solution equals the exact Shapley value (6). Finite-sample bias and variance are characterized in [20], [52].

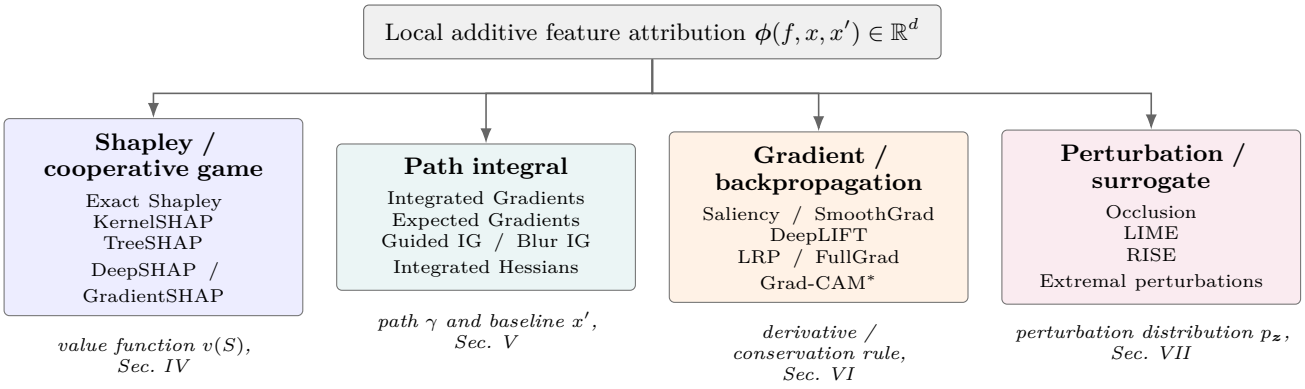


Fig. 4. Taxonomy of local additive feature attribution methods used throughout this survey, with the mathematical object that each family primarily manipulates (italics): a cooperative-game value function, a path integral with its baseline, a backpropagation-style derivative or conservation rule, or a black-box perturbation distribution. Methods marked * (CAM family) are gradient-based but produce spatial attributions through an intermediate-layer projection. Section VIII places all four families in a common axiom-by-method comparison.

**Result 9** (DeepLIFT-Rescale and IG, Ancona et al. [21], §3)**.** *Assumptions:* linear layers and monotone element-wise nonlinearities; no skip connections, concatenations, or non-monotone operators; single fixed baseline. Under these assumptions, DeepLIFT with the Rescale rule and Integrated Gradients (with $K \to \infty$) yield equivalent attributions in the small-input-increment limit. The reduction is sensitive to architectural assumptions and does not extend automatically to transformer or residual networks.

**Result 10** (Grad-CAM as class-conditional gradient projection; reformulation, not reduction)**.** *Assumptions:* CNN with global average pooling followed by a linear classification head. Under these assumptions, the Grad-CAM map of (28) is a rewriting (not a reduction) of a class-conditional spatial average of the final-feature-map gradient [4], and coincides with the original CAM [69] on the GAP-then-linear architecture. We label this a *reformulation* because the two computations yield the same output by construction; it is not a reduction between distinct methods.

**Result 11** (LRP-$\epsilon$ and gradient-$\times$-input, Ancona et al. [21], §3.2–3.3)**.** *Assumptions:* deep ReLU network with no bias terms; LRP with the $\epsilon$-rule applied uniformly to every layer; the input is viewed as the difference $x - \mathbf{0}$ from the zero baseline. Under these assumptions, LRP-$\epsilon$ reduces to gradient $\times$ input, which is the $K = 1$ Riemann approximation of Integrated Gradients with $x' = \mathbf{0}$. The reduction fails when biases are present, when LRP composition (different rules at different layers) is used, or when the baseline is non-zero.

These reductions matter in practice: they imply that empirical disagreements between methods should be tracked back to differences in the axioms (which method ignores which), the value function (which notion of feature absence is in play), or the path (which trajectory through input space is integrated), not to differences in mathematical sophistication.

### E. Taxonomy Table

Table XII situates the methods on three orthogonal axes: their mathematical object (the primary computation), whether they are local or global, and whether they are model-specific or model-agnostic.

### F. The Central Disagreement, Geometrically

A useful way to summarize the comparison is geometric. Each attribution method is a function

$$\Phi \colon \mathcal{F} \times \mathcal{X} \to \mathbb{R}^d \tag{37}$$

mapping a (model, input) pair to an attribution vector, parameterized by a choice of value function $v$, baseline distribution $\mathcal{D}_{x'}$, path $\gamma$, and perturbation distribution $p_{\boldsymbol{z}}$. Two methods that disagree on $\boldsymbol{\phi}(x)$ disagree because they live at different points in

$$\mathcal{C} = \{\text{value fn}\} \times \{\text{baselines}\} \times \{\text{paths}\} \times \{\text{perturbations}\}.$$

The empirical question "which method should I use?" is therefore not a question about $\Phi$ but about which neighbourhood of $\mathcal{C}$ matches the user's intended question. We close this section, and the methods half of the survey, with the following statement.

*Principle* 1. A disagreement between attribution methods $\Phi_1$ and $\Phi_2$ on a prediction $f(x)$ is informative about the model only to the extent that $\Phi_1$ and $\Phi_2$ share a value function, baseline, path, and perturbation distribution. Disagreement under different choices reflects the choices, not the model.

Principle 1 summarizes the comparison: attribution choices are modelling choices, and explanation comparison should hold those choices constant.

## IX. Evaluation Theory and Metrics

Evaluating an attribution method is harder than producing one. There is no ground-truth attribution against which to measure, and the purposes for which attributions are used, including model debugging, user trust, regulatory

TABLE IX

AXIOM SATISFACTION MATRIX. ✓= SATISFIED UNCONDITIONALLY UNDER THE STATED ORIGINAL ASSUMPTIONS; ○ = CONDITIONAL (ARCHITECTURE, LIMIT, RULE, OR METHOD VARIANT); BLANK = NOT SATISFIED; N/A = INAPPLICABLE TO THE METHOD'S OUTPUT TYPE. ABBREVIATIONS: COMP = COMPLETENESS; S(A)/S(B) = SENSITIVITY-(A)/(B); II = IMPLEMENTATION INVARIANCE; CONS = CONSISTENCY; MONO = MONOTONICITY; SYM = SYMMETRY-PRESERVATION; LIN = LINEARITY; CONT = CONTINUITY IN $x$. PER-CELL JUSTIFICATIONS ARE IN APPENDIX A. THE ENTRIES REFER TO CANONICAL FORMULATIONS UNLESS OTHERWISE STATED; SOFTWARE IMPLEMENTATIONS MAY DIFFER.

| Method | Comp | S(a) | S(b) | II | Cons | Mono | Sym | Lin | Cont |
|---|---|---|---|---|---|---|---|---|---|
| *Shapley family* | | | | | | | | | |
| Exact Shapley [44] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ○ |
| KernelSHAP [2] | ✓ | ✓ | ✓ | ✓ | ○ | ○ | ✓ | ✓ | ○ |
| TreeSHAP [45] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ○ |
| DeepSHAP [2] | ○ | ○ | ✓ | ○ | ○ | | ✓ | ✓ | ○ |
| GradientSHAP [38] | ✓ | ✓ | ✓ | ✓ | ○ | | ✓ | ✓ | ○ |
| *Path family* | | | | | | | | | |
| Integrated Gradients [1] | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | ✓ |
| Expected Gradients [38] | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | ✓ |
| Guided IG [41] | ✓ | ✓ | ✓ | ○ | | | | ✓ | ✓ |
| Blur IG [42] | ✓ | ✓ | ✓ | ✓ | | | | ✓ | ✓ |
| Integrated Hessians [56] | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | ✓ |
| *Gradient & backpropagation family* | | | | | | | | | |
| Saliency [62] | | | ✓ | ✓ | | | | ✓ | ○ |
| Grad × Input | ○ | ○ | ✓ | ✓ | | | | ✓ | ○ |
| SmoothGrad [64] | | | ✓ | ✓ | | | | ✓ | ✓ |
| Guided BP [65] | | | | | | | | | ○ |
| Deconvnet [66] | | | | | | | | | ○ |
| DeepLIFT [6] | ✓ | ✓ | ✓ | ○ | | | ✓ | ✓ | ✓ |
| LRP [5] | ✓ | ○ | ○ | ○ | | | | ✓ | ✓ |
| FullGrad [68] | ✓ | ✓ | ✓ | ✓ | | | | ✓ | ✓ |
| Grad-CAM [4] | | ○ | ○ | ○ | | | | ✓ | ✓ |
| HiResCAM [75] | ○ | ✓ | ✓ | ○ | | | | ✓ | ✓ |
| Score-CAM [71] | | ✓ | ○ | ○ | | | | ✓ | ✓ |
| *Perturbation family* | | | | | | | | | |
| Occlusion [66] | | ✓ | ○ | ✓ | | | ✓ | ✓ | ✓ |
| LIME [3] | ○ | ✓ | ○ | ✓ | | | | ✓ | ✓ |
| Meaningful Pert. [87] | | ✓ | ○ | ✓ | | | | | ○ |
| RISE [89] | | ✓ | ○ | ✓ | | | ✓ | ✓ | ✓ |
| *Boundary cases (not local additive on $\mathbb{R}^d$)* | | | | | | | | | |
| Anchors [86] | N/A | ✓ | ✓ | ✓ | | | | N/A | N/A |
| SAGE [53] (global) | N/A | N/A | ✓ | ✓ | ○ | | ✓ | ✓ | N/A |

compliance, and scientific discovery, impose different evaluation criteria. This section surveys the principal evaluation frameworks that have emerged. We organize them around the property each metric purports to measure: faithfulness to the model, stability under input changes, alignment with ground-truth annotations, and the validity of the metric itself.

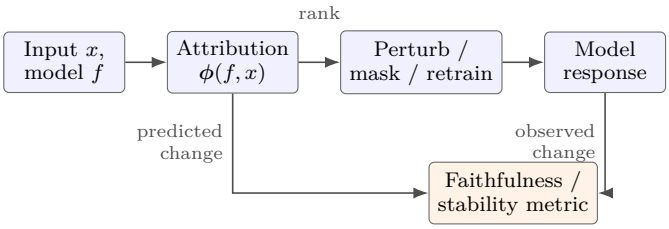


Fig. 5. Generic evaluation pipeline for attribution methods. The attribution $\phi(f,x)$ ranks or selects features; the model is queried under perturbations or retrainings indexed by the ranking; the observed change in the model's response is compared against the change the attribution predicts, producing a faithfulness or stability metric. Section IX catalogues the instantiations.

### A. Faithfulness: Does the Attribution Reflect the Model?

*Faithfulness* measures whether the attribution reports what the model actually does, as opposed to what a human would do given the same input. Faithfulness has been operationalized in several inequivalent ways.

**Perturbation response.**
Remove the top-$k$ features by attribution magnitude and measure the drop in prediction. Used in [92] and codified by DeYoung et al. [93] as "comprehensiveness" (drop when top features are removed) and "sufficiency" (drop when only top features are retained).

**Functional faithfulness.**
$\sum_i \phi_i \approx f(x) - f(x')$, i.e. completeness as an evaluation metric distinct from the axiom. Methods that satisfy completeness pass trivially; methods that do not are penalized.

TABLE X
EMPIRICAL-EVIDENCE COMPANION TO TABLE IX. EACH ROW REPORTS PUBLISHED EMPIRICAL EVIDENCE THAT THE METHOD PASSES (✓) OR FAILS (×) THE MODEL-RANDOMIZATION SANITY CHECK [9] AND THE DATA-RANDOMIZATION SANITY CHECK [9]. ENTRIES MARKED ∘ ARE PARTIAL OR METHOD-VARIANT-DEPENDENT (E.G., GRAD-CAM VARIANTS DIFFER). BLANK EVIDENCE CELLS MAKE NO RANDOMIZATION-CHECK CLAIM FOR THAT METHOD; BLANK SOURCE CELLS ARE INCLUDED ONLY FOR MATRIX COMPLETENESS. SANITY-CHECK BEHAVIOUR DEPENDS ON IMPLEMENTATION, ARCHITECTURE, AND DATASET; ENTRIES HERE ARE NOT AXIOMATIC CLAIMS.

| Method | Model-rand. evidence | Data-rand. evidence | Source(s) |
|---|---|---|---|
| Saliency | ✓ | ✓ | [9] |
| Grad × Input | ✓ | ✓ | [9] |
| SmoothGrad | ✓ | ✓ | [9] |
| Guided BP | × | × | [9] |
| Deconvnet | × | | [9] |
| DeepLIFT (Rescale) | ✓ | | [9] |
| LRP-$\epsilon$ | ✓ | | [9] |
| Integrated Gradients | ✓ | ✓ | [9] |
| Grad-CAM family | ∘ | | [9], [75] |
| HiResCAM | ✓ | | [75] |
| KernelSHAP / TreeSHAP | | | |
| LIME | | | |
| RISE | | | source reports perturbation metrics, not randomization checks |

TABLE XI
COMPUTATIONAL COMPLEXITY OF REPRESENTATIVE METHODS. net DENOTES ONE FORWARD+BACKWARD PASS THROUGH THE MODEL. $T, L, D$ ARE THE NUMBER OF TREES, LEAVES, AND TREE DEPTH; $K$ IS THE NUMBER OF PATH STEPS; $M$ THE NUMBER OF SAMPLES; $N$ THE NUMBER OF SMOOTHGRAD NOISE DRAWS.

| Method | Complexity |
|---|---|
| Saliency / Grad × Input | $O(\text{net})$ |
| SmoothGrad / VarGrad | $O(N \cdot \text{net})$ |
| DeepLIFT, LRP, FullGrad | $O(\text{net})$ |
| Grad-CAM family | $O(\text{net})$ |
| Score-CAM | $O(K \cdot \text{net}_{\text{fwd}})$ |
| Integrated Gradients | $O(K \cdot \text{net})$ |
| Expected Gradients / GradientSHAP | $O(K \cdot M \cdot \text{net})$ |
| Integrated Hessians | $O(K^2 \cdot \text{net})$ |
| TreeSHAP | $O(TLD^2)$ |
| KernelSHAP (sampled) | $O(M \cdot d \cdot \text{net}_{\text{fwd}})$ |
| Exact Shapley | $O(2^d \cdot \text{net}_{\text{fwd}})$ |
| LIME (sampled) | $O(M \cdot \text{net}_{\text{fwd}})$ |
| RISE | $O(N \cdot \text{net}_{\text{fwd}})$ |
| Anchors | $O(M \cdot \text{net}_{\text{fwd}})$ (rule search) |

**Causal faithfulness.**
Defined by Jacovi and Goldberg [94], who distinguish faithfulness from plausibility and argue that the two are often conflated in NLP attribution.

### *B. Infidelity and Sensitivity*

Yeh et al. [95] proposed two metric families that have become standard.

**Definition 1** (Infidelity)**.** *For an attribution $\boldsymbol{\phi}$ and a perturbation distribution $\mu_I$,*

$$\text{INFD}(\boldsymbol{\phi}, f, x) = \mathbb{E}_{I\sim\mu_I}\Big[\big(I^\top \boldsymbol{\phi} - (f(x) - f(x-I))\big)^2\Big]. \quad (38)$$

Infidelity measures the mean-squared error between (i) the attribution's prediction of how the model output changes under perturbation $I$ and (ii) the actual change. Methods that satisfy completeness in expectation under $\mu_I$ minimize (38). Different choices of $\mu_I$ yield different metrics: a Gaussian $\mu_I$ tests robustness to noise; a sparse $\mu_I$ tests feature removal.

**Definition 2** (Max-Sensitivity)**.** $\text{SENS}_{\max}(\boldsymbol{\phi}, f, x, r) = \sup_{\|y-x\|\le r} \|\boldsymbol{\phi}(f, y) - \boldsymbol{\phi}(f, x)\|$.

Max-sensitivity bounds how much the attribution can change when the input is perturbed within radius $r$. It is closely related to the Lipschitz stability axiom (Axiom 16); a Lipschitz attribution has bounded max-sensitivity for every $r$.

Yeh et al. also characterize the optimum: for a fixed perturbation distribution $\mu_I$, the attribution minimizing infidelity (38) is the generalized least-squares solution

$$\boldsymbol{\phi}^* = \mathbb{E}_{I\sim\mu_I}\big[II^\top\big]^{-1}\,\mathbb{E}_{I\sim\mu_I}\big[I\,\big(f(x) - f(x-I)\big)\big], \quad (39)$$

a smoothed, kernel-weighted gradient of $f$ around $x$. They further show that kernel smoothing of a given attribution lowers its max-sensitivity and, under the conditions stated in their paper, does not worsen its infidelity; empirically it often improves both. Stability and fidelity are therefore not strictly opposed. The real tension is between fidelity to $f$ at the point $x$ and fidelity averaged over the neighbourhood that $\mu_I$ defines.

### *C. ROAR and KAR: Retraining-Based Faithfulness*

Hooker et al. [12] argued that perturbation-response metrics like (30) are confounded by distribution shift: removing features may simply push the input off-manifold, in which case a prediction drop may reflect manifold violation and fail to measure feature importance.

**ROAR (RemOve and Retrain)** addresses this by retraining the model after feature removal. Given an attribution method, mask the top-$k$% of features, train a fresh model on the masked data, and compare its test accuracy to a baseline that masks $k$% randomly. A faithful attribution should produce a substantially larger accuracy drop than random.

TABLE XII
TAXONOMY OF ATTRIBUTION METHODS.

| Family | Math. object | Local/Global | Model-specific | Input type | Output |
|---|---|---|---|---|---|
| Shapley / cooperative game | coalition value function | local (mostly) | no | any | vector $\in \mathbb{R}^d$ |
| Path-based | line integral of gradient | local | yes (diff.) | continuous | vector $\in \mathbb{R}^d$ |
| Gradient / backprop | 1st-order derivative / rules | local | yes (diff.) | continuous | vector / heatmap |
| CAM family | spatial gradient projection | local | yes (CNN) | image | 2D heatmap |
| Perturbation / occlusion | black-box query | local | no | any | vector / heatmap |
| Surrogate (LIME) | local linear regression | local | no | any | vector |
| Rule-based (Anchors) | rule predicate | local | no | any | rule set |
| SAGE | global Shapley over loss | global | no | any | vector $\in \mathbb{R}^d$ |

**KAR (Keep And Retrain)** is the dual: retain only the top-$k$% features. A faithful attribution should still allow the model to learn.

ROAR/KAR is expensive because each evaluation point requires a full retraining, and so has been applied at small scale, but it is one of the more defensible faithfulness criteria in the literature because it explicitly controls for distribution shift.

### D. Sanity Checks

Adebayo et al. [9] proposed the *model-randomization* and *data-randomization* sanity checks:

- *Model-randomization:* replace the model's weights (layer-by-layer or in cascade) with random values; the attribution should change substantially. An attribution that does not is not explaining the model.
- *Data-randomization:* train the model on data with permuted labels; the attribution should change. An attribution that does not is not sensitive to what the model has learned.

These checks are *necessary* for an attribution method to be called an explanation, but they are not sufficient: a method can pass both and still be uninformative for downstream tasks.

The most-discussed empirical finding from Adebayo et al. is the failure of guided backpropagation and deconvolution under the model-randomization check. We incorporated this finding directly into Table X.

### E. Insertion / Deletion Curves

For image attribution, Petsiuk et al. [89] introduced two curves:

**Insertion.**
Start from a baseline (e.g., a blurred image) and add features in decreasing order of attribution magnitude. Plot the prediction as a function of the number of features inserted; a faithful attribution should rise quickly.

**Deletion.**
Start from $x$ and delete features in decreasing order of attribution. The prediction should fall quickly.

The area under the resulting curves (AUC-Ins, AUC-Del) summarizes the method's performance; Figure 6 sketches the geometry. The metric has the virtue of being entirely model-internal: no ground-truth annotation is required.

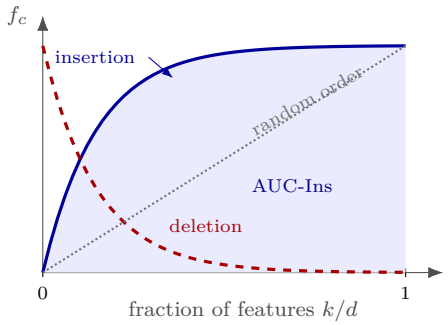


Fig. 6. Insertion and deletion curves [89]. Features are inserted into a baseline input (solid) or deleted from $x$ (dashed) in decreasing order of attribution magnitude. A faithful attribution makes the prediction rise quickly under insertion and fall quickly under deletion, relative to a random ordering (dotted); the shaded area is AUC-Ins. Both curves depend on the baseline used for insertion and on the replacement value used for deletion, which is the same baseline-sensitivity caveat as in Section X.

### F. Localization Metrics

When ground-truth annotations are available, such as bounding boxes for object detection or rationales for text classification, the attribution can be compared directly to them:

**Pointing game.**
Does the maximum-attribution pixel lie within the ground-truth bounding box? (Standard in CAM evaluation.)

**IoU.**
Treat the top-$k$ attributions as a predicted mask and compute intersection-over-union with the ground-truth mask.

**Token F1 / IOU.**
In NLP, treat selected tokens as a predicted rationale and compare against human rationales [93].

Localization metrics conflate plausibility (alignment with human annotations) with faithfulness (alignment with model behaviour). A method can have perfect IoU while being unfaithful (e.g., if the model relies on features outside the ground-truth box but the attribution defers to the human annotation).

### G. Using Attribution During Training

Another evaluation route is to make explanations actionable during training. Ross, Hughes, and Doshi-Velez [96] penalize input gradients at annotated irrelevant dimensions, training models that are accurate and at the same time constrained away from known spurious reasons. Rieger et al. [97] extend this idea through contextual-decomposition explanation penalization, including feature interactions. These methods do not replace post-hoc faithfulness tests; they show that an attribution method can be validated by whether its signal can guide a model away from documented confounders under a stated training objective.

### H. Human-Grounded Evaluation

Doshi-Velez and Kim [23] categorized evaluations into three tiers: *functionally-grounded* (model-internal, no human required), *human-grounded* (simplified tasks with non-expert users), and *application-grounded* (deployed tasks with domain experts). Most attribution evaluations have been functionally-grounded; application-grounded evaluations remain rare and are often the most decision-relevant.

DeYoung et al. [93] introduced the ERASER benchmark, which provides human rationales for several NLP tasks and treats attribution evaluation as a token-selection problem. The benchmark exposes a striking gap: even attribution methods that score well on functional metrics often fail to identify the tokens that humans annotate as evidence.

For model debugging in text classification, Bastings et al. [98] propose a complementary shortcut-based protocol: inject known lexical shortcuts, verify that the trained model uses them, and evaluate whether salience methods rank the shortcut tokens near the top. This protocol is important because it provides a controlled ground truth for model reliance, something a human rationale alone cannot supply.

### I. Sanity Checks on the Metrics

A second-order concern, raised by Tomsett et al. [11], is that the evaluation metrics *themselves* can fail sanity checks. They showed that several popular faithfulness metrics correlate weakly across methods. A method ranked highest by one metric is often ranked lowest by another, and some metrics depend strongly on hyperparameters in ways the original authors did not document. The methodological implication is that no single metric should be treated as authoritative; a method should be evaluated against multiple, ideally diverse, metrics, with the disagreements among them reported.

### J. Summary

Table XIII summarizes the principal evaluation metrics. No single metric dominates: faithfulness, stability, and plausibility are different quantities, and a method can score well on one while failing another. What the metrics jointly require is stated as items R8–R9 of the reporting checklist (Section XII).

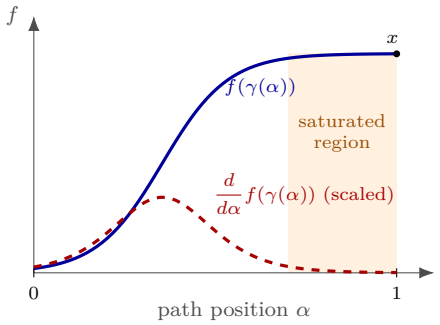


Fig. 7. Gradient saturation along an attribution path. For a sigmoidal $f$, the gradient (dashed) is concentrated in the middle of the path and nearly zero at the input $x$ itself (shaded band). Saliency evaluated at $x$ misses the feature; the path integral (15) accumulates the mid-path gradient mass and recovers the full prediction difference, as in the worked example (40).

## X. Failure Modes and Theoretical Gaps

The standard analyses of attribution failure modes by Adebayo et al. [9], Kindermans et al. [10], Ghorbani et al. [14], Slack et al. [15], and Kumar et al. [18] present these failures as discoveries about specific methods. We recast them as mathematical phenomena tied to specific choices of value function, baseline, path, or perturbation distribution. This recasting clarifies which failures are artefacts of method design and which are intrinsic to the attribution problem itself.

### A. Gradient Saturation and Shattering

Let $f$ be differentiable. *Saturation* occurs when $\partial f/\partial x_i \approx 0$ in a region around $x$, even though $f$ is locally sensitive to $x_i$ in a broader sense. The canonical example is a ReLU network in which a unit's gradient is zero on its inactive side; on long inputs the cumulative effect is that important features receive zero saliency.

For example, take $f(x) = 1 - e^{-5x}$ at $x = 1$ with baseline $x' = 0$. The prediction difference is $\Delta f = 1 - e^{-5} \approx 0.993$, yet the local gradient is $f'(1) = 5e^{-5} \approx 0.034$: saliency and gradient $\times$ input both report the feature as nearly irrelevant. The path integral repairs this by averaging the gradient over the whole trajectory and avoiding reliance on the saturated endpoint:

$$\mathrm{IG}_1(1;0) = (1-0)\int_0^1 5e^{-5\alpha}\,d\alpha = 1 - e^{-5} = \Delta f. \quad (40)$$

Figure 7 plots the same effect for a sigmoidal model along its attribution path.

Saturation is a property of the local geometry of $f$, so any purely local method, including raw saliency, gradient $\times$ input, and the $K = 1$ Riemann approximation of IG, is vulnerable. The mathematical fix is to integrate: IG with sufficiently large $K$ reduces saturation effects by averaging over the path and recovering completeness as the numerical integral converges. Path methods reduce this failure mode, although the result still depends on the baseline and path.

*Gradient shattering* is a stronger failure: in deep networks, gradients become high-frequency and approximately

TABLE XIII
EVALUATION METRICS FOR ATTRIBUTION METHODS. "REQUIRED" IS WHAT THE METRIC NEEDS BEYOND THE ATTRIBUTION AND THE MODEL; "LIMITATIONS" LISTS THE PRINCIPAL KNOWN FAILURE MODES.

| Metric | What it measures | Required | Principal limitation |
|---|---|---|---|
| Completeness check | functional faithfulness | none | trivial for methods satisfying it as axiom |
| Comprehensiveness / Sufficiency [93] | top-$k$ perturbation response | none | off-manifold distortion |
| Infidelity [95] | MSE between attribution and $\Delta f$ | noise distribution $\mu_I$ | depends on choice of $\mu_I$ |
| Max-sensitivity [95] | local stability of $\phi$ | radius $r$ | infidelity/sensitivity trade-off |
| ROAR / KAR [12] | retraining-based faithfulness | retraining budget | very expensive |
| Sanity (model-rand.) [9] | non-degeneracy w.r.t. model | weight randomization | necessary, not sufficient |
| Sanity (data-rand.) [9] | non-degeneracy w.r.t. data | label permutation | necessary, not sufficient |
| Insertion / Deletion AUC [89] | ordering quality of attribution | baseline (e.g., blur) | strong baseline dependence |
| Pointing game / IoU | alignment with bounding box | ground-truth box | conflates plausibility, faithfulness |
| ERASER token F1/IoU [93] | alignment with human rationale | human rationale | limited to text tasks |
| Shortcut protocol [98] | recovery of known model shortcut | controlled shortcut data | task construction must match debugging hypothesis |
| Application study | downstream utility | human users + task | expensive, low statistical power |

independent across inputs, with white-noise-like behaviour inconsistent with a smooth function. The effect grows with depth and is a known property of deep networks [99]. SmoothGrad addresses shattering by averaging gradients over a noise neighbourhood, effectively replacing $f$ with a smoothed version $\tilde{f}_\sigma$. The trade-off is between locality (small $\sigma$) and stability (large $\sigma$).

### B. Baseline Sensitivity

For any method that requires a baseline, including IG, EG, DeepLIFT, LRP, and Shapley methods with $v^{\text{int}}$, the attribution depends on the baseline choice. Sturmfels et al. [59] showed how strong this dependence can be: for the same model and input, IG with a zero baseline and IG with a mean baseline can yield attributions that disagree on the sign for most features.

This is not strictly speaking a failure because a baseline is a modelling choice and different baselines answer different questions. However, the empirical literature often treats it as one. We propose the following classification:

*Remark* 2. A baseline-dependence "failure" is meaningful only if (i) the researcher does not state which baseline they used, or (ii) the researcher claims their attribution is canonical when it depends on a hyperparameter. Otherwise, baseline dependence is the correct behaviour.

### C. Correlated and Dependent Features

Real features are correlated, and most attribution methods are defined assuming independence at one or more points. The interventional value function (1) produces inputs with implausibly extreme features when applied to highly correlated coordinates; the conditional value function (2) requires estimating $d$-dimensional conditional distributions, which is intractable except in special cases.

The mathematical statement of the problem, paraphrasing Theorem 3 of Janzing et al. [24], is:

**Result 12** (Paraphrase of Janzing et al. [24], Thm. 3)**.** For some $f$ and some pairs of features $X_i, X_j$ that are perfectly correlated under $p_X$, the Shapley value $\phi_i^{\text{Shap}}$ under $v^{\text{int}}$ and under $v^{\text{cond}}$ can differ by an arbitrary factor. The precise sufficient conditions are in the cited paper.

The implication is sharp: a researcher who reports "the Shapley value of feature $i$" without specifying the value function has reported an underdetermined quantity.

Approximate solutions: Aas et al. [39] estimate $v^{\text{cond}}$ via a multivariate Gaussian or copula model; Frye et al. [25] use a learned generative model to sample on-manifold coalition completions; Janzing et al. [24] argue that the question itself should be reformulated as a causal one.

### D. Off-Manifold Perturbations

Many evaluations perturb the input by setting features to zero, blurring them, or replacing them with samples from a marginal distribution. The resulting inputs typically lie far from the data manifold, and the model's prediction at these inputs is undefined in a deeper sense: the model was never trained on such inputs and its behaviour there is extrapolation, not interpretation.

Mathematically, off-manifold perturbation breaks the implicit assumption of all perturbation-based methods that $f$ behaves locally well outside $\text{supp}(p_X)$. Hooker et al. [12] showed that the apparent "faithfulness" of an attribution measured by perturbation response can be entirely explained by the model's degradation on off-manifold inputs, not by the attribution's identification of important features.

The principled fix is to constrain perturbations to the data manifold: generative-model perturbations (FIDO-CA [43], Schulz et al. [91]) and on-manifold Shapley (Frye et al. [25]) both implement this. The cost is an additional model whose own validity must be verified.

### E. Causal Misinterpretation

A pervasive failure lies in interpretation: reading attribution as a causal effect. The following statements have all appeared in published applied papers:

- "Feature $i$ caused the prediction $f(x) > 0$."
- "Increasing feature $i$ by 1 unit would change the prediction by $\phi_i$ units."
- "Feature $i$ is more important than feature $j$ for the underlying phenomenon."

None of these is supported by any standard attribution method without strong additional assumptions about the data-generating process. Attribution methods compute *associations* (in the loose sense of sensitivity, marginal contribution, or conditional expectation), typically with respect to a model trained by maximum likelihood. They do not compute causal effects on the world.

Janzing et al. [24] give a constructive bridge: a class of attribution methods that admit a causal interpretation under the assumption that the model is a fair approximation of the conditional expectation and that the user has correctly specified the causal graph. The conditions are strong; the broader point is that attribution-as-causation requires a separate, explicit causal model.

### F. Adversarial Manipulation of Explanations

Adversarial attacks on explanations show that the attribution map $\boldsymbol{\phi}(f, x)$ can be made nearly arbitrary by a small, nearly imperceptible perturbation $\delta$ that preserves the prediction $f(x + \delta) \approx f(x)$.

**Result 13** (Adversarial manipulability of explanations; Ghorbani et al. [14] and Dombrowski et al. [13])**.** For standard gradient-based attributions on ReLU networks, including saliency, gradient $\times$ input, and Grad-CAM, the cited papers construct, for typical inputs $x$, a perturbation $\delta$ of small norm $\|\delta\| \leq \epsilon$ such that the prediction is nearly preserved, $|f(x + \delta) - f(x)|$ is at most $O(\epsilon)$, while the attribution map changes by a substantial amount $\|\boldsymbol{\phi}(x + \delta) - \boldsymbol{\phi}(x)\| \geq C$, where $C$ can be made arbitrarily large. The constructions exploit the piecewise-linear geometry of ReLU networks; we state this as a summary grounded in the cited constructions and do not supply a self-contained proof.

The geometric intuition due to Dombrowski et al. [13] is that for ReLU networks, the prediction is piecewise linear but the attribution is piecewise *constant*; small movements across pieces change the attribution discontinuously while leaving the prediction nearly unchanged. The mathematical fix is to enforce Lipschitz stability (Axiom 16), which SmoothGrad approximates and expected-gradient methods enforce in expectation.

Slack et al. [15] extended adversarial attacks to LIME and SHAP: a model can be trained that behaves discriminatorily on the data distribution but produces non-discriminatory explanations under LIME or KernelSHAP. The attack exploits the off-manifold queries that LIME and KernelSHAP make: a model that is anomalous on those queries can hide its real behaviour. The fix is the same as for the off-manifold problem: constrain queries to the data manifold.

### G. Disagreement Among Methods

A practical concern, often the first one users encounter, is that different attribution methods disagree on the same prediction. Krishna et al. [7] formalized this as the *disagreement problem*, measured across six methods and four benchmarks; their results show median rank correlations below 0.5 between methods that are widely treated as interchangeable. Bilodeau et al. [8] provided the complementary theoretical statement: no attribution method can simultaneously satisfy a small list of intuitive properties (linearity, completeness, and a faithfulness-style criterion), so disagreement is not an empirical accident of any particular method.

Section VIII (Principle 1) established that disagreement is informative only when the methods share a value function, baseline, path, and perturbation distribution. The most common form of reported disagreement, for example between vanilla saliency and KernelSHAP, fails this condition: the methods compute different quantities, so it would be strange if they agreed.

The implication for practice is that disagreement calls for further investigation before it supports any conclusion. "IG and SHAP disagree" is not the same finding as "IG with the same baseline, value function, and number of samples produces a different ordering on different runs"; only the latter indicates a method-level instability. We propose a short protocol for investigating disagreement: (i) verify that both methods use the same scalar output $f_c(x)$ and the same feature granularity; (ii) compare value functions (interventional vs. conditional) and baselines; (iii) quantify approximation error in each method against a higher-budget reference; (iv) only if (i)–(iii) match, report the methods as substantively disagreeing.

### H. Specification-to-Failure Map

Table XIV summarizes the methodological lesson of the preceding subsections: many failures are traceable to an unstated choice in the attribution specification. The table is a diagnostic map for deciding what a paper must report before an attribution claim can be interpreted.

### I. Summary: Failure Modes Are Specifications of Assumptions

Table XV summarizes the failure modes. The unifying view is that each failure occurs when an attribution method is asked a question it was not designed to answer. Saturation occurs when a method is asked about a region in which the gradient is locally uninformative. Baseline sensitivity occurs when a method is asked for a result without a baseline being specified. Off-manifold failure occurs when a method is asked to extrapolate. Causal misinterpretation occurs when a method is asked for a causal answer it does not provide.

TABLE XIV
HIDDEN CHOICES AND THE FAILURE MODES THEY COMMONLY INDUCE.

| Hidden choice | Failure mode | Example |
|---|---|---|
| Off-manifold perturbation | Implausible explanations | occlusion, SHAP, LIME |
| Baseline choice | Baseline sensitivity | IG, DeepLIFT |
| Gradient saturation | Missing important features | saliency, low-$K$ IG |
| Rule-modified backpropagation | Sanity-check failure | guided backprop |
| Sampling distribution | Instability / variance | LIME, KernelSHAP, RISE |

Each failure has a structural fix: path integration for saturation, explicit baselines for sensitivity, manifold-aware perturbations for off-manifold, and explicit causal modelling for causal claims. None of those fixes is method-agnostic. The implication is the following: an attribution paper claiming general superiority over prior methods should identify which failure mode it addresses and what assumptions or trade-offs the mitigation introduces.

## XI. Applications and Model Families

The methods of Sections IV–VII differ in their suitability across model families and input modalities. This section is brief by design: our focus is the mathematical analysis, not the application landscape. We summarize, for each major model family, which method choices are mathematically appropriate, which are empirically established, and which open problems remain.

### A. Tabular Models

Tabular models, including gradient-boosted trees [85], random forests [84], generalized linear models, and deep tabular networks, are the original domain of SHAP and remain its strongest fit. TreeSHAP [45] gives *exact* Shapley values in polynomial time for tree ensembles, removing the principal obstacle (exponential cost) that applies to other Shapley implementations.

For tabular data the choice between interventional and conditional value functions is unusually consequential: real tabular features are often highly correlated (e.g., age, tenure, and account balance in credit scoring), and the two value functions can produce contradictory explanations. The current recommendation, following [17], [24], is to report interventional TreeSHAP as the default and to discuss conditional TreeSHAP when correlation between features is structural and meaningful (e.g., when two columns are functionally related). LIME and KernelSHAP remain competitive for non-tree tabular models.

### B. Convolutional Image Models

For CNNs, the attribution literature has converged on a small set of methods: Grad-CAM and HiResCAM for coarse spatial attribution; Integrated Gradients and Expected Gradients for pixel-level attribution; LRP and DeepLIFT for backpropagation-style attribution; RISE and Score-CAM for gradient-free attribution; and XRAI for region-aggregated attribution.

The key methodological points specific to images are:

- The zero baseline is almost always wrong for image inputs; a blurred or mean-image baseline is more appropriate [59].
- Pixel-level attribution is often dominated by high-frequency artefacts; SmoothGrad smoothing or XRAI region aggregation are usually required for human-readable maps.
- Adebayo et al. [9] sanity-checks should be run for any image attribution method; failure of the sanity check implies the heatmap does not depend on what the model has learned.

### C. Transformers and NLP

For transformer language models, attention rollout [80] and transformer relevance propagation [81] are the principal model-aware attribution methods. For NLP classification specifically, the ERASER benchmark [93] provides the standard evaluation; LIME, IG (with a token-embedding baseline), DeepLIFT, and LRP all have NLP-specific implementations [100].

The chief NLP-specific difficulty is that token-level attribution ignores compositional structure: a sentence's meaning depends on the interaction of tokens, not their individual contributions. Hierarchical methods [101], contextual decomposition [102], and feature-interaction attribution [103] address this directly. The debate between attention-as-explanation [77]–[79] has not been settled but has clarified that attention is at best a partial input-feature attribution: it captures which tokens the model attends to but not why those attentions matter for the prediction. Shortcut-based evaluation [98] is especially useful in NLP because many relevant confounders are lexical or template-like, and therefore can be injected, verified, and evaluated under controlled data modifications.

For *large* language models, attribution is largely an open problem (Section XIV): the same input may produce different generations under different sampling, gradient computations are expensive at trillion-parameter scale, and the meaningful unit of explanation may be sub-token, token, span, or document. Token-level attribution for large language models inherits the same issues surveyed here but adds complications from discrete inputs, subword tokenization, prompt dependence, retrieval context, and generation-time decoding. We treat these as extensions within the local attribution problem and outside the survey's separate targets.

### D. Graph Neural Networks

For GNNs, attribution must respect graph structure: features include node attributes, edge attributes, and the graph topology itself. Extensions of LIME (GraphLIME), SHAP (GraphSHAP), and Integrated Gradients (GNN-IG) to graphs have been developed in dedicated lines of work,

TABLE XV
FAILURE MODES OF ATTRIBUTION METHODS, THE METHODS PRINCIPALLY AFFECTED, THE FORMAL CAUSE, AND THE STRUCTURAL MITIGATION.

| Failure mode | Methods affected | Formal cause | Mitigation |
|---|---|---|---|
| Gradient saturation | saliency, $\nabla f \times x$, low-$K$ IG | $\nabla f \approx 0$ on the path | $K \to \infty$, DeepLIFT-Rescale |
| Gradient shattering | many gradient methods on deep nets | high-frequency $\nabla f$ | SmoothGrad ($\sigma > 0$) |
| Baseline sensitivity | IG, EG, DeepLIFT, LRP, $v^{\text{int}}$ Shapley | answer depends on $x'$ | report $x'$; use expected baseline |
| Correlated features | $v^{\text{int}}$ Shapley, KernelSHAP | off-manifold coalition completions | conditional Shapley / manifold-aware |
| Off-manifold perturbations | LIME, KernelSHAP, RISE, occlusion | $f$ undefined outside $\text{supp}(p_X)$ | generative perturbations |
| Causal misinterpretation | associational methods | associational $\neq$ causal | causal Shapley + DAG |
| Adversarial manipulation | non-Lipschitz methods | Result 13 | Lipschitz / smoothed attribution |
| Method disagreement | cross-family comparisons | differing value fn, path, $\mathcal{D}_{x'}$ | hold method-parameters constant |

which we do not survey here. The mathematical points relevant to our axiomatic framing are: (i) the value function must condition on both the graph adjacency structure and the node features; (ii) permutation symmetries induced by graph isomorphism interact with the symmetry axiom; (iii) message-passing computations admit a layer-wise relevance propagation extension analogous to LRP for feed-forward networks.

### *E. Biomedical and High-Stakes Domains*

In biomedical applications, the failure modes of Section X are not abstract concerns. A clinical decision-support system that attributes credit to a feature for spurious reasons (gradient saturation, off-manifold perturbation, adversarial weight choice during training) can produce confident, visually convincing explanations that are systematically misleading.

The methodological consensus emerging in this domain prefers:

- Methods that satisfy completeness and sensitivity-(a) as axioms (IG, SHAP variants, LRP, DeepLIFT) over methods that satisfy them only approximately;
- Methods with documented sanity-check behaviour;
- Multiple methods reported in parallel, with disagreement analyzed under Principle 1 of Section VIII;
- Training-time explanation constraints when domain knowledge identifies invalid reasons for a prediction [96], [97];
- Application-grounded evaluation [23] involving domain experts in addition to functional metrics.

These are the same principles that emerge in any high-stakes decision domain: legal, financial, hiring. The mathematical machinery of axioms, value functions, and failure modes is the common ground; domain expertise determines which axioms are required for which decisions.

### *F. Application-Domain Matrix*

Table XVI summarizes which methods are preferred across application domains, the principal cautions, and standard evaluation conventions.

## XII. A PROPOSED REPORTING CHECKLIST FOR ATTRIBUTION STUDIES

Attribution methods encode modelling assumptions, and comparison or use of an attribution requires those assumptions to be stated. This section converts that observation into a proposed reporting checklist: a paper or applied study that uses an attribution method should report enough information that another researcher can (i) reproduce the attribution from the same inputs and model, (ii) place the result correctly on the axiom matrix (Table IX) and the failure-mode table (Table XV), and (iii) decide whether the reported attribution is appropriate for the user's question.

The checklist has ten items, organized into three blocks: *specification of the question*, *specification of the method*, and *specification of the evidence*.

### *A. Block A: Specifying the Question*

**R1. Model and scalar output.** State (i) the model $f$ being explained, including its architecture, training data, and whether it is treated as a black box or as a differentiable function; (ii) the scalar output $f_c(x)$ being attributed, including the class logit, post-softmax probability, regression output, loss, or other quantity; and (iii) the input $x$ and whether the attribution is local (per-input) or aggregated.

**R2. Feature definition, grouping, and display transform.** State what counts as a feature: raw pixels, super-pixels, tokens, sub-tokens, structured groups, one-hot categories, or graph nodes/edges. Feature granularity is part of the attribution problem definition: attributions over pixels, superpixels, tokens, words, and domain variables are not directly comparable unless the grouping map is stated. Also state whether attributions are signed, absolute-valued, normalized, clipped, smoothed, thresholded, aggregated, or recolored before visualization.

**R3. Reference / baseline distribution.** State the baseline $x'$ or reference distribution $\mathcal{D}_{x'}$ used. If a single $x'$ is used, state its choice (zero, mean, blurred image, paraphrase, etc.). If an expected baseline is used, state the

TABLE XVI
APPLICATION-DOMAIN MATRIX FOR ATTRIBUTION METHODS.

| Domain | Preferred methods | Principal cautions | Standard evaluation |
|---|---|---|---|
| Tabular | TreeSHAP, KernelSHAP, LIME | feature correlation; value-function choice | ROAR, insertion/deletion |
| Image (CNN) | IG, EG, Grad-CAM, HiResCAM, XRAI, RISE | baseline choice; sanity checks; off-manifold | Insertion/Deletion AUC, sanity checks, IoU |
| Text (transformer) | IG (token), DeepLIFT, transformer LRP, attention rollout | attention vs. attribution; compositional structure | ERASER token F1/IoU, comprehensiveness/sufficiency |
| Graph (GNN) | GNN-IG, GraphSHAP, GNN-LRP | graph structure as feature; permutation symmetry | ablation by edge / node subset |
| Biomedical | TreeSHAP / IG with documented baselines | off-manifold; adversarial trojans; multi-method reporting | ROAR + application-grounded |
| Sequence | Contextual decomposition, hierarchical, LRP-RNN | long-range dependence; vanishing gradients | ERASER-style + perturbation |
| LLM (open problem) | N/A | cost; sampling randomness; meaningful unit | (see Section XIV) |

TABLE XVII
TEN-ITEM REPORTING CHECKLIST FOR LOCAL ADDITIVE ATTRIBUTION. THE ITEMS SPECIFY THE MINIMUM INFORMATION NEEDED TO REPRODUCE AND INTERPRET AN ATTRIBUTION CLAIM.

| Item | What to report | Why it matters |
|---|---|---|
| R1 | Model, training context, and scalar output | fixes the target of attribution |
| R2 | Feature granularity, grouping, and display transform | prevents comparing pixels, tokens, words, and groups as if identical |
| R3 | Baseline or reference distribution | defines the comparison point |
| R4 | Value function or perturbation distribution | defines feature absence and locality |
| R5 | Path, coalition strategy, or sampling design | determines approximation and axioms |
| R6 | Axioms satisfied and not satisfied | states the mathematical guarantees |
| R7 | Approximation budget, stochasticity, seeds, and uncertainty | makes sampled attributions reproducible |
| R8 | Sanity-check behaviour | tests model and data dependence |
| R9 | Faithfulness or stability metric | evaluates the attribution claim |
| R10 | Known failure modes for the setting | prevents overinterpretation |

distribution $p_X$ from which references are drawn and the number of samples.

### B. Block B: Specifying the Method

**R4. Value function or perturbation distribution.** State explicitly which value function is used: interventional ((1)), conditional ((2)), single-reference ((3)), or other. For perturbation methods, state the perturbation distribution $p_{\boldsymbol{z}}$. This item fixes the meaning of feature absence and locality.

**R5. Path or coalition strategy.** For path methods, state the path $\gamma$ (straight-line, adaptive, scale-space) and the number of Riemann steps $K$. For coalition methods, state the sampling strategy (KernelSHAP weighting, permutation sampling, TreeSHAP recursion) and the sample budget $M$.

**R6. Axioms satisfied / not satisfied.** State which axioms from Section III the chosen method satisfies, under the chosen value function and baseline. Where a known reduction or equivalence applies (e.g., Result 8, Result 9), cite it.

**R7. Computational approximation and stochasticity.** State the approximation error budget and the corresponding sample size, integration steps, or network-pass count. For sampling-based methods, report the number of samples, random seeds, variance estimates, and convergence diagnostics where feasible.

### C. Block C: Specifying the Evidence

**R8. Sanity-check results.** Report the outcome of at least one of the model-randomization and data-randomization sanity checks of Adebayo et al. [9]. A method that is largely insensitive to model parameters should not, by itself, be used to support strong claims about model-specific reasoning.

**R9. Faithfulness or stability metric.** Report at least one metric of faithfulness (infidelity (38), comprehensiveness/sufficiency [93], ROAR/KAR accuracy drop [12], or insertion/deletion AUC [89]) and at least one metric of stability (max-sensitivity, empirical Lipschitz, or VarGrad-style variance). We recommend the Quantus toolkit [104] as a reference implementation supporting reproducible computation of these metrics.

**R10. Known failure modes that apply.** Map the chosen method onto Table XV and state which failure modes are plausibly active for the chosen setting. For example, an interventional Shapley method on a tabular dataset with correlated features should explicitly acknowledge the correlated-features failure mode and the value-function ambiguity.

### D. Compliance Examples

We illustrate the checklist with two short compliance examples; the purpose is to show how short the report can be when the methodological choices are unambiguous.

*a) Example 1: TreeSHAP on a credit-risk model.: R1*: gradient-boosted tree ensemble, target = probability of default for input record $x$; local attribution. *R2*: features = the 14 tabular columns of the credit dataset. *R3*: interventional baseline distribution = empirical training distribution, 1000 reference samples. *R4–R5*: interventional value function with TreeSHAP exact recursion. *R6*: satisfies efficiency, symmetry, dummy, additivity, local accuracy, missingness, consistency (see Table IX). *R7*: exact computation; no sampling error. *R8*: model-randomization sanity check passes (cosine similarity $< 0.1$ to attribution under random weights). *R9*: infidelity = 0.012 at Gaussian noise $\sigma = 0.1$. *R10*: features "income" and "debt-to-income ratio" are correlated ($\rho = 0.71$); interventional Shapley under-attributes their joint effect (Table XV, row "correlated features").

*b) Example 2: Integrated Gradients on an image classifier.: R1*: ResNet-50 ImageNet classifier, target = pre-softmax logit for the predicted class. *R2*: pixel-level attribution. *R3*: blurred-input baseline (Gaussian blur, $\sigma = 30$ pixels). *R4–R5*: single-reference value function; straight-line path with $K = 200$ Riemann steps. *R6*: satisfies sensitivity-(a), sensitivity-(b), implementation invariance, completeness, linearity, symmetry (Result 4). *R7*: integration error empirically below 1% on test images relative to $K = 2000$ reference. *R8*: sanity-check pass under cascading randomization. *R9*: max-sensitivity = 0.18 at $r = 0.02$; insertion AUC = 0.74. *R10*: gradient saturation possible on inactive ReLU paths (Table XV); off-manifold artefacts at intermediate $\alpha$ on the blur path.

### E. Summary

The proposed compact, checkable checklist converts the assumptions encoded by attribution methods into a reproducible reporting practice. The checklist does not restrict which attribution methods may be used. It specifies the information needed for attribution results to support the conclusions drawn from them. It is offered to authors and reviewers as a proposal, not as a community-ratified standard.

## XIII. Scope and Boundary Conditions

The principal scope boundaries and their rationale keep the mathematical object of study fixed: local additive attribution operators for predictive models.

### A. Explicit Out-of-Scope

The following topics are adjacent to the survey but outside its main scope because they manipulate mathematical objects distinct from the local additive attribution operator.

- **Mechanistic interpretability.** Circuit-level explanations of neural network internals (induction heads, sparse autoencoders, polysemantic neurons) target a different question: "what does the model compute?" This differs from the question "what features mattered for this prediction?" These methods do not produce a per-feature attribution vector and are not comparable to the methods we survey within the frame of Sections II–III.
- **Concept-based and concept-bottleneck explanations.** TCAV [26], network dissection [27], and concept-bottleneck models replace input features with named human concepts. They are treated briefly in Section I as the natural contrast to feature attribution; a full survey of the concept literature is a separate project.
- **Counterfactual and example-based explanations.** The literature on counterfactual explanations ("what would have to change about $x$ for the prediction to flip?") and on example-based explanations (influence functions, prototypes, training-data attribution) addresses questions complementary to but distinct from local additive attribution. We mention FIDO-CA and counterfactual generators (Section VII) where they intersect the attribution literature; we do not survey the broader counterfactual research programme.
- **Global interpretability and rule extraction.** Methods that produce a global summary of a black-box model (decision lists, partial dependence plots, ALE plots, rule extraction [105]) target a different deliverable than per-prediction attribution. SAGE [53] is the closest member of the Shapley family to a global method and is covered for that reason; the broader global-interpretation literature is surveyed in [32], [37].
- **Causal explanation methods.** Methods grounded in structural causal models or Pearl's do-calculus (causal Shapley [24], do-attribution) are discussed in their connection to interventional value functions (Section IV) and as an open problem (Section XIV), but a full survey of causal attribution would require a separate exposition of causal-graph machinery beyond the scope of this paper.
- **Comprehensive LLM-specific attribution.** The LLM attribution literature has grown rapidly in 2023–2025; we discuss its open status in Section XIV. A full treatment would need to cover token-level vs. span-level attribution for autoregressive generation, in-context-learning attribution, sparse attention probes, and emerging mechanistic-interpretability tools for transformers.

### B. Topics Treated Briefly

The following topics receive a section or subsection in this paper and have dedicated literatures of their own.

- *Graph neural network attribution* (Section XI): permutation symmetry, edge-attribution, and graph-conditional value functions deserve a separate treatment.
- *Biomedical and high-stakes-domain applications* (Section XI): the domain-specific evaluation criteria are surveyed lightly; we focus on the general methodological consensus.
- *Evaluation in NLP* (Section IX): the ERASER [93] benchmark is the standard reference, and the broader debate on plausibility vs. faithfulness in NLP is ongoing; we summarize but do not extend the faithfulness debate.

### C. Interpretive Boundaries of the Taxonomy

The taxonomy is a conceptual comparison and does not provide an empirical ranking. The scoring of prior surveys in Table II depends on judgment about what counts as substantive coverage; for this reason, the scoring rule is stated explicitly and ambiguous cases are assigned in favour of the prior survey. Several attribution methods also have implementation-dependent variants. An axiom entry may hold for the canonical formulation but not for every software implementation or propagation rule. Claims about sanity checks, infidelity, robustness, and method disagreement are therefore synthesized as reported properties of the cited studies, with implementation-dependent cases marked separately from mathematical axioms.

### D. Threats to Validity of the Taxonomy

Four limitations qualify the conclusions that should be drawn from the taxonomy and its tables.

- *Corpus limitation.* The corpus prioritizes methodological and axiomatic relevance over exhaustive bibliometric coverage; counts and coverage claims should be read accordingly.
- *Axiom-classification uncertainty.* Several methods exist in multiple variants (LRP rules, DeepLIFT rules, CAM layers), so matrix entries depend on which variant and implementation is taken as canonical; the conditional marks and Appendix A record these dependencies, but borderline judgments remain.
- *Domain dependence.* Attribution behaviour differs across images, text, tabular data, graphs, and biological sequences; properties observed in one modality do not automatically transfer.
- *Evaluation non-equivalence.* Faithfulness, stability, sanity checks, and human usefulness measure different quantities (Section IX); the taxonomy does not assume that any one of them subsumes the others.

### E. Methodological Boundaries

The survey is structured and taxonomy-driven. Its corpus construction prioritizes papers that define attribution operators, establish axioms, evaluate attribution behaviour, or document failure modes.

- The axiom-by-method matrix (Table IX) encodes the authors' reading of each method's axiomatic profile; where a method's behaviour depends on architectural details or on a choice of rule (LRP rules, DeepLIFT rules), we mark the entry conditionally and refer to the cited paper.
- Several reductions reported in Section VIII as "Results" are paraphrases of theorems in the cited papers, with assumptions made explicit. We have tried to flag every assumption that is critical to the reduction.
- Empirical results we cite (e.g., the sanity-check failure of guided backpropagation) are reported as such, not as theorems. Where quantitative claims are made, they reflect the original authors' experimental setting, not a meta-analysis.

## XIV. Open Problems and Future Directions

We close with six directions in which the methods of Sections IV–VII are incomplete. Each is stated as a question that current theory answers partially or not at all.

### A. Causal Attribution

The clearest open problem is the rigorous integration of attribution methods with causal modelling. Janzing et al. [24] provided the first systematic formulation, but the conditions under which Shapley-style attribution admits a causal interpretation remain restrictive. Open questions include:

- How does attribution interact with confounding? An attribution method may attribute credit to feature $i$ that, in the underlying causal graph, is mediated by an unobserved confounder.
- Can attribution methods be characterized as estimating *some* causal estimand, such as average treatment effect, conditional ATE, or individualized treatment effect, under the model's implicit causal assumptions?
- For interventional Shapley, what is the relationship between the do-operator and the value function? Recent work suggests a tight correspondence, but a full theorem remains to be stated.

### B. Correlated Features and Manifold-Aware Attribution

The conditional vs. interventional value-function debate (Section IV) remains live. Manifold-aware methods [25], [39], [91] constrain perturbations or coalition completions to the data manifold, but they require an additional generative model whose own faithfulness is now a question. Two open problems:

- How should the additional generative model be trained, evaluated, and validated? Standard generative-model evaluation (FID, likelihood) is not obviously appropriate for an attribution sub-routine.
- Is there an axiomatic characterization of on-manifold Shapley analogous to Result 1? The Lundstrom-Razaviyayn-style characterization [16] of IG does not yet have a manifold-aware analogue.

### C. Uncertainty in Explanations

Attribution methods are deterministic functions of $(f, x)$. Yet the model itself is a sample from a learning algorithm, and the attribution should plausibly carry uncertainty propagated from the model's training process. Existing approaches include Bayesian-network attribution, ensemble disagreement, and posterior-distribution attribution, but the principled framework remains open. The natural question is whether the attribution should be a distribution over $\mathbb{R}^d$, and if so what its marginals should mean.

### D. Attribution for Large Language Models

LLM attribution is the most active open area in XAI today. Four specific difficulties:

- **What is the prediction?** LLMs produce sequences, not scalars. Attribution targets include per-token logits, sequence probabilities, and aggregate metrics like answer correctness; the choice matters.
- **What is the feature?** Sub-tokens, tokens, spans, sentences, documents, system prompts, and retrieved context are all candidate units. Hierarchical methods [101] extend cleanly here but require deciding the granularity in advance.
- **Computational cost.** A trillion-parameter forward pass makes IG-style $K$-step path integration prohibitively expensive; gradient methods are similarly costly at scale.
- **In-context learning attribution.** For few-shot LLM predictions, the explanation may need to attribute credit to specific demonstrations in the context. This is a new problem class without an obvious analogue in the methods we have surveyed.

### *E. Rigorous Benchmarks and Human-Aligned Axioms*

The benchmarks of Section IX are largely functionally-grounded and often disagree with one another. Two directions seem productive:

- **Application-grounded benchmarks.** ERASER [93] is one such; analogues are needed for vision (clinical image cohorts), biology (gene-perturbation prediction), and structured prediction (graph classification). The principal obstacle is the cost of expert annotation.
- **Human-aligned axioms.** The axioms surveyed in Section III are mathematically motivated. A complementary line of work would establish which axioms humans actually endorse for an explanation, and whether the satisfaction of those axioms predicts downstream utility.

### *F. Explanation Robustness*

Result 13 establishes that gradient-based attributions are intrinsically fragile. The constructive question is how to build attributions that are stable by design. Two directions:

- **Lipschitz-constrained attribution.** Enforce Lipschitz stability (Axiom 16) directly, either by smoothing the model or by smoothing the attribution. SmoothGrad does the latter in expectation; principled methods that enforce a Lipschitz constant explicitly would be a substantial advance.
- **Certifiable attribution.** Analogous to certifiable robustness for predictions, the question is whether one can certify that the attribution map has a bounded change under any $\|\delta\| \leq \epsilon$. Existing approaches based on randomized smoothing provide a partial answer, but tight bounds remain open.

### *G. Conclusion*

We began this survey by arguing that the mathematical objects manipulated by feature-attribution methods provide the appropriate basis for comparison. The intervening sections have substantiated that argument: the methods discussed here, from exact Shapley to Grad-CAM to LIME, can be situated in a common frame defined by a value function, a baseline distribution, a path or perturbation distribution, and a conservation rule. The axiomatic analysis (Section III, Table IX) and the failure-mode formalization (Section X, Table XV) make precise the conditions under which each method's attribution is well-defined, the questions it can and cannot answer, and the assumptions it requires of the underlying data and model.

The thesis stated in Section I, restated here as a closing principle:

> *There is no assumption-free feature-attribution method. Every local additive attribution method defines feature importance through choices about value functions, references, paths, perturbation distributions, or conservation rules. Trustworthy use therefore requires reporting a heatmap or ranking together with the assumptions under which the attribution was computed and interpreted.*

For practice, this means that both method papers and applied papers should state the assumptions behind the reported attribution: a new attribution method should state its value function, baseline, path, and conservation rule, and an applied paper that reports attributions without specifying these choices has reported an underdetermined quantity.

The resulting catalogue supports a more explicit style of attribution reporting: method comparisons should name the mathematical choices they hold fixed, and applied studies should state the choices on which their attributions depend.

## Reproducibility Statement

**Reproducible artefacts.** (i) The axiom-by-method matrix (Table IX) is derivable from the per-cell justifications in Appendix A; each cell references either the original method paper, an axiomatic characterization, or a stated conditional assumption. A reader can verify each cell against the cited paper without further implementation. (ii) The empirical-evidence companion (Table X) cites the published sources for each sanity-check outcome; these external empirical claims were not generated in this survey. (iii) The checklist crosswalk in Appendix B records the source literature that motivates each reporting item and shows how the items apply across method families.

**Supplementary materials.** The supplementary material includes a machine-readable reporting checklist, a review-corpus summary, an axiom-matrix CSV, and a CSV recording the scoring rationale for Table II. These materials provide stable artefacts for checking the taxonomy, the proposed reporting checklist, and the survey-positioning table.

## Acknowledgments

The authors thank the maintainers of the open-source attribution libraries (Captum, SHAP, iNNvestigate, Quantus, Grad-CAM++) whose implementations underlie much of the empirical XAI literature.

## Appendix A
## Per-Cell Justification of the Axiom Matrix

This appendix supplies, for each row of Table IX, a short justification or citation for each non-trivial axiom claim. We use the following abbreviations: *Cmp* (Completeness), *Sa/Sb* (Sensitivity-(a)/(b)), *II* (Implementation Invariance), *Cns* (Consistency), *Mono* (Monotonicity), *Sym* (Symmetry-preservation), *Lin* (Linearity), and *Cont* (Continuity in $x$). Entries in the matrix marked ✓ are justified here unconditionally; entries marked ∘ are conditional, with the condition stated.

*Shapley Family*

**Exact Shapley.** Cmp, Sym, dummy and Lin are Axioms 1–4; satisfied by construction via [44]. Cns is a direct consequence of monotone marginal contributions. II follows from $v$ depending only on the input–output function of $f$. Cont in $x$ holds for any continuous value function $v$.

**KernelSHAP [2].** In the limit of infinite samples, equals exact Shapley (Result 8); inherits its axioms. Cns and Mono are conditional (∘) because finite-sample KernelSHAP can exhibit variance-induced violations, documented in [20], [52].

**TreeSHAP [45].** Inherits all Shapley axioms exactly; Cns is provably satisfied (Theorem 1 of [45]). Cont is conditional because TreeSHAP attributions are piecewise constant in $x$.

**DeepSHAP [2].** Cmp is approximate (one forward+backward pass; full Shapley requires multiple). Sa and Cns are conditional (∘) and II is conditional because both depend on the DeepLIFT propagation rule used (Rescale vs. RevealCancel).

**GradientSHAP [38].** Equation (12): the inner expectation over $\alpha$ is an IG operator, the outer expectation over $x'$ is a sampling step. Inherits Cmp, Sa, Sb, II, Lin, Sym from IG. Cns conditional on sampling adequacy.

*Path Family*

**Integrated Gradients [1].** Cmp by fundamental theorem of calculus (Result 4); Sa, Sb, II, Lin, Sym by Theorem 1 of the cited paper. Cont follows from the integrability of the gradient.

**Expected Gradients [38].** Same as IG, with the baseline replaced by an expectation. Cmp, Sa, Sb, II, Lin, Sym all preserved by linearity of the expectation operator.

**Guided IG [41].** Cmp preserved because the integration is still over a path. II is conditional because the path depends on the gradient magnitude, which can break strict invariance under network reparameterization. Sym is not preserved.

**Blur IG [42].** Cmp on the scale-space path; inherits II, Lin from IG. Sym fails because the path is not permutation-equivariant.

**Integrated Hessians [56].** Cmp over feature pairs; inherits II, Sym, Lin from the IG construction. Sa, Sb at the interaction level.

*Gradient and Backpropagation Family*

**Saliency [62].** Sb by definition ($\phi_i = 0$ if $\partial f/\partial x_i \equiv 0$). II by invariance of partial derivatives. Cmp fails (sum of gradients $\neq \Delta f$). Cont conditional (∘) due to ReLU non-smoothness.

**Gradient × Input.** Cmp is satisfied only when $f$ is linear (∘). II by invariance of $\nabla f$. Sa conditional on nonzero gradient.

**SmoothGrad [64].** Smoothing preserves II and Sb; gives Cont by construction (averaging Gaussian kernel). Cmp not satisfied.

**Guided BP [65] / Deconvnet [66].** Do not satisfy II because the rule-modified backward pass is non-standard. Their model-randomization behaviour is reported separately in Table X.

**DeepLIFT [6].** Cmp by summation-to-delta (Axiom 15). Sa, Sb, Sym, Lin by construction. II conditional on Rescale rule and architecture.

**LRP [5].** Cmp by conservation (Axiom 14). Sa, Sb conditional on the propagation rule. II conditional on standard rules and absence of bias terms.

**FullGrad [68].** Cmp by explicit bias accounting. Sa, Sb, II, Cont by the smooth differentiability of the total gradient plus biases.

**Grad-CAM [4].** Cmp not satisfied at the pixel level. II conditional on the architecture having global average pooling.

**HiResCAM [75].** Cmp at the layer level (∘). Sa, Sb by direct gradient application (no pooling). II is conditional (∘): the map is defined on a chosen internal layer, and functionally equivalent networks need not share internal representations.

**Score-CAM [71].** Does not require gradients, but the weights are computed from selected activation maps of a target layer, so II is conditional (∘): strict implementation invariance holds only for methods that depend on input–output queries alone. Sa by construction (perturbation-based).

*Perturbation Family*

**Occlusion [66].** Sa for any patch with nonzero marginal contribution. II by black-box queries. Sym by permutation-equivariance of patch sliding. Cmp fails (single coalition, not exhaustive).

**LIME [3].** Cmp conditional (∘) on the kernel; satisfied only when the kernel is the Shapley kernel (10). II by black-box queries. Cont by smoothness of the local linear fit.

**Anchors [86].** Anchors output rules, not real-valued attributions; Cmp, Lin, Cont are N/A. Sa and Sb apply analogously (rule-based).

**Meaningful Perturbations [87].** Sa via optimization objective. II by black-box queries. Cont conditional on the regularization $\lambda$.

**RISE [89].** Sa by construction. Sym by mask distribution being permutation-symmetric. II by black-box queries.

*Notes on Conditional Entries*

The ∘ marks in Table IX signal that the axiom holds under specific conditions: choice of rule (LRP, DeepLIFT), architectural restrictions (Grad-CAM family, DeepLIFT), sample adequacy (KernelSHAP, GradientSHAP), or method variant (CAM family). Readers extracting machine-readable claims from the matrix should consult the per-method paragraph above for the precise qualification.

*Interpretation of the Matrix*

The matrix entries above summarize mathematical claims and implementation-dependent empirical statuses reported in the cited papers. Where a method's behaviour depends on a configuration (e.g., LRP-$\epsilon$ vs. LRP-composition), the matrix records the most commonly used configuration as the default and the alternative as a ∘.

## Appendix B
## Checklist Crosswalk for the Analytic Corpus

This appendix connects the proposed reporting checklist of Section XII to the source literature used throughout the survey. The goal is to show how each checklist item follows from established attribution practice, axiomatic characterization, or documented failure modes. The crosswalk also provides a compact guide for applying the checklist across method families.

*Item-Level Source Anchors*

Table XVIII records, for each item R1–R10, the specification target, representative source anchors in the analytic corpus, and the item's role in the checklist. Each anchor is a paper that either motivated the item or documented the consequence of leaving it unreported.

*Cross-Family Reporting Profile*

Table XIX groups the same ten items by method family, separating the items that are usually necessary to make the attribution mathematically specified from the items that support its empirical use.

*Use of the Crosswalk*

The crosswalk gives the reporting checklist a literature-backed interpretation. For a new attribution study, the relevant method family in Table XIX identifies the items that must be specified before the attribution is mathematically defined. The source anchors in Table XVIII then identify the literature that explains why each item matters. For example, a KernelSHAP study requires the scalar output (R1), feature grouping (R2), the interventional or conditional value function (R4), the coalition sampling design (R5), the approximation budget (R7), and the correlated-feature failure mode (R10). An Integrated Gradients study requires the scalar output (R1), baseline (R3), path and integration budget (R5, R7), axioms used (R6), sanity-check behaviour (R8), and baseline/path sensitivity (R10).

## References


[1] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. ICML*, ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 3319–3328.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, vol. 30, 2017.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. ICCV*, 2017, pp. 618–626.

[5] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, 2015.

[6] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. ICML*, 2017.

[7] S. Krishna, T. Han, A. Gu, S. Wu, S. Jabbari, and H. Lakkaraju, "The disagreement problem in explainable machine learning: A practitioner's perspective," *Transactions on Machine Learning Research*, 2024.

[8] B. Bilodeau, N. Jaques, P. W. Koh, and B. Kim, "Impossibility theorems for feature attribution," *Proceedings of the National Academy of Sciences*, vol. 121, no. 2, 2024, art. no. e2319169121.

[9] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. NeurIPS*, vol. 31, 2018, pp. 9525–9536.

[10] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (un)reliability of saliency methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham, Switzerland: Springer, 2019, pp. 267–280.

[11] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, "Sanity checks for saliency metrics," in *Proc. AAAI*, vol. 34, no. 4, 2020, pp. 6021–6029.

[12] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in *Proc. NeurIPS*, vol. 32, 2019.

[13] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel, "Explanations can be manipulated and geometry is to blame," in *Proc. NeurIPS*, vol. 32, 2019, pp. 13 567–13 578.

[14] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proc. AAAI*, 2019.

[15] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post-hoc explanation methods," in *Proc. AAAI/ACM Conf. on AI, Ethics, and Society (AIES)*, 2020, pp. 180–186.

[16] D. Lundstrom and M. Razaviyayn, "Four axiomatic characterizations of the integrated gradients attribution method," *Journal of Machine Learning Research*, vol. 26, no. 177, pp. 1–31, 2025.

[17] M. Sundararajan and A. Najmi, "The many Shapley values for model explanation," in *Proc. ICML*, 2020.

[18] I. E. Kumar, C. Scheidegger, S. Venkatasubramanian, and S. A. Friedler, "Problems with Shapley-value-based explanations as feature importance measures," in *Proc. ICML*, 2020.

[19] T. Han, S. Srinivas, and H. Lakkaraju, "Which explanation should i choose? A function approximation perspective to characterizing post hoc explanations," in *Proc. NeurIPS*, 2022.

[20] I. Covert and S.-I. Lee, "Improving KernelSHAP: Practical Shapley value estimation using linear regression," in *Proc. AISTATS*, 2021.

[21] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *Proc. ICLR*, 2018.

[22] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.

[23] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv:1702.08608*, 2017.

[24] D. Janzing, L. Minorics, and P. Blöbaum, "Feature relevance quantification in explainable AI: A causal problem," in *Proc. AISTATS*, ser. Proceedings of Machine Learning Research, vol. 108, 2020, pp. 2907–2916.

TABLE XVIII
SOURCE ANCHORS FOR THE REPORTING CHECKLIST ACROSS THE ANALYTIC CORPUS. EACH ROW IDENTIFIES REPRESENTATIVE PAPERS THAT MOTIVATE THE CORRESPONDING CHECKLIST ITEM.

| Item | Specification target | Representative source anchors | Role in the standard |
|---|---|---|---|
| R1 | Model and scalar output | Saliency [62], Grad-CAM [4], ERASER [93] | Fixes whether the attribution explains a logit, probability, loss, or task-specific score. |
| R2 | Feature unit and display transform | LIME [3], TCAV [26], hierarchical attribution [101], transformer attribution [81] | Prevents direct comparison of pixels, superpixels, tokens, spans, concepts, and grouped variables. |
| R3 | Baseline or reference distribution | Integrated Gradients [1], DeepLIFT [6], baseline studies [59] | Defines the counterfactual reference point against which feature contributions are measured. |
| R4 | Value function or perturbation distribution | Shapley values [44], SHAP [2], conditional Shapley [39], manifold Shapley [25], causal attribution [24], QII [51] | Specifies feature absence, locality, and whether correlated features are treated interventionally or conditionally. |
| R5 | Path, coalition strategy, or sampling design | IG [1], Guided IG [41], KernelSHAP [2], [20], RISE [89] | Determines numerical approximation, sampling variance, and the applicable axiom guarantees. |
| R6 | Axioms satisfied and not satisfied | Shapley [44], Banzhaf [46], Aumann-Shapley [48], IG axioms [1], IG characterization [16], impossibility results [8] | States which mathematical guarantees the attribution can legitimately support. |
| R7 | Approximation budget and stochasticity | SmoothGrad [64], KernelSHAP analysis [20], Shapley algorithms [52], RISE [89] | Makes sampled, noisy, or discretized attributions reproducible. |
| R8 | Sanity-check behaviour | Sanity checks [9], sanity-check extensions [11], HiResCAM [75] | Tests whether the attribution depends on learned model parameters and training labels. |
| R9 | Faithfulness or stability metric | Infidelity [95], ROAR [12], ERASER [93], RISE metrics [89], robustness [49], shortcut tests [98], Quantus [104] | Connects attribution values to model behaviour under controlled perturbation or stability tests. |
| R10 | Known failure modes | Input invariance [10], fragility [14], manipulation [13], fooling attacks [15], disagreement [7], explanation constraints [96], [97] | Links each attribution claim to the failure modes most relevant to its method family and data setting. |

TABLE XIX
CHECKLIST EMPHASIS BY METHOD FAMILY. "PRIMARY" ITEMS ARE USUALLY NECESSARY TO MAKE THE ATTRIBUTION MATHEMATICALLY SPECIFIED; "EVIDENCE" ITEMS SUPPORT EMPIRICAL USE OF THE ATTRIBUTION.

| Method family | Primary specification items | Evidence items | Representative sources |
|---|---|---|---|
| Shapley and value-function methods | R1–R5, especially R4 for conditional/interventional choice | R6, R7, R9, R10 | [2], [20], [24], [25], [39], [44], [45], [50], [51] |
| Path-integral methods | R1–R3 and R5 for baseline and path | R6, R7, R8, R10 | [1], [16], [38], [41], [42] |
| Gradient and backpropagation methods | R1, R2, R6 and the propagation rule | R8, R9, R10 | [5], [6], [21], [62], [65], [68] |
| CAM-style spatial methods | R1, R2 and target layer / projection rule | R8, R9, R10 | [4], [69]–[71], [75], [76] |
| Perturbation and surrogate methods | R1, R2, R4 and R5 for perturbation design | R7, R9, R10 | [3], [43], [83], [86], [87], [89], [91] |
| Transformer and NLP attribution | R1 and R2 for scalar target and token/span unit | R8, R9, R10 | [77], [78], [80], [81], [93], [94], [98], [100], [102] |


[25] C. Frye, D. de Mijolla, T. Begley, L. Cowton, M. Stanley, and I. Feige, "Shapley explainability on the data manifold," in *Proc. ICLR*, 2021.

[26] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viégas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. ICML*, 2018.

[27] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proc. CVPR*, 2017.

[28] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, 2018.

[29] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[30] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE DSAA*, 2018.

[31] C. Molnar, *Interpretable Machine Learning*, 2nd ed. Independently published, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[32] R. Saleem, B. Yuan, F. Kurugollu, A. Anjum, and L. Liu, "Explaining deep neural networks: A survey on the global interpretation methods," *Neurocomputing*, vol. 513, pp. 165–180, 2022.

[33] Y. Wang, T. Zhang, X. Guo, and Z. Shen, "Gradient based feature attribution in explainable AI: A technical review," *arXiv:2403.10415*, 2024.

[34] A. Cremades, S. Hoyas, and R. Vinuesa, "Additive-feature-attribution methods: A review on explainable artificial intelligence for fluid dynamics and heat transfer," *arXiv:2409.11992*,

2024.
[35] M. Li, H. Sun, Y. Huang, and H. Chen, "Shapley value: From cooperative game to explainable artificial intelligence," *Autonomous Intelligent Systems*, vol. 4, no. 1, 2024, art. no. 2.
[36] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI," *ACM Computing Surveys*, vol. 55, no. 13s, 2023, art. no. 295.
[37] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2021.
[38] G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, and S.-I. Lee, "Improving performance of deep learning models with axiomatic attribution priors and expected gradients," *Nature Machine Intelligence*, vol. 3, pp. 620–631, 2021.
[39] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values," *Artificial Intelligence*, vol. 298, 2021.
[40] L. Merrick and A. Taly, "The explanation game: Explaining machine learning models using Shapley values," in *Proc. CD-MAKE*, 2020.
[41] A. Kapishnikov, S. Venugopalan, B. Avci, B. Wedin, M. Terry, and T. Bolukbasi, "Guided integrated gradients: An adaptive path method for removing noise," in *Proc. CVPR*, 2021.
[42] S. Xu, S. Venugopalan, and M. Sundararajan, "Attribution in scale and space," in *Proc. CVPR*, 2020.
[43] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," in *Proc. ICLR*, 2019.
[44] L. S. Shapley, "A value for n-person games," in *Contributions to the Theory of Games, Volume II*, ser. Annals of Mathematics Studies, H. W. Kuhn and A. W. Tucker, Eds. Princeton University Press, 1953, no. 28, pp. 307–317.
[45] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
[46] J. F. Banzhaf, "Weighted voting doesn't work: A mathematical analysis," *Rutgers Law Review*, vol. 19, pp. 317–343, 1965.
[47] G. Owen, "Multilinear extensions of games," *Management Science*, vol. 18, no. 5, pp. P64–P79, 1972.
[48] R. J. Aumann and L. S. Shapley, *Values of Non-Atomic Games*. Princeton University Press, 1974.
[49] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," in *ICML Workshop on Human Interpretability*, 2018.
[50] E. Štrumbelj and I. Kononenko, "An efficient explanation of individual classifications using game theory," *Journal of Machine Learning Research*, vol. 11, pp. 1–18, 2010.
[51] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *Proc. IEEE Symposium on Security and Privacy*, 2016, pp. 598–617.
[52] H. Chen, I. C. Covert, S. M. Lundberg, and S.-I. Lee, "Algorithms to estimate Shapley value feature attributions," *Nature Machine Intelligence*, vol. 5, pp. 590–601, 2023.
[53] I. Covert, S. Lundberg, and S.-I. Lee, "Understanding global feature contributions with additive importance measures," in *Proc. NeurIPS*, 2020.
[54] D. Fryer, I. Strümke, and H. Nguyen, "Shapley values for feature selection: The good, the bad, and the axioms," *arXiv:2102.10936*, 2021.
[55] K. Dhamdhere, A. Agarwal, and M. Sundararajan, "The Shapley-Taylor interaction index," in *Proc. ICML*, 2020.
[56] J. D. Janizek, P. Sturmfels, and S.-I. Lee, "Explaining explanations: Axiomatic feature interactions for deep networks," *Journal of Machine Learning Research*, vol. 22, no. 104, pp. 1–54, 2021.
[57] M. Tsang, D. Cheng, and Y. Liu, "Detecting statistical interactions from neural network weights," in *Proc. ICLR*, 2018.
[58] A. Ghorbani and J. Zou, "Neuron Shapley: Discovering the responsible neurons," in *Proc. NeurIPS*, 2020.
[59] P. Sturmfels, S. Lundberg, and S.-I. Lee, "Visualizing the impact of feature attribution baselines," *Distill*, 2020.
[60] A. Kapishnikov, T. Bolukbasi, F. Viégas, and M. Terry, "XRAI: Better attributions through regions," in *Proc. ICCV*, 2019.
[61] K. Dhamdhere, M. Sundararajan, and Q. Yan, "How important is a neuron?" in *Proc. ICLR*, 2019.
[62] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv:1312.6034*, 2013.
[63] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, 2010.
[64] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," *arXiv:1706.03825*, 2017.
[65] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR Workshop*, 2015.
[66] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.
[67] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.
[68] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," in *Proc. NeurIPS*, 2019.
[69] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, 2016.
[70] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Improved visual explanations for deep convolutional networks," in *Proc. WACV*, 2018, pp. 839–847.
[71] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. CVPR Workshops*, 2020.
[72] S. Desai and H. G. Ramaswamy, "Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization," in *Proc. WACV*, 2020.
[73] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Trans. Image Processing*, vol. 30, pp. 5875–5888, 2021.
[74] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," *arXiv:2008.00299*, 2020.
[75] R. L. Draelos and L. Carin, "Use HiResCAM instead of Grad-CAM for faithful explanations of CNNs," *arXiv:2011.08891*, 2020.
[76] Q. Zheng, Z. Wang, J. Zhou, and J. Lu, "Shap-CAM: Visual explanations for convolutional neural networks based on Shapley value," in *Proc. ECCV*, 2022.
[77] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proc. NAACL*, 2019.
[78] S. Serrano and N. A. Smith, "Is attention interpretable?" in *Proc. ACL*, 2019.
[79] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proc. EMNLP*, 2019.
[80] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proc. ACL*, 2020.
[81] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. CVPR*, 2021.
[82] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," in *Proc. ICLR*, 2017.
[83] G. Plumb, D. Molitor, and A. S. Talwalkar, "Model agnostic supervised local explanations," in *Proc. NeurIPS*, 2018.
[84] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[85] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[86] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI*, 2018.
[87] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. ICCV*, 2017.

[88] R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *Proc. ICCV*, 2019.
[89] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in *Proc. BMVC*, 2018.
[90] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," in *Proc. NeurIPS*, 2017.
[91] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," in *Proc. ICLR*, 2020.
[92] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, 2017.
[93] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "ERASER: A benchmark to evaluate rationalized NLP models," in *Proc. ACL*, 2020.
[94] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" in *Proc. ACL*, 2020.
[95] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. Ravikumar, "On the (in)fidelity and sensitivity of explanations," in *Proc. NeurIPS*, 2019.
[96] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proc. IJCAI*, 2017.
[97] L. Rieger, C. Singh, W. J. Murdoch, and B. Yu, "Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge," in *Proc. ICML*, 2020.
[98] J. Bastings, S. Ebert, P. Zablotskaia, A. Sandholm, and K. Filippova, ""will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification," *arXiv:2111.07367*, 2022.
[99] D. Balduzzi, M. Frean, L. Leary, J. P. Lewis, K. W.-D. Ma, and B. McWilliams, "The shattered gradients problem: If resnets are the answer, then what is the question?" in *Proc. ICML*, ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 342–350.
[100] L. Arras, G. Montavon, K.-R. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," in *EMNLP WASSA Workshop*, 2017.
[101] C. Singh, W. J. Murdoch, and B. Yu, "Hierarchical interpretations for neural network predictions," in *Proc. ICLR*, 2019.
[102] W. J. Murdoch, P. J. Liu, and B. Yu, "Beyond word importance: Contextual decomposition to extract interactions from LSTMs," in *Proc. ICLR*, 2018.
[103] M. Tsang, S. Rambhatla, and Y. Liu, "How does this interaction affect me? Interpretable attribution for feature interactions," in *Proc. NeurIPS*, 2020.
[104] A. Hedström, L. Weber, D. Bareeva, D. Krakowczyk, F. Motzkus, W. Samek, S. Lapuschkin, and M. M.-C. Höhne, "Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond," *Journal of Machine Learning Research*, vol. 24, no. 34, pp. 1–11, 2023.
[105] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proc. ACM SIGKDD*, 2016.