PULLSHAP: PRINCIPLED ATTRIBUTION PULLBACK THROUGH LEARNED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned representations create an attribution mismatch: predictions are made after transforming the original features, while explanations are often required in the original feature space. We study when Shapley attribution can be pulled back through a factorization $F = f \circ p$. We first show that no generic local chain rule exists: two smooth representations can agree on neighborhoods of both the baseline and explicand, and hence have identical endpoint derivatives of every finite order, yet induce different composite input-feature Shapley values. Moreover, representation compression alone is insufficient; even a smooth scalar representation can realize any normalized coalition game on the relevant hybrid vertices. We then establish sufficient conditions for pullability. Mixed cross-feature curvature of the representation controls its deviation from an additive coalition surrogate, showing that nonlinearity alone need not cause representation-side coalition distortion. On this surrogate, a distinct obstruction is the discrepancy between Aumann–Shapley path allocation and Shapley–Shubik allocation. These results lead to PULLSHAP, which combines integrated representation contributions with integrated downstream sensitivity, together with an error decomposition separating representation distortion, downstream allocation mismatch, and numerical approximation.

Controlled experiments confirm the predicted exact and failure regimes. Across 1,250 exact attribution instances nested within successful learned-representation runs, median relative-ℓ_1 error is 0.60%, with a non-negligible upper tail. Within dense MLP encoder tracks across three datasets, representation-side distortion is positively associated with exact attribution error. A separate exact-reference benchmark compares the differentiable approximation with established SHAP estimators while keeping black-box model-evaluation budgets distinct from gradient/Jacobian computation.

1 INTRODUCTION

Many modern predictors factor as $F = f \circ p$, where $p : \mathbb{R}^d \to \mathbb{R}^r$ is a learned representation and f predicts from the latent state. Yet explanations are often required in the original input coordinates. This creates a fundamental attribution problem: an explanation constructed through the representation need not recover the Shapley values of the composite model in input space. We study this problem as *Shapley attribution pullback*: when, and with what error, can attribution through a learned representation be pulled back to the input-feature Shapley values of $F = f \circ p$? Prior work has developed Shapley-style propagation through composed and graphical models and latent-to-input attribution mappings (Chen et al., 2022; Wang et al., 2021; Singal et al., 2021; Hu et al., 2024). Our focus is different: we characterize the structural conditions governing recovery of a fixed input-space Shapley target through the representation, and quantify the error when exact recovery fails.

For a baseline x^0, those Shapley values depend on $F(x_S)$ over hybrid coalitions S, not only on endpoint behavior. We first establish two boundary results that delimit what local or compressed representations can guarantee. First, two smooth representations can agree on neighborhoods of both x^0 and x, and hence have identical endpoint derivatives of every finite order, yet induce different composite Shapley vectors (T1). Second, even a scalar representation can realize any normalized set function on the relevant hybrid vertices (T2). Thus, neither endpoint-local differential information nor a low-dimensional latent space is sufficient, by itself, to guarantee exact pullback. Since neither controls the coalition game, we next ask which structural property of the representation does.

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

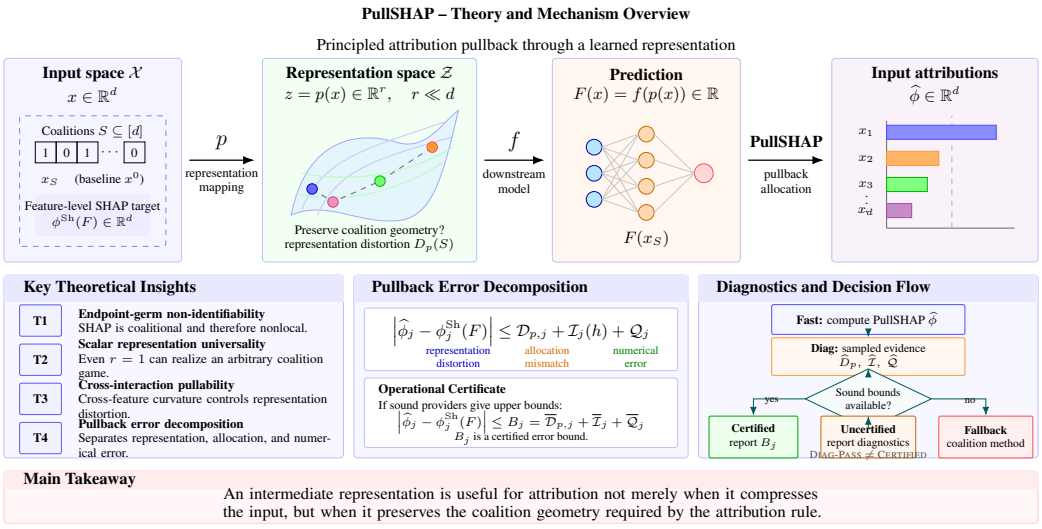

Figure 1: **PullSHAP theory and mechanism overview.** Exact input-feature SHAP is coalitional, so neither endpoint-local information nor low representation dimension yields a generic pullback (T1–T2). Pullability instead depends on representation-side coalition distortion controlled by cross-feature interaction (T3) and downstream Aumann–Shapley versus Shapley–Shubik allocation mismatch. T4 combines these effects with numerical approximation error. Diagnostics provide sampled evidence but become certificates only when sound upper bounds on all three terms are supplied.

We then identify the positive structure. Integrated encoder Jacobians define complete feature-wise representation atoms and an additive latent surrogate. Mixed cross-partials of the encoder along the baseline-to-input box control the surrogate's coalition distortion (T3): nonlinear but feature-separable representations have zero representation-side coalition distortion. A second obstruction is downstream. Even when the encoder preserves the additive coalition surrogate exactly, path allocation through that surrogate need not equal its discrete Shapley allocation. On the additive surrogate, our path allocation is Aumann–Shapley (AS), whereas surrogate SHAP is Shapley–Shubik (SS); their classical coincidence on multilinear-plus-additively-separable functions (Sun & Sundararajan, 2011) yields an exact regime.

These results lead to PULLSHAP and the organizing bound

$$|\widehat{\phi}_j - \phi_j^{\mathrm{Sh}}(F)| \le \underbrace{\mathcal{D}_{p,j}}_{\text{representation}} + \underbrace{\mathcal{I}_j(h)}_{\text{allocation}} + \underbrace{\mathcal{Q}_j}_{\text{numerics}} . \tag{1}$$

Theorem 4 formalizes this decomposition. Corollary 5 gives a conditional certificate template: it becomes operational only when sound upper bounds for all three terms are available; we do not claim generic inexpensive providers. Controlled experiments validate exact and failure regimes and isolate the two structural mechanisms. On learned representations across three datasets, representation-side distortion is positively associated with exact attribution error within the nonlinear-encoder tracks. A targeted exact-reference benchmark then places the differentiable approximation alongside established SHAP estimators while keeping black-box coalition/model-row budgets distinct from gradient/Jacobian computation.

Contributions. Our primary contributions can be given as follows: (i) a mixed-curvature pullability bound showing that cross-feature interaction bounds representation-side coalition distortion, while nonlinearity alone need not cause it; and (ii) PULLSHAP with a target-relative error decomposition separating representation distortion, AS–SS allocation mismatch, and numerical error, together with a conditional certificate template when sound upper bounds are available. Two boundary results—endpoint-germ non-identifiability and scalar-representation universality—show why neither endpoint-local information nor low latent dimension alone can guarantee exact pullback. Finally, controlled and learned-representation mechanism-validation experiments test the two structural error

mechanisms separately, and a targeted exact-reference benchmark evaluates practical approximations while keeping black-box coalition/model-row budgets distinct from gradient/Jacobian computation.

2 Problem Setup

Let $p : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}^r$, $f : \mathbb{R}^r \to \mathbb{R}$, and $F = f \circ p$. For an explicand $x \in \mathcal{X}$ and a baseline $x^0 \in \mathcal{X}$, assume that the axis-aligned box between x^0 and x is contained in $\mathcal{X}$. Let $[d] = \{1, \dots, d\}$, write $\Delta x = x - x^0$, and, for $S \subseteq [d]$, define

$$x_S = x^0 + \mathbf{1}_S \odot \Delta x, \qquad v(S) = F(x_S) - F(x^0). \tag{2}$$

Our target is the baseline Shapley value

$$\phi_j^{\mathrm{Sh}}(F; x, x^0) = \mathbb{E}_{S \sim Q_j^{\mathrm{Sh}}}\big[v(S \cup \{j\}) - v(S)\big], \tag{3}$$

where Q_j^{Sh} is the distribution over $S \subseteq [d] \setminus \{j\}$ induced by the set of features preceding j in a uniformly random permutation of $[d]$. This is the *composite input-space* game: a latent attribution is not itself the target. We study single-reference baseline (hybrid-coalition) semantics; conditional SHAP is outside the present theory.

3 Why Pullback Is Not Generic

The factorization $F = f \circ p$ exposes internal structure but does not localize the coalition target.

Theorem 1 (Endpoint-Germ Non-Identifiability)**.** *Let at least two coordinates of $x - x^0$ be nonzero. There exist smooth scalar representations p_0, p_1 and a smooth f such that $p_0 = p_1$ on neighborhoods of both x^0 and x, but*

$$\phi^{\mathrm{Sh}}(f \circ p_0; x, x^0) \neq \phi^{\mathrm{Sh}}(f \circ p_1; x, x^0).$$

Consequently, no universal operator depending only on the endpoint germs of p can recover exact composite baseline SHAP for all smooth representations.

Proof idea. Restrict to the active features, choose a proper hybrid vertex of the active-coordinate cube distinct from both endpoints, and place a smooth bump in a sufficiently small neighborhood containing no other active hybrid vertex and disjoint from neighborhoods of the endpoints. The representations therefore have identical endpoint germs, while the coalition game is perturbed at a single proper active coalition; this perturbation has a nonzero Shapley vector. Appendix A gives the full proof. Thus *SHAP pullback is nonlocal because its target is coalitional*.

Compression alone does not remove this obstruction.

Theorem 2 (Scalar Representation Universality)**.** *Suppose $x_j \neq x_j^0$ for all j. For every normalized set function $v : 2^{[d]} \to \mathbb{R}$ with $v(\varnothing) = 0$, there exist a smooth scalar representation $p : \mathbb{R}^d \to \mathbb{R}$ and $f(z) = z$ such that*

$$(f \circ p)(x_S) - (f \circ p)(x^0) = v(S) \qquad \textit{for every } S \subseteq [d].$$

Proof idea. With $u_j(\xi) = \frac{\xi_j - x_j^0}{x_j - x_j^0}$, let $P_v(u) = \sum_{T \subseteq [d]} \mu_v(T) \prod_{j \in T} u_j$ be the Möbius-form multilinear extension of v. Since $P_v(\mathbf{1}_S) = v(S)$ by Möbius inversion, $p(\xi) = P_v(u(\xi))$ realizes the game with $r = 1$. This is a structural expressivity statement, not an oracle lower bound: low latent dimension alone does not restrict the coalition games realizable on the hybrid vertices and therefore cannot, by itself, guarantee exact pullback.

4 Cross-Interaction Pullability

Normalize the baseline–explicand box by $q(u) = p(x^0 + D_{\Delta x} u)$, where $u \in [0, 1]^d$, and assume that p is C^2 on this box. Fix any norm on $\mathbb{R}^r$. Define the complete representation atoms

$$a_j = \int_0^1 \partial_j q(t\mathbf{1})\, dt = \Delta x_j \int_0^1 \partial_j p(x^0 + t\Delta x)\, dt, \qquad \sum_j a_j = p(x) - p(x^0), \tag{4}$$

and the additive coalition surrogate

$$\widetilde{q}(S) = q(0) + \sum_{j \in S} a_j, \qquad D_p(S) = \|q(\mathbf{1}_S) - \widetilde{q}(S)\|.$$

For $i \neq j$, define

$$\kappa_{ij} = \sup_{u \in [0,1]^d} \|\partial_{ij}^2 q(u)\|, \qquad \mathcal{K}_p = \sum_{i<j} \kappa_{ij}, \tag{5}$$

and, for a coalition S,

$$K_{\text{int}}(S) = \sum_{\substack{i<j \\ i,j \in S}} \kappa_{ij}, \qquad K_{\text{cross}}(S) = \sum_{\substack{i \in S \\ j \notin S}} \kappa_{ij}.$$

Theorem 3 (Cross-Interaction Pullability)**.** *For every $S \subseteq [d]$,*

$$D_p(S) \leq 2K_{\text{int}}(S) + \tfrac{1}{2} K_{\text{cross}}(S). \tag{6}$$

Consequently,

$$\max_{S \subseteq [d]} D_p(S) \leq 2\mathcal{K}_p. \tag{7}$$

The result holds for vector-valued p under any fixed norm and requires neither invertibility nor full rank.

Proof idea. Anchored single-feature atoms reconstruct each coalition up to its internal mixed curvature, while comparison with the complete diagonal atoms contributes the crossing term. Summing the resulting pairwise bounds gives Equation 6; Appendix B gives the full proof. In particular, if all mixed partials vanish, then q is additively separable across input coordinates on the box: there exist vector-valued univariate functions g_j with $g_j(0) = 0$ such that $q(u) = q(0) + \sum_{j=1}^{d} g_j(u_j)$. Consequently, $D_p(S) = 0$ for every coalition. Thus nonlinearity itself is not the obstruction; cross-feature interaction is.

Downstream allocation. Let $z^0 = p(x^0)$ and $\Delta z = p(x) - p(x^0) = \sum_j a_j$, and define

$$h(u) = f\left(z^0 + \sum_j u_j a_j\right). \tag{8}$$

The path allocation of this surrogate is

$$\phi_j^{\text{AS}}(h) = \int_0^1 \partial_j h(t\mathbf{1})\, dt = a_j^\top \int_0^1 \nabla f(z^0 + t\Delta z)\, dt =: \phi_j^{\text{Pull}}.$$

Exact SHAP of the surrogate coalition game is instead its Shapley–Shubik allocation $\phi_j^{\text{SS}}(h)$. Define

$$\mathcal{I}_j(h) = \left|\phi_j^{\text{AS}}(h) - \phi_j^{\text{SS}}(h)\right|. \tag{9}$$

If $D_p(S) = 0$ for every S, then the composite and surrogate games agree on all hybrid coalitions. By the classical AS–SS coincidence result (Sun & Sundararajan, 2011), if in addition h is multilinear plus additively separable, then

$$\phi^{\text{Pull}} = \phi^{\text{Sh}}(F; x, x^0).$$

Representation-side coalition distortion and downstream allocation mismatch are therefore distinct structural obstructions.

5 PullSHAP and Error Control

Let $z^0 = p(x^0)$ and $\Delta z = p(x) - p(x^0)$. PullSHAP separately averages representation transport and downstream sensitivity:

$$A_x = \int_0^1 J_p(x^0 + t\Delta x)\, dt\, D_{\Delta x}, \qquad \bar{g} = \int_0^1 \nabla f(z^0 + s\Delta z)\, ds, \qquad \phi^{\text{Pull}} = A_x^\top \bar{g}, \tag{10}$$

Algorithm 1 PULLSHAP-FAST

Require: Encoder p, downstream predictor f, explicand x, baseline x^0, quadrature orders K_p, K_f
Ensure: Input-feature attribution $\widehat{\phi} \in \mathbb{R}^d$

$\Delta x \leftarrow x - x^0$
$z^0 \leftarrow p(x^0)$, $\Delta z \leftarrow p(x) - z^0$
Construct Gauss–Legendre rules $\{(t_k, w_k^p)\}_{k=1}^{K_p}$ and $\{(s_\ell, w_\ell^f)\}_{\ell=1}^{K_f}$ on $[0,1]$
$\widehat{g} \leftarrow \sum_{\ell=1}^{K_f} w_\ell^f \, \nabla f(z^0 + s_\ell \Delta z)$
$\widehat{\phi} \leftarrow \mathbf{0}$
for $k = 1, \ldots, K_p$ **do**
  $x_k \leftarrow x^0 + t_k \Delta x$
  $r_k \leftarrow J_p(x_k)^\top \widehat{g}$ {vector–Jacobian product}
  $\widehat{\phi} \leftarrow \widehat{\phi} + w_k^p \, (\Delta x \odot r_k)$
end for
return $\widehat{\phi}$

where $D_{\Delta x} = \mathrm{diag}(\Delta x)$. The jth column of A_x is precisely the complete representation atom a_j from Equation 4. Hence,

$$A_x \mathbf{1} = \Delta z,$$

and exact integration gives the completeness identity

$$\mathbf{1}^\top \phi^{\mathrm{Pull}} = \Delta z^\top \bar{g} = F(x) - F(x^0). \tag{11}$$

In practice, we approximate the two integrals independently by numerical quadrature. Algorithm 1 gives the resulting fast implementation using vector–Jacobian products rather than explicitly materializing the encoder Jacobian.

Implementation and cost. Algorithm 1 does not require materializing the $r \times d$ encoder Jacobian. After computing $\widehat{g}$, the quadrature approximation to the average downstream gradient, each encoder quadrature node requires one vector–Jacobian product $J_p(x_k)^\top \widehat{g}$. Thus the fast path uses K_f downstream gradient evaluations and K_p encoder VJPs, together with endpoint forward evaluations. These computational primitives are kept distinct from black-box coalition/model evaluations in our empirical comparisons. The fast approximation requires no coalition enumeration.

Define the feature-wise numerical approximation error by

$$\mathcal{Q}_j = |\widehat{\phi}_j - \phi_j^{\mathrm{Pull}}|. \tag{12}$$

Let f be L_f-Lipschitz on the relevant latent region and define the Shapley-weighted representation term

$$\mathcal{D}_{p,j} = L_f \, \mathbb{E}_{S \sim Q_j^{\mathrm{Sh}}} \left[D_p(S) + D_p(S \cup \{j\}) \right]. \tag{13}$$

Theorem 4 (Pullback Error Decomposition)**.** *For every feature j,*

$$|\widehat{\phi}_j - \phi_j^{\mathrm{Sh}}(F; x, x^0)| \le \mathcal{D}_{p,j} + \mathcal{I}_j(h) + \mathcal{Q}_j. \tag{14}$$

Consequently, Theorem 3 yields

$$|\widehat{\phi}_j - \phi_j^{\mathrm{Sh}}(F; x, x^0)| \le 4 L_f \mathcal{K}_p + \mathcal{I}_j(h) + \mathcal{Q}_j. \tag{15}$$

Proof idea. Compare the true and additive-surrogate coalition marginals using Lipschitz continuity. The resulting representation discrepancy is bounded by $\mathcal{D}_{p,j}$. Surrogate SHAP is its Shapley–Shubik allocation, whereas ideal PullSHAP is its Aumann–Shapley allocation, leaving the term $\mathcal{I}_j(h)$. Finally, add the numerical approximation error $\mathcal{Q}_j$. Appendix B gives the full proof.

Corollary 5 (Conditional Pullback Certificate Template)**.** *If available sound bounds satisfy*

$$\mathcal{D}_{p,j} \le \overline{\mathcal{D}}_{p,j}, \qquad \mathcal{I}_j(h) \le \overline{\mathcal{I}}_j, \qquad \mathcal{Q}_j \le \overline{\mathcal{Q}}_j,$$

then

$$|\widehat{\phi}_j - \phi_j^{\mathrm{Sh}}(F; x, x^0)| \le B_j := \overline{\mathcal{D}}_{p,j} + \overline{\mathcal{I}}_j + \overline{\mathcal{Q}}_j. \tag{16}$$

The decomposition is therefore not automatically an operational certificate. Corollary 5 is a conditional certificate template: certification additionally requires sound upper-bound providers for every term, and we do not claim a generic inexpensive construction of such providers.

Relation to composite IG. Composite integrated gradients couples encoder and downstream derivatives at the same path time:

$$\phi_j^{\mathrm{IG}} = \Delta x_j \int_0^1 \partial_j p(x^0 + t\Delta x)^\top \nabla f(p(x^0 + t\Delta x))\, dt. \tag{17}$$

In contrast, Equation 10 separately averages representation transport and downstream sensitivity before combining them. Equivalently, PullSHAP integrates over two independent path parameters, whereas composite IG couples the two derivatives through a single parameter. They coincide for affine p but generally differ for nonlinear representations; neither universally dominates the other as a SHAP approximation.

Fast, diagnostic, and certified modes. FAST runs Algorithm 1 to compute $\widehat{\phi}$ without coalition enumeration. DIAG additionally samples Shapley-prefix coalitions to estimate representation distortion and AS–SS allocation mismatch. These diagnostics are screening statistics, not certificates:

$$\text{DIAG-PASS} \neq \text{CERTIFIED}.$$

CERTIFIED is available when Corollary 5 is instantiated with sound upper bounds for all error terms. Its cost depends on the corresponding bound providers and is accounted for separately from Algorithm 1. Thus, the computational advantage of FAST is coalition-free approximation of the fixed Shapley target, not faster exact SHAP computation.

6 EXPERIMENTS

We organize the empirical study as mechanism validation rather than broad validation across representation families. Four research questions **RQs** mirror the theory: **RQ1** when is the pullback exact; **RQ2** can representation distortion and downstream allocation mismatch be isolated as distinct fsubmit ailure mechanisms; **RQ3** are the corresponding diagnostics associated with exact error in learned representations; and **RQ4** how does the resulting differentiable approximation compare with practical estimators of the same fixed-baseline SHAP game? We do not seek universal method dominance. Appendix C specifies the exact oracle, population rules, seeds, metrics, and accounting conventions.

RQ1–RQ2: exactness and controlled failure mechanisms. In the affine sanity case, the maximum PULLSHAP relative-ℓ_1 error over 20 runs is 8.08×10^{-16}. More importantly, a nonlinear but additively separable encoder with the quadratic downstream construction used here, for which the induced surrogate lies in the AS–SS coincidence class, remains exact up to numerical precision (9.58×10^{-16} maximum error), while the Pull–IG discrepancy reaches 0.306. Thus nonlinearity alone does not obstruct pullback.

Increasing only encoder cross-interaction raises PULLSHAP error from numerical zero to 3.02×10^{-2} and maximum representation distortion to 0.295, with Spearman $\rho = 0.877$ between distortion and attribution error. Conversely, with affine p and increasing cubic downstream interaction, representation distortion stays at numerical zero while allocation discrepancy tracks attribution error with $\rho = 0.976$. Figure 2 summarizes these two interventions. They empirically isolate the two structural mechanisms appearing in Theorem 4; they do not test tightness of its constants.

RQ3: learned representations. We train smooth linear, componentwise additive nonlinear, and dense MLP encoders on Diabetes, Breast Cancer, and Adult. The primary analysis fixes $r = 4$ and includes only models passing a predictive-quality gate fixed before attribution aggregation. Exact composite SHAP is computed on $d = 10$ models: Diabetes uses its full ten features, whereas Breast Cancer and Adult use separately trained ten-feature exact-track models rather than the corresponding full $d = 30$ and $d = 108$ fitted models.

The successful-model population contains 1,250 exact attribution instances nested within successful trained runs; these are not 1,250 independently trained models. The balanced additive/MLP comparison contains 900 attribution instances.

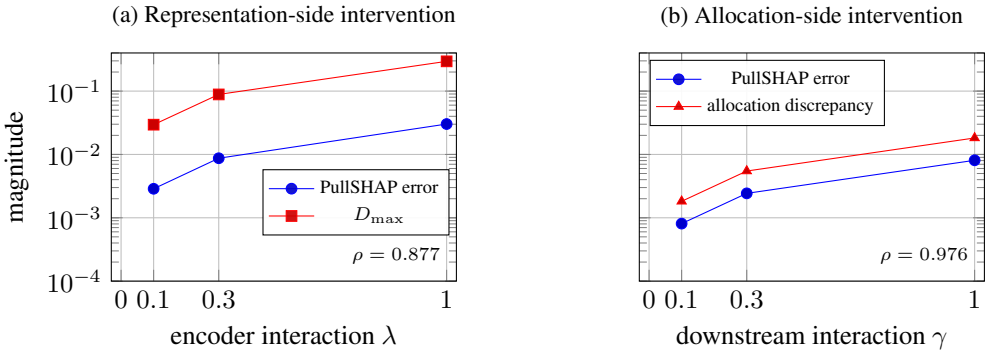

Figure 2: **Controlled mechanism isolation.** (a) Increasing only encoder cross-feature interaction increases both representation distortion and exact-SHAP error. (b) With an affine encoder, representation distortion remains at numerical zero while downstream allocation mismatch increases with exact-SHAP error. Values are medians over five seeds. Points at floating-point scale fall below the displayed log range.

Table 1: **Learned-representation exact track at $r = 4$.** Medians over 150 attribution instances (three trained seeds) per cell. $\widehat{D}_{\mathrm{rep}}$ is the sampled representation-distortion diagnostic, not a certified bound. ρ_{rep} is computed only within MLP strata; correlations among floating-point-scale additive diagnostics are not interpreted.

Dataset	Encoder	Pull err.	IG err.	$\widehat{D}_{\mathrm{rep}}$	ρ_{rep}
Diabetes	Additive	5.65×10^{-7}	3.16×10^{-7}	5.04×10^{-15}	–
Diabetes	MLP	9.57×10^{-3}	9.57×10^{-3}	0.316	0.574
Breast Cancer	Additive	2.70×10^{-3}	4.52×10^{-3}	1.41×10^{-15}	–
Breast Cancer	MLP	6.53×10^{-3}	5.56×10^{-3}	0.084	0.554
Adult	Additive	6.31×10^{-3}	8.37×10^{-3}	9.75×10^{-16}	–
Adult	MLP	9.61×10^{-3}	9.21×10^{-3}	0.070	0.664

Across all three datasets, learned additive encoders have representation distortion at numerical precision, whereas dense MLP encoders have nonzero distortion (Table 1). Within each MLP dataset, the sampled representation diagnostic is positively associated with exact PULLSHAP error ($\rho = 0.554$–0.664). The sampled allocation diagnostic provides complementary association evidence, with within-dataset Spearman correlations 0.365, 0.525, and 0.594 on Diabetes, Breast Cancer, and Adult, respectively.

Across all 1,250 successful $r = 4$ attribution instances, median relative-ℓ_1 error is 0.60%, but the upper tail is non-negligible (Table 2). This is why we treat pullback as a conditional approximation rather than a uniform guarantee.

RQ4: practical exact-SHAP approximation. On a targeted 20-explicand exact-reference benchmark, we compare black-box estimators of the same single-fixed-baseline game: KernelSHAP, PermutationSHAP, LeverageSHAP (Musco & Witter, 2025), and TreeMSR (Witter et al., 2025). Figure 3 reports only the verified common-budget range $m \in \{100, 200, 400, 800\}$. Across this verified budget grid, the observed median errors decrease for all four estimators.

Separately, the differentiable path computation saturates by four to eight Gauss–Legendre nodes on this benchmark. *The two panels deliberately use different horizontal axes*: coalition/model-row budgets for black-box estimators and quadrature nodes (with gradient/Jacobian primitives) for PULLSHAP/IG. We therefore do not equate these costs or claim a common query budget.

Diagnostics and scope. In a small synthetic diagnostic study, a screening score calibrated on 64 cases reaches Spearman $\rho = 0.636$ on 48 unseen interaction cases, with one false-safe among 31 accepted cases. We retain this as screening evidence only; sampled diagnostics are not sound

Table 2: **Error tail over the primary successful-run population at $r = 4$.** N counts attribution instances nested within successful trained runs, not independently trained models. Errors are relative-ℓ_1 errors against exact composite SHAP.

N	Median	Q_{90}	Q_{95}	Q_{99}	$P(E > 2\%)$
1250	0.60%	2.16%	3.20%	7.61%	11.76%

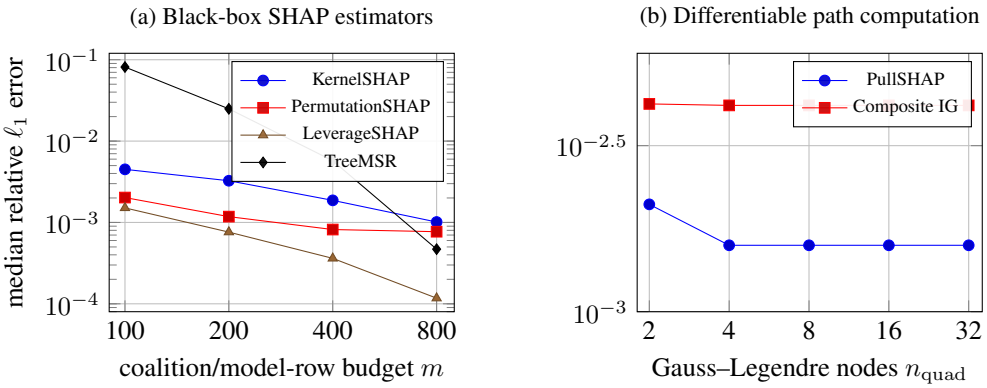

Figure 3: **Practical approximation to the same exact fixed-baseline SHAP target.** (a) Black-box estimators versus coalition/model-row budget. (b) PULLSHAP and composite IG versus quadrature nodes. The horizontal axes represent different computational primitives and are *not* interchangeable query budgets. Each black-box point aggregates 20 explicands and five estimator seeds (100 runs); path points aggregate 20 explicands.

upper bounds. Full screening results, bottleneck ablations, failed predictive runs, linear controls, and execution audits are deferred to the appendix.

In particular, no empirical scalability claim is made from the full $d = 30$ or $d = 108$ models because the exact-reference evaluation uses separate $d = 10$ tracks.

7 Related Work

Shapley semantics and representation-mediated attribution. Shapley values underlie many feature-attribution methods, including SHAP (Lundberg & Lee, 2017), but the resulting explanation depends not only on the Shapley aggregation rule but also on how missing features and reference information define the coalition game (Sundararajan & Najmi, 2020; Covert et al., 2021; Aas et al., 2021). We therefore fix a single-reference baseline game and study recovery of its input-space Shapley values through a learned representation.

Prior work has exploited internal model structure through reference-relative propagation, approximate Shapley propagation, graphical credit assignment, and latent-space explanations (Shrikumar et al., 2017; Ancona et al., 2019; Chen et al., 2022; Wang et al., 2021; Singal et al., 2021; Hu et al., 2024). These works establish the utility of compositional and representation-mediated attribution. Our focus is instead target-relative: when does attribution through $F = f \circ p$ recover, or approximate with controlled error, the Shapley values of the original fixed input-space game? Our theory separates generic non-identifiability, representation-side coalition distortion, and downstream allocation mismatch.

Path allocation and interactions. Gradient-based path methods include Integrated Gradients and Expected Gradients (Sundararajan et al., 2017; Erion et al., 2021). Our downstream analysis uses the classical relationship between Aumann–Shapley path allocation and Shapley–Shubik random-order allocation, including their coincidence on multilinear-plus-additively-separable functions (Sun & Sundararajan, 2011). Higher-order attribution methods instead allocate credit to feature interactions, including Shapley–Taylor, Integrated Hessians, Faith-Shap, and SHAP-IQ (Sundararajan et al., 2020; Janizek et al., 2021; Tsai et al., 2023; Fumagalli et al., 2023). Our mixed cross-partials have a

different role: they are not interaction attributions, but control representation-side coalition distortion in Theorem 3.

Efficient approximation and pullback viewpoints. A complementary literature seeks computationally efficient Shapley approximations through weighted regression, propagation, amortized explainers, and improved sampling or interaction estimators (Lundberg & Lee, 2017; Ancona et al., 2019; Jethani et al., 2022; Fumagalli et al., 2023; 2024). PULLSHAP instead exploits differentiability and the factorization $F = f \circ p$ to construct a pathwise approximation without coalition enumeration; this is not a claim of reduced complexity for exact SHAP, and we keep gradient/Jacobian primitives distinct from black-box coalition evaluations.

Pullback terminology has also appeared in recent neural explanation work (Satkiewicz et al., 2025). Accordingly, our contribution is not pullback terminology, gradient mapping, compositional propagation, or interaction analysis itself. It is a target-relative theory showing why endpoint-local information and latent compression alone are insufficient, when representation geometry controls coalition distortion, and how representation, downstream allocation, and numerical errors combine in approximation to a fixed input-space Shapley target.

8 DISCUSSION AND LIMITATIONS

PULLSHAP targets the fixed-baseline game in Equation 2; conditional SHAP changes the coalition semantics and is outside the present theory. Our positive results assume a smooth representation p and a differentiable downstream predictor f, with mixed second derivatives required for the curvature-based pullability bound. Arbitrary trees, discrete representations, and generic nonsmooth models therefore fall outside the present smooth theory. The exact composite SHAP target is invariant to how F is factorized, whereas the PULLSHAP approximation is factorization dependent because its representation atoms, surrogate, and associated error terms depend on the chosen p. Thus two factorizations of the same input–output function can induce different approximation behavior even though the target Shapley vector is unchanged.

The empirical study is designed primarily as mechanism validation, and its scope is deliberately narrower than the class of differentiable factorizations covered by the method. Exact-reference learned-representation results use $d = 10$ models; the Breast Cancer and Adult tracks are separately trained ten-feature reductions and therefore do not establish exact-SHAP accuracy for the corresponding full $d = 30$ and $d = 108$ fitted models. The practical estimator comparison uses 20 fixed explicands and is a targeted benchmark rather than evidence of universal method dominance. Its two cost axes are intentionally distinct: black-box coalition/model-row evaluations and gradient/Jacobian primitives are different computational resources and are not treated as interchangeable query budgets.

Finally, sampled diagnostics are screening statistics rather than certificates. A CERTIFIED output requires sound upper-bound providers for every term in Corollary 5; such providers may be expensive or unavailable for large neural networks. The decomposition therefore exposes failure modes without making certification automatic. Likewise, low latent dimension alone does not make exact SHAP easy: Theorem 2 shows that even a scalar representation can realize any normalized coalition game on the relevant hybrid vertices. PULLSHAP should instead be viewed as a differentiable, factorization-dependent approximation whose error can be separated into representation distortion, downstream allocation mismatch, and numerical approximation.

More broadly, our results suggest that the suitability of a learned representation for attribution should be assessed not merely by how much it compresses the input, but by how well it preserves the coalition geometry relevant to the target attribution rule. For fixed-baseline SHAP, mixed cross-feature curvature provides one sufficient quantitative route to controlling this representation-side distortion. Extending this perspective to conditional coalition semantics, nonsmooth representations, and richer representation geometries remains open.

Reproducibility Statement

Complete proofs are provided in Appendices A–B, and experimental protocols, exact-reference construction, population accounting, metrics, and additional results are documented in Appendices C–E. An anonymized implementation and reproducibility package is available at `https://anonymous.4open.science/r/pullshap-release-BF61/`. The package contains executable scripts that regenerate the reported empirical tables and figures covered by the package and automatically verify their numerical values.

AI Use Statement

Generative AI tools were used during this project to assist with literature exploration, proof checking, software implementation and debugging, grammar checking, and editing of the manuscript. All mathematical claims, proofs, experimental protocols, code, numerical results, citations, and AI-assisted text were reviewed and verified by the authors. The authors take full responsibility for the scientific content and final presentation of the work.

References

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021. doi: 10.1016/j.artint.2021.103502.

Marco Ancona, Cengiz Öztireli, and Markus Gross. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 272–281. PMLR, 2019.

Hugh Chen, Scott M. Lundberg, and Su-In Lee. Explaining a series of models by propagating shapley values. *Nature Communications*, 13:4512, 2022. doi: 10.1038/s41467-022-31384-3.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott M. Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3:620–631, 2021. doi: 10.1038/s42256-021-00343-w.

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

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. KernelSHAP-IQ: Weighted least square optimization for shapley interactions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14308–14342. PMLR, 2024.

Xuran Hu, Mingzhe Zhu, Zhenpeng Feng, and Ljubiša Stanković. Manifold-based shapley explanations for high dimensional correlated features. *Neural Networks*, 180:106634, 2024. doi: 10.1016/j.neunet.2024.106634.

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

Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-time shapley value estimation. In *International Conference on Learning Representations*, 2022.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Christopher Musco and R. Teal Witter. Provably accurate shapley value estimation via leverage score sampling. In *International Conference on Learning Representations*, 2025.

Maciej Satkiewicz, Roberto Corizzo, and Marcin Pietroń. Pulling back the curtain on deep networks. *arXiv preprint arXiv:2507.22832*, 2025.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR, 2017.

Raghav Singal, George Michailidis, and Hoiyi Ng. Flow-based attribution in graphical models: A recursive shapley approach. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9733–9743. PMLR, 2021.

Yi Sun and Mukund Sundararajan. Axiomatic attribution for multilinear functions. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, pp. 177–178. ACM, 2011. doi: 10.1145/1993574.1993601.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9269–9278. PMLR, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

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

Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. Shapley flow: A graph-based approach to interpreting model predictions. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 721–729. PMLR, 2021.

R. Teal Witter, Yurong Liu, and Christopher Musco. Regression-adjusted monte carlo estimators for shapley values and probabilistic values. In *Advances in Neural Information Processing Systems*, 2025.

A PROOFS FOR THE FUNDAMENTAL LIMITS

This appendix gives complete proofs of Theorems 1 and 2. Throughout, the baseline game associated with a scalar composite F is

$$v_F(S) = F(x_S) - F(x^0), \qquad S \subseteq [d]. \tag{18}$$

We write

$$w_\ell^{(n)} = \frac{\ell!(n-\ell-1)!}{n!}, \qquad 0 \le \ell \le n-1, \tag{19}$$

for the usual Shapley weight in an n-player game.

A.1 PROOF OF THEOREM 1

Proof. Let

$$A = \{j \in [d] : x_j \neq x_j^0\} \tag{20}$$

be the active-feature set and put $m = |A|$. By assumption $m \ge 2$. Coordinates outside A do not change under any hybrid intervention and are therefore dummy players for every composite considered below.

Choose a nonempty proper subset $T^\star \subset A$. The 2^m active hybrid points

$$\{x_T : T \subseteq A\}$$

are pairwise distinct. Indeed, if $T \neq R$, choose $j \in T \triangle R$; the jth coordinates of x_T and x_R are x_j and x_j^0 in some order, and these differ because $j \in A$. In particular, $x_{T^\star}$ is distinct from both endpoints x^0 and x.

Because the set of active hybrid points is finite, the separation

$$\delta = \min_{\substack{T \subseteq A \\ T \neq T^\star}} \|x_T - x_{T^\star}\| \tag{21}$$

is strictly positive. Choose $0 < \rho < \delta/3$. Then, $B_\rho(x_{T^\star})$ contains no other active hybrid point, and its closure is separated from both x^0 and x.

Choose a standard C^∞ bump $b : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$\operatorname{supp}(b) \subset B_\rho(x_{T^\star}), \qquad b(x_{T^\star}) = 1. \tag{22}$$

Thus, b is identically zero on neighborhoods of both x^0 and x.

Set

$$p_0(\xi) \equiv 0, \qquad p_1(\xi) = \varepsilon b(\xi), \qquad f(z) = z, \tag{23}$$

where $\varepsilon \neq 0$. Both representations and the downstream map are smooth. Since b vanishes on neighborhoods of the endpoints, $p_0 = p_1$ on neighborhoods of both x^0 and x. Hence the two representations have identical endpoint germs and, in particular, identical endpoint jets of every finite order.

It remains to show that their baseline Shapley vectors differ. The game induced by $F_0 = f \circ p_0$ is identically zero, so

$$\phi^{\mathrm{Sh}}(F_0; x, x^0) = 0. \tag{24}$$

Consider the reduced m-player game on the active player set A. For $T \subseteq A$, define

$$\bar{v}_1(T) = F_1(x_T) - F_1(x^0). \tag{25}$$

Because $T^\star$ is nonempty and the support of b contains no active hybrid point other than $x_{T^\star}$, we have

$$\bar{v}_1(T) = \begin{cases} \varepsilon, & T = T^\star, \\ 0, & T \neq T^\star, \end{cases} \qquad T \subseteq A. \tag{26}$$

Let $t = |T^\star|$, so $1 \le t \le m-1$. For any $j \in T^\star$, the only predecessor set yielding a nonzero marginal contribution is $T^\star \setminus \{j\}$. Therefore,

$$\phi_j^{\mathrm{Sh}}(\bar{v}_1) = w_{t-1}^{(m)} \varepsilon = \frac{(t-1)!(m-t)!}{m!}\, \varepsilon \neq 0. \tag{27}$$

For any $j \in A \setminus T^\star$, the only predecessor set yielding a nonzero marginal contribution is $T^\star$, for which

$$\bar{v}_1(T^\star \cup \{j\}) - \bar{v}_1(T^\star) = -\varepsilon.$$

Hence,

$$\phi_j^{\mathrm{Sh}}(\bar{v}_1) = -w_t^{(m)}\varepsilon = -\frac{t!(m-t-1)!}{m!}\,\varepsilon \neq 0. \tag{28}$$

Thus, the reduced active-player Shapley vector is nonzero.

For the original d-player game, every coordinate outside A is a dummy player. Deleting the inactive players from a uniformly random permutation of $[d]$ leaves a uniformly random permutation of A, and the marginal contribution of an active feature depends only on its active predecessors. Hence dummy-player extension preserves the Shapley values of all active players and assigns zero to inactive ones. Consequently,

$$\phi^{\mathrm{Sh}}(F_1; x, x^0) \neq \phi^{\mathrm{Sh}}(F_0; x, x^0), \tag{29}$$

which proves the theorem. □

Remark (finite-order consequence). Because the construction makes p_0 and p_1 identical on neighborhoods of both endpoints, their endpoint jets agree to every finite order. Thus, Theorem 1 immediately rules out, as a special case, any universal exact pullback based only on endpoint values and finitely many endpoint derivatives. Additional information about the intervened coalition geometry is necessary.

A.2 Proof of Theorem 2

Proof. Assume $x_j \neq x_j^0$ for every $j \in [d]$ and define the affine normalized coordinates

$$u_j(\xi) = \frac{\xi_j - x_j^0}{x_j - x_j^0}, \qquad j \in [d]. \tag{30}$$

For every coalition $S \subseteq [d]$, the hybrid definition gives

$$u(x_S) = \mathbf{1}_S \in \{0,1\}^d. \tag{31}$$

Let $v : 2^{[d]} \to \mathbb{R}$ be arbitrary with $v(\varnothing) = 0$, and define its Möbius coefficients by

$$\mu_v(T) = \sum_{R \subseteq T} (-1)^{|T|-|R|} v(R), \qquad T \subseteq [d]. \tag{32}$$

Consider the Möbius-form multilinear polynomial

$$P_v(u) = \sum_{T \subseteq [d]} \mu_v(T) \prod_{j \in T} u_j. \tag{33}$$

At a Boolean vertex $\mathbf{1}_S$, the monomial indexed by T equals one exactly when $T \subseteq S$ and zero otherwise. Therefore,

$$P_v(\mathbf{1}_S) = \sum_{T \subseteq S} \mu_v(T). \tag{34}$$

Substituting Equation 32 and interchanging the finite sums yields

$$\sum_{T \subseteq S} \mu_v(T) = \sum_{T \subseteq S} \sum_{R \subseteq T} (-1)^{|T|-|R|} v(R) \tag{35}$$

$$= \sum_{R \subseteq S} v(R) \sum_{T : R \subseteq T \subseteq S} (-1)^{|T|-|R|}. \tag{36}$$

For fixed $R \subseteq S$, write $T = R \cup U$ with $U \subseteq S \setminus R$. The inner sum becomes

$$\sum_{U \subseteq S \setminus R} (-1)^{|U|} = (1-1)^{|S \setminus R|}, \tag{37}$$

which equals one if $R = S$ and zero otherwise. Hence,

$$P_v(\mathbf{1}_S) = v(S) \qquad \text{for all } S \subseteq [d]. \tag{38}$$

This is precisely Möbius inversion on the Boolean lattice.

Now define

$$p(\xi) = P_v(u(\xi)), \qquad f(z) = z. \tag{39}$$

Because u is affine and P_v is a polynomial, $p : \mathbb{R}^d \to \mathbb{R}$ is a smooth scalar representation. Equation 31–Equation 38 imply

$$(f \circ p)(x_S) = v(S) \qquad \text{for every } S \subseteq [d]. \tag{40}$$

In particular, $u(x^0) = \mathbf{0}$ and

$$(f \circ p)(x^0) = P_v(\mathbf{0}) = \mu_v(\varnothing) = v(\varnothing) = 0. \tag{41}$$

Therefore,

$$(f \circ p)(x_S) - (f \circ p)(x^0) = v(S) \qquad \text{for every } S \subseteq [d], \tag{42}$$

which proves Theorem 2. Since the codomain of p is one-dimensional, the construction establishes scalar representation universality on the relevant hybrid vertices. □

Remark (what the theorem does and does not imply). The proof uses the standard Möbius-form multilinear polynomial representation of a set function; the mathematical novelty claimed in the main paper is not Möbius inversion or this representation itself. The theorem instead exposes a structural limitation of representation compression: the condition $r = 1$ alone places no restriction on the coalition game realizable on the relevant hybrid vertices. It does not establish an oracle lower bound for exact Shapley computation, nor does it assert that every scalar representation is attributionally difficult.

B Proofs for Pullability and Error Control

B.1 Proof of Theorem 3

Proof. We prove the sharper coalition-dependent inequality Equation 6; the global bound then follows immediately. Throughout this proof,

$$q(u) = p(x^0 + D_{\Delta x} u), \qquad u \in [0,1]^d.$$

The argument applies to vector-valued q under any fixed norm on $\mathbb{R}^r$, since it uses only the fundamental theorem of calculus and the triangle inequality.

Step 1: anchored atoms. Define

$$b_j = q(e_j) - q(0) = \int_0^1 \partial_j q(te_j)\, dt.$$

Fix $S = \{j_1, \ldots, j_s\}$ in any order and let

$$v_\ell = \sum_{m=1}^{\ell} e_{j_m}, \qquad v_0 = 0.$$

Then,

$$q(\mathbf{1}_S) - q(0) = \sum_{\ell=1}^{s} \big(q(v_\ell) - q(v_{\ell-1})\big) \tag{43}$$

$$= \sum_{\ell=1}^{s} \int_0^1 \partial_{j_\ell} q(v_{\ell-1} + te_{j_\ell})\, dt. \tag{44}$$

Subtracting $\sum_{j\in S} b_j$ gives a sum of differences of first derivatives. For fixed ℓ, move the coordinates $j_1, \ldots, j_{\ell-1}$ one at a time from zero to one. Repeated application of the fundamental theorem of calculus gives

$$\|\partial_{j_\ell} q(v_{\ell-1} + te_{j_\ell}) - \partial_{j_\ell} q(te_{j_\ell})\| \tag{45}$$

$$\leq \sum_{m<\ell} \int_0^1 \left\|\partial^2_{j_m j_\ell} q(\zeta_{m,\ell}(s,t))\right\| \, ds \leq \sum_{m<\ell} \kappa_{j_m j_\ell}, \tag{46}$$

where each $\zeta_{m,\ell}(s,t)$ lies in $[0,1]^d$. Integrating in t, summing over ℓ, and applying the triangle inequality counts each unordered pair internal to S exactly once. Hence

$$\left\| q(\mathbf{1}_S) - q(0) - \sum_{j\in S} b_j \right\| \leq \sum_{\substack{i<j \\ i,j\in S}} \kappa_{ij} = K_{\text{int}}(S). \tag{47}$$

Step 2: complete versus anchored atoms. By Equation 4,

$$a_j - b_j = \int_0^1 [\partial_j q(t\mathbf{1}) - \partial_j q(te_j)]\, dt.$$

Fix $t \in [0,1]$. Connect te_j to $t\mathbf{1}$ by changing each coordinate $k \neq j$ from 0 to t, one coordinate at a time. For the segment corresponding to coordinate k, the fundamental theorem of calculus gives

$$\|\Delta_k \partial_j q\| \leq \int_0^t \left\|\partial^2_{kj} q(\zeta_{k,j}(s,t))\right\| \, ds \leq t\, \kappa_{jk},$$

where $\zeta_{k,j}(s,t) \in [0,1]^d$. Telescoping over all $k \neq j$ therefore yields

$$\|\partial_j q(t\mathbf{1}) - \partial_j q(te_j)\| \leq t \sum_{k\neq j} \kappa_{jk}.$$

Integrating in t gives

$$\|a_j - b_j\| \leq \int_0^1 t\, dt \sum_{k\neq j} \kappa_{jk} = \frac{1}{2} \sum_{k\neq j} \kappa_{jk}. \tag{48}$$

Step 3: coalition distortion. By definition,

$$D_p(S) = \left\| q(\mathbf{1}_S) - q(0) - \sum_{j\in S} a_j \right\|.$$

Adding and subtracting the anchored atoms gives

$$D_p(S) \leq \left\| q(\mathbf{1}_S) - q(0) - \sum_{j\in S} b_j \right\| + \sum_{j\in S} \|a_j - b_j\|. \tag{49}$$

Applying Equation 47 and Equation 48, observe that in

$$\frac{1}{2} \sum_{j\in S} \sum_{k\neq j} \kappa_{jk},$$

each unordered pair internal to S is counted twice, whereas each pair crossing from S to S^c is counted once. Therefore,

$$D_p(S) \leq K_{\text{int}}(S) + K_{\text{int}}(S) + \frac{1}{2} K_{\text{cross}}(S) \tag{50}$$

$$= 2K_{\text{int}}(S) + \frac{1}{2} K_{\text{cross}}(S), \tag{51}$$

which proves the coalition-dependent inequality in Equation 6.

Moreover,

$$2K_{\text{int}}(S) + \frac{1}{2}K_{\text{cross}}(S) \le 2\big(K_{\text{int}}(S) + K_{\text{cross}}(S)\big) \le 2\sum_{i<j}\kappa_{ij} = 2\mathcal{K}_p.$$

Hence,

$$\max_{S\subseteq[d]} D_p(S) \le 2\mathcal{K}_p,$$

completing the proof. □

B.2 Additive-separability consequence

Suppose

$$\partial^2_{ij} q(u) = 0 \qquad \text{for all } i \neq j \text{ and } u \in [0,1]^d.$$

Then, $\partial_i q(u)$ is independent of every coordinate u_j with $j \neq i$. Hence there exists a one-variable vector-valued function γ_i such that

$$\partial_i q(u) = \gamma_i(u_i).$$

Define

$$G_i(t) = \int_0^t \gamma_i(s)\, ds.$$

Integrating coordinate by coordinate from 0 to u yields

$$q(u) = q(0) + \sum_{i=1}^{d} G_i(u_i), \qquad G_i(0) = 0. \tag{52}$$

Thus, q, equivalently the restriction of p to the baseline–explicand box under the normalized coordinates, is additively separable across input coordinates.

The coalition consequence also follows directly from Theorem 3: if all mixed partials vanish, then $\mathcal{K}_p = 0$, and therefore

$$D_p(S) = 0 \qquad \text{for every } S \subseteq [d].$$

This implication concerns representation-side coalition distortion; downstream AS–SS allocation mismatch is a separate condition.

B.3 Proof of Theorem 4

Proof. We separate the approximation into the same three terms appearing in the theorem. Let

$$v(S) = f(p(x_S)) - f(z^0), \qquad \widetilde{v}(S) = f\left(z^0 + \sum_{i\in S} a_i\right) - f(z^0).$$

Let $\mathcal{Z}$ be a latent set containing

$$\{p(x_S) : S \subseteq [d]\} \;\cup\; \left\{z^0 + \sum_{i\in S} a_i : S \subseteq [d]\right\}, \tag{53}$$

and assume that f is L_f-Lipschitz on $\mathcal{Z}$. By definition of $D_p(S)$, the two corresponding latent arguments differ by norm $D_p(S)$. Therefore

$$|v(S) - \widetilde{v}(S)| \le L_f D_p(S) \qquad \text{for every } S \subseteq [d]. \tag{54}$$

Step 1: representation distortion. Fix feature j. Using the predecessor-set representation of the Shapley value and Eq. equation 54,

$$\begin{aligned}
&|\phi_j^{\text{Sh}}(v) - \phi_j^{\text{Sh}}(\widetilde{v})| && (55)\\
&= \left|\mathbb{E}_{S\sim Q_j^{\text{Sh}}}\Big[(v(S\cup\{j\}) - v(S)) - (\widetilde{v}(S\cup\{j\}) - \widetilde{v}(S))\Big]\right| && (56)\\
&\le \mathbb{E}_{S\sim Q_j^{\text{Sh}}}\left[|v(S\cup\{j\}) - \widetilde{v}(S\cup\{j\})| + |v(S) - \widetilde{v}(S)|\right] && (57)\\
&\le L_f \mathbb{E}_{S\sim Q_j^{\text{Sh}}}\left[D_p(S\cup\{j\}) + D_p(S)\right] = \mathcal{D}_{p,j}. && (58)
\end{aligned}$$

This is a feature-wise, Shapley-weighted stability bound; no maximum over coalitions has been taken.

Step 2: downstream allocation. For

$$h(u) = f\left(z^0 + \sum_i u_i a_i\right),$$

the restriction of h to Boolean vertices satisfies

$$\widetilde{v}(S) = h(\mathbf{1}_S) - h(\mathbf{0}).$$

The additive constant does not affect Shapley values, so

$$\phi_j^{\mathrm{Sh}}(\widetilde{v}) = \phi_j^{\mathrm{SS}}(h).$$

Moreover, Equation 10 and the fact that the jth column of A_x is a_j give

$$\phi_j^{\mathrm{Pull}} = a_j^\top \int_0^1 \nabla f(z^0 + t\Delta z)\, dt \tag{59}$$

$$= \int_0^1 \partial_j h(t\mathbf{1})\, dt = \phi_j^{\mathrm{AS}}(h). \tag{60}$$

Therefore, by Equation 9,

$$|\phi_j^{\mathrm{Sh}}(\widetilde{v}) - \phi_j^{\mathrm{Pull}}| = |\phi_j^{\mathrm{SS}}(h) - \phi_j^{\mathrm{AS}}(h)| = \mathcal{I}_j(h). \tag{61}$$

Unlike Step 1, this is an identity rather than an upper bound.

Step 3: numerical approximation. By definition,

$$|\widehat{\phi}_j - \phi_j^{\mathrm{Pull}}| = \mathcal{Q}_j. \tag{62}$$

Insert ϕ_j^{Pull} and $\phi_j^{\mathrm{Sh}}(\widetilde{v})$ between $\widehat{\phi}_j$ and $\phi_j^{\mathrm{Sh}}(v)$. The triangle inequality together with Equation 58–Equation 62 gives

$$|\widehat{\phi}_j - \phi_j^{\mathrm{Sh}}(v)| \le \mathcal{D}_{p,j} + \mathcal{I}_j(h) + \mathcal{Q}_j.$$

Since

$$\phi_j^{\mathrm{Sh}}(v) = \phi_j^{\mathrm{Sh}}(F; x, x^0),$$

this proves Equation 14.

Step 4: global cross-curvature bound. Theorem 3 gives

$$D_p(T) \le 2\mathcal{K}_p \qquad \text{for every } T \subseteq [d].$$

Consequently,

$$\mathcal{D}_{p,j} = L_f \mathbb{E}_{S\sim Q_j^{\mathrm{Sh}}}\left[D_p(S) + D_p(S\cup\{j\})\right] \tag{63}$$

$$\le L_f \mathbb{E}_{S\sim Q_j^{\mathrm{Sh}}}\left[2\mathcal{K}_p + 2\mathcal{K}_p\right] \tag{64}$$

$$= 4L_f\mathcal{K}_p. \tag{65}$$

Substitution into Equation 14 gives

$$|\widehat{\phi}_j - \phi_j^{\mathrm{Sh}}(F; x, x^0)| \le 4L_f\mathcal{K}_p + \mathcal{I}_j(h) + \mathcal{Q}_j,$$

completing the proof. □

Conditional-certificate corollary. Corollary 5 follows immediately from Theorem 4. If sound upper bounds satisfy

$$\mathcal{D}_{p,j} \le \overline{\mathcal{D}}_{p,j}, \qquad \mathcal{I}_j(h) \le \overline{\mathcal{I}}_j, \qquad \mathcal{Q}_j \le \overline{\mathcal{Q}}_j,$$

then

$$|\widehat{\phi}_j - \phi_j^{\mathrm{Sh}}(F; x, x^0)| \le \overline{\mathcal{D}}_{p,j} + \overline{\mathcal{I}}_j + \overline{\mathcal{Q}}_j. \tag{66}$$

Thus, the corollary is conditional on the availability of sound upper-bound providers; the theorem itself does not supply generic inexpensive providers for these quantities.

If a sound provider additionally supplies

$$\mathcal{K}_p \le \overline{\mathcal{K}}_p,$$

then Theorem 3 implies

$$D_p(S) \le 2\overline{\mathcal{K}}_p \qquad \text{for every } S,$$

and hence,

$$\mathcal{D}_{p,j} \le 4L_f\overline{\mathcal{K}}_p.$$

This provides a sufficient representation-side contribution to the certificate but does not by itself control the downstream allocation or numerical terms.

Exact-regime check. If $D_p(S) = 0$ for every coalition, if the AS–SS allocation mismatch also vanishes, and if the path integrals are evaluated exactly, then

$$\mathcal{D}_{p,j} = \mathcal{I}_j(h) = \mathcal{Q}_j = 0.$$

Theorem 4 therefore gives

$$\phi_j^{\mathrm{Pull}} = \phi_j^{\mathrm{Sh}}(F; x, x^0) \qquad \text{for every } j.$$

For example, the allocation mismatch vanishes under the classical AS–SS coincidence conditions stated in Section 4. This check also makes explicit that $\mathcal{D}_{p,j}$ and $\mathcal{I}_j(h)$ are distinct structural terms, whereas $\mathcal{Q}_j$ is numerical.

C Experimental Details

Exact oracle and target semantics. For every exact-track problem, we evaluate the fixed-baseline game

$$v(S) = F(x_S) - F(x^0)$$

and compute Shapley values by exact weighted coalition enumeration. All exact-track models have $d = 10$, hence $2^{10} = 1024$ hybrid coalitions are evaluated per explicand. All exact-oracle calculations use float64. Diabetes uses its full ten input features. Breast Cancer and Adult use separately trained ten-feature reduced models; their exact-track results are therefore not exact SHAP values of the corresponding full $d = 30$ and $d = 108$ fitted models.

Synthetic families. E0 uses an affine encoder and a linear downstream function. E1 uses a componentwise nonlinear additive encoder together with a quadratic downstream construction chosen so that the induced surrogate lies in the AS–SS coincidence regime; representation cross-interaction is therefore zero and PullSHAP is exact up to numerical precision. E5 uses the multilinear cubic

$$h(u) = u_1 u_2 u_3.$$

The representation-interaction sweep (E3) augments an additive encoder by smooth pairwise cross-feature terms with strength $\lambda \in \{0, 0.1, 0.3, 1\}$ while using a quadratic downstream map. The allocation sweep (E4) uses an affine encoder and varies a downstream component with strength $\gamma \in \{0, 0.1, 0.3, 1\}$ to induce controlled AS–SS allocation mismatch. A joint (λ, γ) sweep checks that representation distortion and downstream allocation mismatch can coexist. E12 compares PullSHAP and composite IG in the separable exact family and in the interacting-encoder family.

Learned-representation population freeze. The learned study uses Diabetes (regression), Breast Cancer (binary classification), and Adult (binary classification), with linear, componentwise additive nonlinear, and dense smooth MLP encoders. The primary analysis fixes $r = 4$ and includes only trained runs that pass a predictive-quality gate frozen before attribution aggregation; the gate criteria and run-level outcomes are retained in the reproducibility archive. All Breast Cancer and Adult runs pass. On Diabetes, two $r = 4$ linear runs fail the gate and are excluded from the primary accuracy population but retained as stress-test records.

The primary successful-model population contains 1,250 exact attribution instances nested within successful trained runs. The balanced additive/MLP structural comparison contains 900 instances: three datasets, two encoder families, three seeds, and 50 explicands per seed. Successful $r = 4$ linear controls contain 350 instances. Unless otherwise stated, attribution instances within a trained model/seed run are not treated as independent training replicates. Uncertainty estimates and resampling procedures operate at the trained-run level rather than independently resampling explicands.

Diagnostics. The sampled representation diagnostic, denoted $\widehat{D}_{\mathrm{rep}}$ in the experiments, evaluates discrepancies between true coalition representations and the additive atom surrogate over sampled Shapley-prefix coalitions. It is an empirical diagnostic and should not be confused with the theorem-level quantity $\mathcal{D}_{p,j}$. The sampled allocation diagnostic compares path allocation with sampled permutation marginals of the additive surrogate game and is likewise an empirical approximation rather than a sound upper bound on $\mathcal{I}_j(h)$.

For a separate screening experiment, these quantities are combined into a normalized empirical score. Its threshold is selected on a 64-case calibration set and transferred unchanged to 48 cases from three unseen interaction families. This experiment evaluates diagnostic usefulness only: the sampled quantities, normalized score, and selected threshold are not theorem-level certificates and are not used to label CERTIFIED outputs.

Practical benchmark and accounting. The exact-reference practical benchmark contains 20 fixed explicands drawn from four audited $d = 10$ model bundles. KernelSHAP, PermutationSHAP, LeverageSHAP (Musco & Witter, 2025), and TreeMSR (Witter et al., 2025) target the same single-reference fixed-baseline game. In the main comparison they are evaluated at coalition/model-row budgets

$$m \in \{100, 200, 400, 800\},$$

with five estimator seeds per explicand and budget. PULLSHAP and composite IG instead use Gauss–Legendre node counts

$$n_{\mathrm{quad}} \in \{2, 4, 8, 16, 32\}.$$

These horizontal axes measure different computational primitives and are not interpreted as interchangeable query budgets. In particular, a gradient or encoder VJP/Jacobian primitive is not equated with a black-box forward model row. Exact SHAP references are precomputed independently and are not exposed to any estimator.

Metrics. The primary attribution metric is relative ℓ_1 error to exact SHAP. The reproducibility archive additionally records relative ℓ_2 error, cosine similarity, Spearman rank correlation, top-k agreement, sign agreement, completeness residual, wall-clock time, and primitive/model evaluation counts. Main-text reporting is restricted to quantities needed for RQ1–RQ4. Timing comparisons use the benchmark protocol reported with the archive; computational primitives are also reported separately so that wall-clock measurements are not used to equate gradient-based and black-box access models.

D ADDITIONAL EXPERIMENTAL RESULTS

Table 3: Controlled synthetic sweeps. Values are medians over five seeds per condition unless stated otherwise. $D_{\max}$ denotes the maximum coalition representation distortion, and I denotes the measured downstream allocation discrepancy.

Condition	Pull Rel-ℓ_1	IG Rel-ℓ_1	$D_{\max}$	I
E3 $\lambda = 0$	2.51×10^{-16}	4.88×10^{-3}	4.01×10^{-16}	2.43×10^{-16}
E3 $\lambda = 0.1$	2.87×10^{-3}	6.43×10^{-3}	2.95×10^{-2}	2.78×10^{-16}
E3 $\lambda = 0.3$	8.71×10^{-3}	9.37×10^{-3}	8.86×10^{-2}	4.09×10^{-16}
E3 $\lambda = 1$	3.02×10^{-2}	1.03×10^{-2}	2.95×10^{-1}	2.28×10^{-16}
E4 $\gamma = 0$	1.93×10^{-16}	2.05×10^{-16}	3.39×10^{-16}	1.46×10^{-16}
E4 $\gamma = 0.1$	8.09×10^{-4}	8.09×10^{-4}	3.39×10^{-16}	1.82×10^{-3}
E4 $\gamma = 0.3$	2.43×10^{-3}	2.43×10^{-3}	3.39×10^{-16}	5.47×10^{-3}
E4 $\gamma = 1$	8.08×10^{-3}	8.08×10^{-3}	3.39×10^{-16}	1.82×10^{-2}

Across the E3 sweep, the Spearman correlation between $D_{\max}$ and PullSHAP error is 0.877. Across E4, the Spearman correlation between the measured allocation discrepancy and PullSHAP error is 0.976. These controlled sweeps are designed to isolate the two error mechanisms rather than establish a general predictive relationship. The joint (λ, γ) sweep is retained as a mechanism check rather than used to claim empirical additivity, since representation and allocation contributions can interact and partially cancel in signed attribution coordinates.

Diagnostic screening. On the 64-case calibration set, the synthetic screening score has Spearman correlation $\rho = 0.641$ with exact attribution error. Without retuning, the same score has $\rho = 0.636$ on 48 cases from three unseen interaction families. Of the 31 cases passing the screening threshold, one is false-safe, corresponding to an empirical false-safe fraction of 3.2% within this small test

set. These quantities characterize a screening experiment only; they do not provide a certification guarantee or a population-level bound on false-safe probability.

Learned-representation tail and controls. At $r = 4$, the primary successful-model population contains 1,250 exact attribution instances nested within successful trained runs. Its median PullSHAP relative-ℓ_1 error is 0.60%, with 90th, 95th, and 99th percentiles of 2.16%, 3.20%, and 7.61%, respectively; 11.76% of instances exceed 2%. Successful $r = 4$ linear controls contain 350 exact attribution instances. PullSHAP and composite IG coincide with each other numerically in the affine-encoder controls, as predicted by their affine-encoder equivalence; this equivalence does not by itself imply exact agreement with SHAP. Failed predictive models are retained in the reproducibility archive but excluded from the primary attribution aggregates.

Diabetes bottleneck ablation. For successful MLP runs at $r \in \{2, 4, 8\}$, median PullSHAP relative-ℓ_1 error is 1.30%, 0.96%, and 0.73%, respectively, while median sampled representation distortion is 0.332, 0.316, and 0.286. The corresponding fractions of instances exceeding 2% relative-ℓ_1 error are 26.7%, 11.3%, and 5.3%. This is an observational ablation within one dataset and the available successful trained runs; it does not establish that increasing bottleneck dimension causally improves pullability.

Practical benchmark: full verified range. Table 4 records the median exact-SHAP error used in Figure 3. Each black-box table entry aggregates 20 fixed explicands and five estimator seeds per explicand, whereas each path-method entry in Table 5 aggregates the same 20 fixed explicands. Higher-budget exhaustive or implementation-specific outputs are omitted from the main comparison when they no longer follow the common verified sampling regime.

Table 4: Verified practical exact-reference benchmark. Columns 100–800 are coalition/model-row budgets. These budgets are not equal-cost counterparts of the differentiable primitives reported in Table 5.

Method	100	200	400	800	Primitive
KernelSHAP	4.495×10^{-3}	3.256×10^{-3}	1.869×10^{-3}	1.012×10^{-3}	model rows
PermutationSHAP	2.023×10^{-3}	1.179×10^{-3}	8.182×10^{-4}	7.699×10^{-4}	model rows
LeverageSHAP	1.510×10^{-3}	7.611×10^{-4}	3.621×10^{-4}	1.173×10^{-4}	model rows
TreeMSR	8.131×10^{-2}	2.491×10^{-2}	5.739×10^{-3}	4.697×10^{-4}	model rows

Table 5: Differentiable path computation on the same 20 exact-reference explicands. Columns are Gauss–Legendre node counts n_{quad}, not black-box model-row budgets.

Method	2	4	8	16	32
PullSHAP	2.105×10^{-3}	1.585×10^{-3}	1.585×10^{-3}	1.585×10^{-3}	1.585×10^{-3}
Composite IG	4.224×10^{-3}	4.181×10^{-3}	4.181×10^{-3}	4.181×10^{-3}	4.181×10^{-3}

TreeMSR execution audit. In our tested invocation of the pinned official TreeMSR implementation, a dedicated forensic sweep verifies the standard sampling path through $m = 1023$ and locates a branch transition at the $d = 10$ coalition-space boundary $m = 2^{10} = 1024$. In the tested one-explicand invocation, this transition enters an enumeration helper that expects a collection of explicands; supplying a one-dimensional explicand changes the input semantics. A correctly batched control recovers the exact target to numerical precision. We therefore retain the common verified benchmark grid $m \in \{100, 200, 400, 800\}$ and exclude the affected high-budget outputs. This observation documents the behavior of the tested implementation/API invocation and is not evidence of mathematical instability of TreeMSR itself.

E Reproducibility Checklist and Implementation Details

Determinism and provenance. All reported experiment families use fixed random seeds recorded with their run artifacts. Raw instance-level attribution outputs are retained alongside aggregate

summaries so that every reported table statistic, quantile, and error distribution can be recomputed without retraining. Experiment configurations, model/encoder settings, preprocessing metadata, and run-level predictive-quality outcomes are retained with the corresponding artifacts. Failed predictive runs are preserved rather than silently replaced or rerun until successful.

Population accounting. The learned-representation archive contains 2,250 exact-attribution instances in total. The primary $r = 4$ accuracy population contains 1,250 exact attribution instances nested within trained runs that pass the predictive-quality gate. The balanced additive/MLP structural comparison contains 900 instances, and successful $r = 4$ linear controls contain 350 instances. Attribution instances within the same trained model/seed run are not treated as independent training replicates; run-level resampling is used where uncertainty across trained runs is reported.

The practical benchmark uses 20 fixed explicands drawn from four audited $d = 10$ model bundles. For every reported black-box budget, each estimator uses five estimator seeds per explicand. Path-method results use the same fixed explicands but are indexed by quadrature order rather than black-box budget.

Exact-reference construction. Every exact-track attribution target is computed for a $d = 10$ model by full enumeration of all

$$2^{10} = 1024$$

hybrid coalitions under the fixed baseline. Diabetes uses its original ten-feature model. Breast Cancer and Adult exact-reference experiments use separately trained ten-feature reduced models; these references are not presented as exact SHAP values for the corresponding full $d = 30$ and $d = 108$ fitted models. Exact references are generated and stored independently of the approximation procedures and are not exposed to an estimator during attribution.

Semantic integrity. Every method in the practical benchmark targets the same single-reference fixed-baseline game

$$v(S) = F(x_S) - F(x^0).$$

For black-box estimators, the benchmark records the requested budget, actual model rows, unique coalitions where available, runtime, random seed, attribution vector, relative error, and completeness residual. For differentiable path methods, gradient, encoder VJP/Jacobian, and forward-evaluation primitives are recorded separately where applicable. A gradient or Jacobian primitive is not converted into an equivalent number of black-box model rows. Wall-clock measurements are therefore reported alongside, rather than as a replacement for, primitive-level accounting.

External estimators. LeverageSHAP uses the official implementation associated with Musco & Witter (2025), and TreeMSR uses the official Regression-MSR implementation associated with Witter et al. (2025). The reproducibility artifacts record the external repository revisions used for the reported experiments together with the verified execution range. The TreeMSR implementation/API behavior at the $d = 10$ coalition-space boundary is documented separately in Section D. No affected high-budget output is included in the estimator-quality comparison.

Numerical implementation. Exact-reference calculations use float64. PULLSHAP evaluates the representation and downstream path integrals using Gauss–Legendre quadrature. The implementation computes the downstream averaged gradient first and then evaluates encoder vector–Jacobian products along the representation path, avoiding explicit materialization of the full $r \times d$ encoder Jacobian on the fast path. Composite IG is evaluated using the corresponding numerical path-integration protocol. Quadrature order is recorded for every path-method result.

Artifact-to-paper traceability. The reproducibility package retains raw records separately from paper-level summaries. Aggregate tables and figures are derived from the frozen raw outputs rather than manually transcribed experimental results. Run identifiers link each reported attribution instance to its dataset, preprocessing configuration, encoder family, latent dimension, training seed, predictive-quality outcome, explicand, baseline, and attribution configuration. This separation allows the exact-reference results, learned-representation aggregates, controlled sweeps, and practical benchmark to be audited independently.

Scope of reproducibility claims. The archived artifacts are intended to reproduce the reported experimental results and their aggregation under the stated configurations. Sampled diagnostics remain empirical screening quantities and are not converted into theorem-level certificates. Likewise, reproduction of the targeted exact-reference benchmark should not be interpreted as evidence of estimator behavior outside the reported models, explicands, budgets, or access regimes.

Code reproducibility. An anonymized implementation and reproducibility package is available at `https://anonymous.4open.science/r/pullshap-release-BF61/`. The repository contains executable scripts that regenerate the empirical tables and figures covered by the package from frozen experimental artifacts and verify the corresponding reported numerical values.