BEYOND FRAME-CORRECTNESS: NON-IDENTIFICATION AND ACTIONABILITY OF EQUIVARIANT ATTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Attributions for structured inputs can depend on arbitrary choices of coordinate frame, gauge, or discretization, and a growing literature addresses this by enforcing *frame-correctness* — requiring explanations to transform consistently with the symmetries of the model. We ask what this property establishes beyond soundness. First, equivariance and completeness do not identify a unique attribution rule: for path attributions on a Lie group, we show that left-covariant path families yield frame-correct attributions, construct a second complete left-covariant family on $SE(3)$, and give an explicit analytic witness showing that two admissible rules can assign different credit. On a docking scorer, two frame-exact rules select different degrees of freedom on 60.4% of inputs. Second, we test whether correcting frame dependence or adding finer attribution improves model-external outcomes. In matched intervention studies, repairing IG removes frame-dependent decision flips but does not measurably improve pose quality; adding four exact-Shapley-derived features to a matched reward-model policy changes Success@1 by -0.33 pp (95% CI $[-1.55, +0.88]$). Together, these results separate frame-correctness from rule identification and downstream decision utility.

1 INTRODUCTION

Post-hoc attribution assigns credit for a model's output to parts of its input. When that input is a geometric or structured object, its numerical representation carries choices the model is indifferent to: a molecule's coordinates depend on the world frame, a reward trained from pairwise preferences is identified only up to a prompt-dependent shift, and a function sampled on a grid depends on the resolution. Call an attribution **frame-correct** if it transforms as the quantity it explains — invariant for a scalar such as a molecular energy under $SE(3)$, equivariant for a vector such as a force, unchanged by the reward gauge, convergent under grid refinement. Standard attributions often are not: per-coordinate integrated gradients (IG) on an $SE(3)$-invariant pair potential has relative equivariance defect 1.22, and on a real docking scorer the top-attributed degree of freedom changes under a pure frame rotation on 57.8% of decisions.

A line of work repairs this defect. Crabbé and van der Schaar (2023) formalize that explanations of invariant models must respect the symmetry and give a way to increase any method's invariance; SITE (Wang and Wang, 2021) learns transformation-equivariant interpretations by regularization; manifold and Riemannian integrated gradients (Zaher et al., 2024; Costanza and Simpson, 2025) integrate along geodesics of a chosen metric; and attributing in a canonical frame (Dym et al., 2024) is the standard route that predates these methods. As recently as 2026, new post-hoc methods impose explanation-level equivariance through Lie-algebra intertwining constraints (Radiuk et al., 2026). We take their goal as given: frame-correctness is a soundness requirement, and establishing it is not our contribution.

Frame-correctness addresses one precise defect: whether an explanation changes merely because the same model and input are represented in a different frame. It does not by itself answer two separate questions. First, **rule selection**: if several attribution rules satisfy the same frame-correctness and completeness requirements, which one should be used? Second, **decision value**: does repairing

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.

the frame defect, or computing a finer attribution, improve a downstream decision made from the explanation?

We study these questions separately. For rule selection, we show that equivariance and completeness leave a non-trivial family of admissible path attributions. For decision value, we use matched intervention experiments in which realized outcomes are external to the explained model. Frame-correctness is necessary for a sound explanation, but selecting an attribution rule and using it for action require additional evidence.

Contributions.

- **Rule-level non-identification.** We show that naturality and completeness admit multiple path-attribution rules. On $SE(3)$, both the screw path and a rotate-then-translate path are left-covariant and complete; an explicit $\mathbb{R}^2$ witness proves that the resulting attributions can disagree. On a docking scorer, the two frame-exact rules select different top degrees of freedom on 60.4% of inputs (Section 4).
- **Decision value.** We test whether correcting frame dependence or adding finer attribution improves model-external outcomes under matched intervention protocols. Repairing IG removes frame-dependent decision flips in docking but does not improve pose quality; in a separate preregistered reward-model study, adding four Shapley-derived features changes Success@1 by -0.33 pp (95% CI $[-1.55, +0.88]$) (Section 5, Appendix M).

We also state the structural scope under which an invariant attribution is constrained by complete invariant descriptors, and where that characterization stops (Section 6). Supporting domains, proofs and full experimental protocols are in the appendices.

2 SETUP

Attribution. For a model $f : \mathcal{X} \to \mathbb{R}$ and input x, an attribution $A(x) \in \mathbb{R}^{\dim x}$ scores input components. We use integrated gradients, $A_i(x; b) = (x - b)_i \int_0^1 \partial_i f(b + \alpha(x - b))\, d\alpha$, because it is completeness-exact and the canonical target of equivariant variants, and exact Shapley values where an algebraic completeness identity is needed.

Frame-correctness is naturality. A group G acts on $\mathcal{X}$. Changing the world frame moves the *model* as well as the input, so the property to ask of a rule Ψ mapping a model and endpoints to an attribution is naturality,

$$\Psi[f^g;\, gb,\, gx] \;=\; \rho(g)\, \Psi[f;\, b, x] \quad \text{for all } f, g, \qquad f^g(\cdot) := f(g^{-1}\cdot),$$

with ρ the representation carried by the attributed quantity. For a G-invariant f this reduces to the familiar $A(gx) = \rho(g)A(x)$; we use *invariant* for $\rho = \mathrm{id}$ (a scalar per element) and *equivariant* otherwise, and *frame-correct* as the umbrella term. Requiring $f(g\cdot) = f(\cdot)$ of the model itself would be the wrong condition here: on a group acting transitively on itself it forces f constant (Appendix A).

Native quantities. Models that respect G expose symmetry-respecting objects as part of the forward computation, and we keep four of them distinct. **(i) Native invariant descriptors** $h = (h_1, \ldots, h_N)$: the local G-invariant features the model computes through. **(ii) A native scalar readout** $n(x)$: an invariant scalar per element or per comparison, such as an atomic energy decomposition $E = \sum_i E_i$, the pooled contribution of a message-passing model, or the quantity fixed by identifiability in a reward model (Skalse et al., 2023). **(iii) A native gradient readout**: an equivariant vector such as the force $-\nabla f$. **(iv) A native-derived decision policy** π_n: a rule that picks an action by querying the model. Two such policies appear below and are *not* interchangeable — a *prefix* rule that scores truncations of the model's own input, and a *candidate-rescoring* rule that scores each post-intervention candidate. Where a comparison uses a policy rather than a readout we say so.

Outcome measures. A *realized* outcome is external to the explained model: a ground-truth answer checked by a symbolic verifier, or RMSD to a crystal pose. Frame-flip rates, rank correlations and localization accuracy are proxies and we label them as such.

3 THE DEFECT THE REPAIR ADDRESSES IS REAL, AND BOUNDED

Minimal synthetic models confirm the frame-dependence that motivates the field. For an $SE(3)$-invariant pair potential, per-coordinate IG has relative equivariance defect ≈ 1.22 while the atom-wise contraction is invariant to machine precision (2.6×10^{-16}); the same pattern appears for reward models (baseline-induced instability 0.33) and for neural operators (cross-resolution drift 0.996 against 0.013 for the attribution density). These probes use randomly-initialized networks; they establish the premise, not a magnitude.

3.1 HOW BIG IS THE DEFECT? A CHANCE-CALIBRATED MEASUREMENT

Frame-flip rates are routinely reported without a reference line, which makes their magnitude uninterpretable: when the conclusion is $\arg\max_k |A_k|$ over d degrees of freedom, a conclusion carrying *no* frame-consistent information already flips at rate $\approx 1 - 1/d$; matching the per-complex marginal instead of assuming a uniform one lowers that line, because conclusions concentrate on a few degrees of freedom. On a learned pose/torsion scorer over $SE(3) \times \mathbb{T}^m$ ($d = 10$–28; 47 complexes $\times\, 12$ draws $= 564$ decisions per condition), the observed flip rate rises from 23.6% to 68.6% as the frame change grows, while the marginal-matched chance line computed within each complex sits at 85.2–89.0%; Cohen's κ falls from 0.72 to 0.23 (Appendix L). The defect is real and grows, and it is also bounded: $\kappa = 0.23$ is far from stable, so such a method is not usable as it stands, but the conclusion retains frame-consistent signal well above chance even under a complete reorientation, so the repair recovers a degraded signal rather than manufacturing a new one. Independent work reports the same chance-calibrated correction in a setting with no explicit group-theoretic formulation (ul Hassan and Ebrahimi, 2026).

4 EQUIVARIANCE DOES NOT IDENTIFY THE ATTRIBUTION RULE

If frame-correctness selected a method, one could read a uniqueness theorem as a justification for it. It does not, and this section makes the failure precise.

(a) A whole family of path attributions is frame-correct. We consider body path-attributions on a Lie group (Appendix A): among rules obtained by integrating the body-frame derivative along a path from b to x, completeness is automatic by the chain rule, and left-covariance of the path family, $\gamma_{gb,gx} = g\,\gamma_{b,x}$, is *sufficient* for naturality for all f. We claim sufficiency only; nothing below needs the converse.

Constant body velocity is an additional path-selection criterion: off the cut locus it singles out the screw path, in which case the rule is the group-path integral. It does not follow from equivariance or completeness, so adopting it is a further choice.

> **Proposition 1 (equivariance does not identify an attribution).** There exist a Lie group G, a G-invariant model F (that is, $F(gc, gq) = F(c, q)$ for all g), and two attribution rules $A^{(1)}, A^{(2)}$ that are both complete and both natural in the sense of Section 2 — naturality holding for all models and all g — such that $A^{(1)}[F(c,\cdot); b, x] \neq A^{(2)}[F(c,\cdot); b, x]$ for some c, b, x.

(b) Construction and witness. On $SE(3)$, write $\Delta = b^{-1}x$, $\omega = \log R_\Delta$ and $v = R_\Delta^{-1} t_\Delta$. The family $\gamma^{(2)}$ that applies $\exp(\omega, 0)$ and then $\exp(0, v)$ depends on the endpoints only through Δ, hence is left-covariant and yields an equivariant, complete rule; its body velocity is piecewise constant but not constant, so by (a) it is not the screw path. That two *paths* differ does not by itself make two *attributions* differ, so we close the claim with an explicit witness in the abelian case. Take $G = \mathbb{R}^2$ acting by translation, $F(c, q) = (q_1 - c_1)(q_2 - c_2)$ — genuinely G-invariant as a model, with $f(x, y) = xy$ at $c = 0$ — baseline $(0, 0)$ and input $(1, 1)$. The straight path gives $A_1 = (\frac{1}{2}, \frac{1}{2})$; the axis-ordered path gives $A_2 = (0, 1)$. Both are left-covariant and both are complete — the components of each sum to $f(1, 1) - f(0, 0) = 1$ — and they disagree. Naturality holds numerically to 2×10^{-16}, while the naive fixed-f reading fails by $O(1)$, as it must (Appendix A).

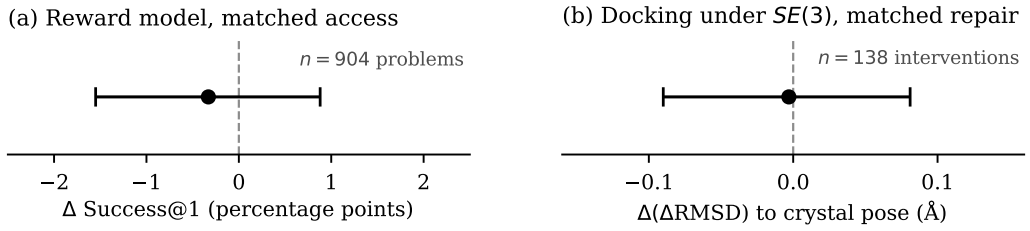


Figure 1: Effect of the attribution on a realized, model-external outcome in the two matched comparisons; point estimate and 95% CI. **(a)** Reward model, 904 preregistered confirmatory problems. Adding four attribution-shape features to an otherwise identical policy — same candidate bank, same candidate scores, same feasible action set, same policy class — changes the selected restart step on 208 problems; on the 202 of those resolved by blind human adjudication the two arms reach the same outcome 171 times, the attribution arm wins 14 and loses 17. **(b)** Docking, 138 held-out interventions. Repairing integrated gradients removes the frame defect completely ($57.8\% \to 0\%$ conclusion flips under a world-frame rotation) and changes the selected degree of freedom on 126 interventions, of which 61 end closer to the crystal pose and 65 further away. The outcome is a verified gold answer in (a) and crystal-pose RMSD in (b); in both panels the interval contains zero.

On the docking scorer of Section 5 (a 12-complex subset $\times$ 4 draws), the screw and rotate-then-translate rules have equivariance defects 3.4×10^{-12} and 4.3×10^{-12} under a random world-frame change, a median relative difference of 0.86, and rank a different degree of freedom first on 60.4% of inputs.

To complement the constructive witness, we compare against attribution in a PCA canonical frame, an independently defined equivariant canonicalization rule. In the experiment of Section 5, it and the group-path rule are frame-exact to machine precision yet recommend different degrees of freedom on 31.9% of interventions. Their realized effects differ by $|\Delta\text{RMSD}_{\text{screw}} - \Delta\text{RMSD}_{\text{canon}}| = 0.086$ Å on average (95% CI $[0.041, 0.137]$), carried entirely by the interventions on which the two rules disagree, since interventions on which they agree contribute exactly zero. The canonicalizer is a different object from the witness of Proposition 1, and it is equivariant only when the canonical frame is chosen equivariantly (Appendix A). Canonical frames also fail on a substantial fraction of inputs: canonicalize-then-IG flips for 75% of shapes under a 0.75π frame change on 200 ModelNet40 test shapes (Appendix F).

Naturality constrains how an attribution transforms, and completeness constrains how its components sum, but neither selects a unique rule (Appendix I). Choosing among admissible rules therefore requires an additional criterion.

5 What the repair buys, and what a finer attribution buys

Soundness and decision value are distinct evaluation targets. We therefore use an explanation-to-intervention protocol: a rule recommends one intervention, the intervention is executed, and we measure the realized effect. Because such comparisons are easy to confound, we state the access structure of every arm before the results (Table 1); Figure 1 summarizes both matched comparisons. Three smaller controlled regimes — a force field, point ablation on an invariant DeepSets network, and a policy shortcut probe — point the same way and are reported in Appendices B–D; all use synthetic or randomly-initialized models, and in each the native arm is structurally favoured, which is why the two experiments below were designed to remove that advantage.

5.1 Does the repair itself improve the intervention? Docking under $SE(3)$

This is the cleanest available test of the repair, because the repaired and unrepaired rules differ in nothing else. On 47 real protein–ligand complexes (crystal ligand and PDB pocket), poses are parameterised as $SE(3) \times \mathbb{T}^m$ relative to the crystal conformer; complexes are split by sorted index into a development half (24) and a pre-designated held-out half (23), opened once. A decoy perturbs

Table 1: Access structure. In docking only the native candidate-rescoring arm observes scorer values of candidate outcomes; the repaired-versus-naive comparison is matched on every column. In the reward model the prefix rule's queries are a *subset* of the Shapley arm's. No arm observes the realized outcome, the verifier or the annotated error before choosing.

Arm	What it queries before choosing	Queries	Outcome first?
$SE(3)$ docking — choose one DOF; step size and move sign shared by all arms			
Naive IG	scorer along a path from a random baseline pose	32 steps	no
Repaired (group-path) IG	*identical* scorer, baseline and budget	32 steps	no
Canonicalize-then-IG	scorer in a PCA canonical frame	32 steps	no
Native gradient	scorer gradient at the decoy	1 bwd	no
Native candidate rescoring	scorer at *every* candidate move	$\approx$38	no
Reward model — choose a restart step j; candidate set identical across arms			
Exact Shapley	reward on all 2^T coalitions of the original solution	70.1	no
Native prefix rule	reward on the $T+1$ prefixes of the original solution	6.3	no

the crystal pose along *exactly one* uniformly drawn degree of freedom, with the step normalised per DOF so each produces the same RMSD displacement δ_0 (mean 0.85 Å to the crystal). Every arm names one DOF; the sign of the move is then fixed for all arms alike by two scorer queries, so the comparison is only about *which* DOF. The scorer is a differentiable implementation of the AutoDock Vina functional — the chance-calibration measurement of Section 3.1 used a learned scorer on the same complex library — and the outcome is ΔRMSD to the crystal, which is geometric and independent of the scorer. Gates were fixed first: over the 282 gate interventions across all 47 complexes, naive IG flips its conclusion under a frame rotation on 57.8% of cases while both frame-correct arms flip on 0.00%; and the held-out half supplies $n = 138 \geq 120$ interventions. Statistics are clustered bootstrap over complexes (5000), with a preregistered minimum effect of 8 pp.

Repaired versus naive IG. Against naive IG — same scorer, same random baseline pose, same 32-step budget, same step and sign rule — the repaired rule changes localization by $\Delta\text{Loc@1} = -1.4$ pp, CI $[-8.0, +5.1]$, and realized outcome by $\Delta(\Delta\text{RMSD}) = -0.003$ Å, CI $[-0.090, +0.081]$. We repaired exactly the defect equivariance was designed to fix, under an otherwise identical protocol, and the downstream intervention did not improve (Table 2).

Table 2: Docking arms on the 138 held-out interventions. The matched comparison is naive IG against repaired IG: same scorer, baseline pose, step budget and sign rule. Repairing the frame defect moves neither localization nor realized outcome. The native counterfactual policy is shown as a reference point only; it queries the scorer on every candidate move and so is not access-matched (Table 1). All arms and the step-size robustness check are in Table 7.

Arm	Loc@1	ΔRMSD (Å)	Success@1
Naive IG	10.1%	-0.261	20.3%
Repaired (group-path) IG	8.7%	-0.264	21.7%
Random	4.3%	-0.269	14.5%
Native counterfactual (unequal access)	$\mathbf{72.5}\%$	$\mathbf{+0.590}$	$\mathbf{74.6}\%$
chance $(\overline{1/d})$	5.9%		

That average is not diluted by cases where the repair did nothing. Two rules that pick the same degree of freedom execute the same intervention, so their paired difference is zero by construction; here the repair changed the chosen degree of freedom on 126 of 138 interventions (91.3%). Restricted to those, the realized difference is -0.003 Å (CI $[-0.096, +0.087]$), with the repaired rule better on 61 and worse on 65. The near-zero average is therefore not explained by unchanged decisions: the repaired rule selected a different intervention in 91.3% of cases, and those changes were

nearly balanced between better and worse realized outcomes. The same analysis applies to the PCA comparison in Section 4: the group-path rule and the PCA-canonicalized rule pick different degrees of freedom on 31.9% of interventions, and on that subset neither is the better choice (22 against 22, -0.049 Å, CI $[-0.212, +0.122]$).

Secondary comparator, with unequal access. The native candidate-rescoring policy is the only arm that improves the pose at all (ΔRMSD $= +0.59$ Å, Success 74.6%, against -0.26 Å and 21.7% for the repaired rule; -52.9 pp, CI $[-63.0, -42.0]$). That comparison is *not* access-matched: it queries the scorer on all $2(6+m)$ candidate moves, while the attribution arms decompose the current pose once (Table 1). We report it as a reference point for what a querying policy achieves at comparable cost, not as evidence about attribution quality. The cheapest native readout — one backward pass — localises the faulty DOF on 21.0% against $\leq 10.1\%$ for the 32-step IG arms (chance 5.9%).

Experimental scope. Both IG arms sit near the floor, so the finding is that the repair does not rescue a method with little signal on this task, not that it could not help a stronger one; the repaired rule is also not distinguishable from picking a DOF at random ($+4.3$ pp, CI $[-1.4, +10.1]$). The scorer is a physics functional rather than a learned model, decoys carry a single injected fault, and one prespecified fault-type balance criterion was narrowly missed, so we report this study as supporting rather than confirmatory evidence (Appendix J).

5.2 Does a finer-than-native attribution act better? A trained reward model

The docking test asks about the repair. This one asks about resolution: what happens when the attribution provides a *step-level refinement whose sum equals* the native scalar. A trained 7B reward model r scores step-wise solutions to competition mathematics; with $n = r(y) - r(\varnothing)$ and exact step-level Shapley values ϕ_t, efficiency gives $\sum_t \phi_t = n$ as an algebraic identity with no quadrature, so $A = (\phi_1, \ldots, \phi_T)$ contains the native scalar by construction while keeping step-level resolution, and is invariant to the reward gauge. The intervention is discrete — regenerate the solution from step j onward — and success is judged by an external symbolic verifier against the benchmark's gold answer. Preregistered gates passed before any outcome was computed: completeness residual 1.3×10^{-14}, gauge invariance 0, first-error position spread $32/38/30\%$, and $n = 154 \geq 120$ actionable problems on the untouched confirmatory split (377 problems, 26,424 coalitions). On a separate audit of 1,000 items of the source benchmark, agreement between our symbolic verifier and the benchmark's own correctness labels is 98.9% $(901/911)$ on the items from which an answer could be extracted, extraction succeeding on $911/1{,}000$; extraction failures count as failed interventions for every arm alike (Appendix H).

Result. Shapley improves error localization over chance, but this localization advantage does not translate into a higher intervention success rate under the corresponding restart policy (Table 3). It identifies the human-annotated first error on 32.1% of problems against a 21.2% chance rate and 17.5% for random selection ($+14.6$ pp, CI $[+8.5, +20.7]$), replicating a development-split observation specified prospectively. Selecting the restart point by $\arg\min_t \phi_t$, however, succeeds on 37.7% of actionable problems against 57.1% for the native rule ($\Delta = -19.5$ pp, CI $[-29.9, -9.1]$), at $11.2\times$ the scoring forwards. The gap is conservative with respect to generation truncation, which harms the native arm more: restricting to non-truncated candidates moves it from -19.5 to roughly -31 pp. A trivial restart-from-scratch policy reaches 56.5%, indistinguishable from the native rule.

Comparison scope. This comparison differs in the action-selection rule and therefore does not isolate the marginal value of attribution features. It is not, however, an access asymmetry. The native rule scores *prefixes of the original response*, $\pi_n(y) = \arg\max_j [r(y_{1:j}) - r(y_{1:T})]$, and those prefix evaluations are a subset of the coalition evaluations exact Shapley already requires: the attribution arm has strictly more model-query access, at $11.2\times$ the cost, and neither arm observes a regenerated candidate, the verifier or the annotated error before choosing (Table 1). The direct test, which holds the action rule fixed as well, is reported next.

Matched policy comparison, development split. We hold the candidate bank, the candidate reward scores $c_{ij} = r(T_j(y)) - r(y)$, the feasible action set, the policy class and the action rule

Table 3: Attribution resolution versus intervention utility on the reward-model stress test. Exact Shapley localizes the annotated first error above chance, but its corresponding restart policy does not improve Success@1 over the native counterfactual policy. This stress-test comparison uses different action-selection rules; the matched marginal-value test follows below. Shapley minus native on Success@1: -19.5 pp (95% CI $[-29.9, -9.1]$); Shapley minus random on Loc@1: $+14.6$ pp ($[+8.5, +20.7]$). The full table, with the preregistered Shapley rule and a trivial restart policy, is Table 5.

Policy	**Loc@1** (n=377)	**Success@1** (n=154)	**scoring cost**
Exact Shapley $\arg\min_t \phi_t$	$\mathbf{32.1}\%$	37.7%	70.1 forwards
Native counterfactual policy	29.4%	$\mathbf{57.1}\%$	6.3 forwards
Random	17.5%	37.0%	0
chance ($\overline{1/T}$)	21.2%		

fixed, and vary only whether four attribution features enter the policy. Both arms are the same zero-initialised residual ranker over a shared candidate set, $\pi_\bullet = \arg\max_{j \in J^{\text{valid}}} [R_c(j) + g_\bullet(\cdot)]$ with R_c the within-problem rank of the candidate score c_{ij} and J^{valid} the feasible action set of Appendix M, so $g = 0$ reproduces the native policy exactly; the native arm receives $(R_c,\ j/T,\ n)$ and the attribution arm those plus A_{ij}^{shape}. Two design requirements follow from the setting. First, because exact Shapley satisfies $\sum_t \phi_t = n$, any attribution feature that retains the completeness direction can implicitly reintroduce the native scalar and invalidate a marginal-value comparison. We therefore give n to both arms and restrict the additional block to features invariant under $\phi \mapsto \alpha\phi + \beta\mathbf{1}$ ($\alpha > 0$), the group that moves n while fixing the across-step shape, verified by a preregistered gate (maximum block change 2.4×10^{-13} while the draws move $|n|$ by a median of 18.6; Appendix M). This is not a proof of independence — shape and level may covary — but the block cannot encode n through the completeness degree of freedom. Second, candidates whose regeneration was truncated or failed are removed from *both* arms' action sets; this is generation metadata known before any policy chooses, unlike gold correctness.

Out-of-fold on the 377 development problems, adding the attribution features changes realised Success@1 by -0.27 pp (95% CI $[-1.86, +1.33]$; McNemar $b = 4$, $c = 5$; policies disagree on 18.0% of problems). No gain is detected, and the data bound any gain to at most $+1.33$ pp. A dimension-matched control that adds four random features to the same architecture is instead significantly *harmful* (-2.39 pp, CI $[-4.24, -0.80]$). The two controls were not tested against each other, so we read this only as a development diagnostic: the random-feature control was more harmful on this split, which suggests the attribution features do not behave like arbitrary added dimensions, but it is not evidence of incremental attribution information. The best policy of all was the parameter-free $\arg\max_j c_{ij}$ (37.1% against 36.3% for the learned native arm; these are over all 377 problems, not the actionable subset of Table 5) — on this task, fitting a residual at all costs accuracy, and the symmetric design makes both arms pay that cost equally rather than removing it.

A separate matched-access confirmatory study. The 377 problems used above — the confirmatory split of the stress test — serve as the *development* split for this one, which has its own confirmatory pool of different problems. We ran it on 904 problems from MATH, OlympiadBench and Omni-MATH, with the policies fitted only on development and every selection fixed before outcome adjudication. The two policies chose different candidates on 208 problems (23.0%), which bounds $|\Delta|$ at 23.0 pp by construction; the remaining problems contribute exactly zero to the paired difference, so only the 406 selected candidates of the 203 discordant problems with a usable gold answer required an outcome (the remaining 5 enter worst-case bounds). Those were judged by two independent annotators who saw the problem, the candidate solution and the gold answer and nothing else — no arm label, no reward score, no attribution, no policy identity — after each passed a qualification gate of 30 development items with known labels. Disagreements and abstentions went to a third adjudicator under the same blinding.

Adding the attribution features changed realised Success@1 by -0.33 pp (95% CI $[-1.55, +0.88]$; McNemar $b = 14$, $c = 17$, exact $p = 0.720$). No gain was detected, and the data bound any positive gain to at most 0.88 pp at the 95% confidence level. 6 problems could not be resolved —

5 because the benchmark supplies no usable gold answer and 1 because the adjudication did not settle it — and they are kept in the denominator through worst-case bounds of $[-1.00, +0.33]$ pp rather than dropped. Per source the estimate was $+0.00$ pp on Omni-MATH ($n = 456$), $+0.27$ pp on OlympiadBench ($n = 374$) and -5.41 pp on MATH ($n = 74$); the MATH subset is the one the policies were fitted on — the problems are disjoint, the source benchmark is not — and is small, so we read it only as an in-distribution sanity check. Annotator agreement was $\kappa = 0.90$; 10 items were disagreements and 11 involved an abstention, so $21/406$ (5.2%) went to adjudication.

In this matched-access comparison, the candidate bank, candidate scores, feasible action set, policy class, training data and action rule are held fixed; only four attribution-shape features vary. The result bounds what those features are worth to this decision, on one reward model and one task family, under the access conditions that could be equalised.

6 The structural condition for native-readout sufficiency

Native-readout sufficiency requires a stronger structural condition than frame-correctness. Proposition 2 makes that condition explicit.

> **Proposition 2.** Suppose **(H1)** the G-invariant model f computes through local descriptors $h = (h_1, \ldots, h_N)$ that form a *complete* G-invariant — only completeness is used below, the factorization of f being what makes a native per-element readout exist at all. Let A be a per-element attribution that is itself G-invariant, with h and A Borel. Then $A = \Lambda(h)$ for a measurable Λ: A carries no frame-invariant information about x beyond h. If moreover **(H2)** $\sigma(n) = \sigma(h)$, then $\sigma(A) \subseteq \sigma(n)$, so the native readout is sufficient for A. Here $\sigma(\cdot)$ is the σ-algebra generated by a map.

Under (H1), any invariant attribution factors through the complete invariant descriptor h; the argument is the elementary one and we claim no novelty for it (Appendix E). Sufficiency of the scalar native readout n, however, additionally requires (H2), $\sigma(n) = \sigma(h)$. That stronger condition does not hold in the three settings examined here (Appendix K): summing a complete per-token attribution returns the identifiable reward difference but is not determined by it, while a per-atom energy and a per-atom attribution are separate functions of the same local descriptor, as are an invariant operator feature and a per-location saliency. Xu et al. (2026) likewise report attribution separating inputs a forward representation does not. The bound is therefore by h, not by n, which is why Section 5 measures the value of the extra resolution rather than settling it by assumption.

7 Related work and scope

Symmetry-respecting explanation. Crabbé and van der Schaar (2023) formalize that a faithful explanation of a G-invariant model must respect G; SITE (Wang and Wang, 2021) learns transformation-equivariant interpretations; geometry-aware IG variants (Zaher et al., 2024; Costanza and Simpson, 2025) and canonicalization (Dym et al., 2024) supply frame-correct constructions. These establish the view we adopt. We study what they do not establish: rule selection and realized decision value. Section 6 gives a separate boundary on native sufficiency.

Non-uniqueness of path attributions. That different integration paths give different credit is classical, and Sundararajan et al. (2017) single out integrated gradients among path methods by an additional symmetry-preservation axiom. Our result is the group-level analogue and is not implied by it: we show that requiring *naturality under* G — the property equivariant attribution is built to satisfy — still leaves a family with more than one member, exhibit that family on a Lie group, and give two admissible members that differ. Xu et al. (2026) prove a complementary obstruction: input automorphisms can make a minimal symmetry-respecting *report* non-canonical. Theirs requires a nontrivial automorphism of the input and concerns which report to surface; ours concerns which *rule* to use and holds away from such ties. Work on this question has appeared quickly through 2026, so we make no general priority claim: to our knowledge the rule-level non-identification statement, within this naturality-constrained path-attribution setting, has not been made elsewhere, and that is the only distinction we assert (Appendix G).

Native readouts and intrinsic alternatives. Jia et al. (2026) state a native-readout hypothesis and audit Attention Rollout, Grad-CAM and RISE by deletion AUC; we adopt their term. Their bound is empirical, resting on which operators a forward pass happens to expose, whereas Proposition 2 is conditional on a symmetry hypothesis; their setting has no symmetry component; and deletion AUC is a faithfulness proxy where we measure a realized intervention effect. The broader post-hoc versus intrinsic debate (Amorim et al., 2023; Moshkovitz et al., 2026) supplies the surrounding frame. Orgad et al. (2026) argue interpretability is valuable when actionable; we adopt that bar and report what passes it and what does not.

Task-grounded evaluation of explanations. A separate line of work evaluates explanations by their utility for downstream tasks and has reported mixed results, including several cases in which explanations do not improve the target decision. Bilodeau et al. (2024) prove that a complete, linear attribution can fail to improve on random guessing for inferring model behaviour, including identifying a spurious feature; Adebayo et al. (2022) find post hoc explanations ineffective at revealing a model's reliance on a spurious signal that is unknown at test time, and prone to indicating such a reliance where there is none; SimEvals (Chen et al., 2022) replace the human study with a simulated decision task and use it to identify which explanation methods are informative for a given downstream use; and ROAR (Hooker et al., 2019) finds that, after removing the inputs a method ranks highest and retraining, several popular methods are no more damaging than a random ranking. The same tradition also records successes: Pahde et al. (2023) use explanations to localize a spurious artefact in medical imaging models and then revise the model, and Pahde et al. (2025) build the same detect-then-mitigate loop into a safety framework; both measure the effect of acting on the explanation rather than the explanation itself. Our decision-value experiments belong to this tradition and differ in two ways. The impossibility of Bilodeau et al. (2024) constrains what a complete linear attribution can reveal; our non-identification result concerns which complete *equivariant* rule is used, a degree of freedom that is introduced by the symmetry requirement rather than restricted by it. And the symmetry structure supplies a well-defined native readout to compare against, so the baseline is a quantity the model already computes rather than a random or degraded ranking.

Limitations. (i) The probes of Section 3 use randomly-initialized networks. (ii) The docking study uses a physics functional rather than a learned model and single-fault decoys, and it misses one preregistered balance gate, so we report it as preregistered supporting rather than confirmatory evidence. (iii) The reward experiments use one task family and one intervention operator; the initial stress test compares decision rules, while the separate matched-access study isolates the contribution of four Shapley-shape features within one policy class (Section 5.2). (iv) We examine three domains; Proposition 2 delimits generalization rather than establishing it.

8 Conclusion

Frame-correctness is necessary for explanations under representation symmetries, but it does not determine a unique attribution rule. We show this formally for path attributions and empirically on $SE(3)$, where two frame-exact rules often select different degrees of freedom. We then test whether these differences matter for action. In docking, removing frame-dependent decision flips changes most selected interventions without improving pose quality. In a matched reward-model study, adding Shapley-derived shape features likewise yields no measurable Success@1 gain. Whether a native readout could have sufficed instead is a scope condition rather than a result: it requires a structural hypothesis that failed in every domain we examined (Section 6). These results motivate evaluating three questions separately: whether an attribution is well defined under representation changes, how a rule is selected among admissible alternatives, and whether using it improves the downstream outcome.

AI use statement

Large language models assisted with experiment implementation, analysis tooling, and drafting and revising the manuscript. They were not used as annotators or outcome evaluators. The authors verified the experimental results, mathematical claims and bibliographic information, and take responsibility for the final manuscript.

ETHICS STATEMENT

The adjudication described in Appendix M used three adult human annotators, who judged whether the final answer of a model-generated solution was mathematically equivalent to a public benchmark's published gold answer. The task involved no personal, sensitive or identifying data: annotators saw only a benchmark problem, a candidate solution and the gold answer, and were blinded to arm identity, reward scores and attribution values. No information about the annotators is collected or reported beyond aggregate agreement rates, and no new dataset is released. All models, benchmarks and protein–ligand structures used here are publicly available and are evaluated as published.

REPRODUCIBILITY STATEMENT

Propositions 1 and 2 are stated in Sections 4 and 6; their hypotheses, proofs, the explicit $\mathbb{R}^2$ witness and the conditions under which each hypothesis fails are given in Appendices A and E. For the decision-value experiments, the access structure of every arm is stated in Table 1 before any result. The preregistered gates, data splits, sample sizes, intervention operator, external verifier and statistical procedure are given in Appendix H for the reward-model stress test, Appendix J for the docking study, and Appendix M for the matched-access study, which also gives the attribution feature definitions, the invariance gate and the blinded adjudication protocol.

REFERENCES

J. Adebayo, M. Muelly, H. Abelson, and B. Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *ICLR*, 2022. arXiv:2212.04629.

J. P. Amorim, P. H. Abreu, J. Santos, and H. Müller. Evaluating post-hoc interpretability with intrinsic interpretability. arXiv:2305.03002, 2023.

B. Bilodeau, N. Jaques, P. W. Koh, and B. Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024.

V. Chen, N. Johnson, N. Topin, G. Plumb, and A. Talwalkar. Use-case-grounded simulations for explanation evaluation. In *NeurIPS*, 2022. arXiv:2206.02256.

F. Costanza and L. Simpson. Riemannian integrated gradients: a geometric view of explainable AI. arXiv:2503.00892, 2025.

J. Crabbé and M. van der Schaar. Evaluating the robustness of interpretability methods through explanation invariance and equivariance. In *NeurIPS*, 2023. arXiv:2304.06715.

N. Dym, H. Lawrence, and J. W. Siegel. Equivariant frames and the impossibility of continuous canonicalization. In *ICML*, 2024. arXiv:2402.16077.

S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. In *NeurIPS*, 2019. arXiv:1806.10758.

S. Jia, Z. Zhao, and K. Li. Architecture-aware explanation auditing for industrial visual inspection. arXiv:2605.14255, 2026.

M. Moshkovitz, S. Srinivas, L. Semenova, N. Frost, C. Rashtchian, V. Boreiko, S. Zhang, H. Lakkaraju, C. Rudin, and J. Wortman Vaughan. Position: Explainability research must prioritize foundations over ad-hoc methods. In *ICML*, PMLR 306, 2026. arXiv:2607.14123.

H. Orgad, F. Barez, T. Haklay, I. Lee, M. Mosbach, A. Reusch, N. Saphra, B. Wallace, S. Wiegreffe, E. Wong, I. Tenney, and M. Geva. Interpretability can be actionable. In *ICML*, 2026. arXiv:2605.11161.

F. Pahde, M. Dreyer, W. Samek, and S. Lapuschkin. Reveal to revise: An explainable AI life cycle for iterative bias correction of deep models. In *MICCAI*, 2023. arXiv:2303.12641.

F. Pahde, T. Wiegand, S. Lapuschkin, and W. Samek. Ensuring medical AI safety: interpretability-driven detection and mitigation of spurious model behavior and associated data. *Machine Learning*, 114:206, 2025. doi:10.1007/s10994-025-06834-w.

P. Radiuk, O. Barmak, L. Bedratyuk, and I. Krak. Equivariant transition matrices for explainable deep learning: A Lie group linearization approach. *Machine Learning and Knowledge Extraction*, 8(4):92, 2026. doi:10.3390/make8040092.

J. Skalse, M. Farrugia-Roberts, S. Russell, A. Abate, and A. Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *ICML*, 2023. arXiv:2203.07475.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. arXiv:1703.01365.

K. M. ul Hassan and M. Ebrahimi. Signal or noise? Auditing rotation-induced saliency drift in medical and aerial imaging. arXiv:2609.02224, 2026.

Y. Wang and X. Wang. Self-interpretable model with transformation-equivariant interpretation (SITE). In *NeurIPS*, 2021. arXiv:2111.04927.

X. Xu, S. Tao, and K. Tan. Automorphism-induced non-canonicity in top-k explanations of graph neural networks. arXiv:2607.26344, 2026.

E. Zaher, M. Trzaskowski, Q. Nguyen, and F. Roosta. Manifold integrated gradients: Riemannian geometry for feature attribution. In *ICML*, 2024. arXiv:2405.09800.

A A SUFFICIENT CONSTRUCTION FOR EQUIVARIANT PATH ATTRIBUTIONS

We state the sufficient construction used in Section 4(a). Let G be a Lie group with algebra $\mathfrak{g}$, and let $f : G \to \mathbb{R}$. Call a **body path-attribution** any $\psi_k(f) = \int_0^1 \eta_k(\alpha)\, g_k(\gamma(\alpha))\, d\alpha$ obtained by integrating the body-frame derivative g along a smooth path γ from R' to R with body velocity η (that is, $\gamma' = \gamma\hat{\eta}$).

Completeness is automatic. $\sum_k \psi_k = f(R) - f(R')$ for every such ψ, directly by the chain rule. No axiom is being imposed here; the property holds for the whole class.[1]

(a) Left-covariance is sufficient for naturality. A change of world frame moves the *model* as well as the endpoints, so the property to ask for is naturality, not invariance of a fixed f. (Requiring $f(gQ) = f(Q)$ for all g would force f constant, since G acts transitively on itself.) Write $f^g(Q) := f(g^{-1}Q)$ — in the docking setting, where the score also depends on a context c, this is $f_{gc}(gR) = f_c(R)$ with c suppressed. Call ψ **natural** — frame-correct with $\rho = \mathrm{id}$, which for body-frame components is invariance of the component vector — if

$$\Psi[f^g;\, gR', gR] \;=\; \Psi[f;\, R', R] \qquad \text{for all } f, g.$$

Left-covariance of the path family, $\gamma_{gR',gR} = g\,\gamma_{R',R}$, is *sufficient* for naturality for all f: the body-frame derivative of f^g at $g\gamma$ equals that of f at γ, so the two integrals agree term by term. We use only this direction; we do not claim that left-covariance is necessary, and Proposition 1 does not require it.

(b) Constant body velocity selects the representative. Off the cut locus of $\exp$ (for $SO(3)$, $\|\log(R'^{-1}R)\| < \pi$), γ has constant body velocity — η independent of α — iff it is the screw path $\gamma(\alpha) = R' \exp(\alpha\xi)$ with $\xi = \log(R'^{-1}R)$ on the principal branch: constant η makes $\gamma' = \gamma\hat{\eta}$ a linear ODE whose unique solution with $\gamma(0) = R'$ is $R' \exp(\alpha\hat{\eta})$, and $\gamma(1) = R$ then forces $\hat{\eta} = \log(R'^{-1}R)$, which is single-valued off the cut locus. The resulting ψ is the group-path integral, and it is invariant to monotone reparametrization, so its value depends on the path *image* rather than the parametrization.

[1] Completeness is automatic *analytically*; its finite-step numerical realization is not, which is why the reward-model study of Section 5.2 uses exact Shapley values, whose completeness is algebraic rather than quadrature-dependent (Appendix H).

(c) A second left-covariant family: rotate, then translate. The witness used in Section 4(b). Write $\Delta = R'^{-1}R$ and, on $SE(3)$, $\Delta = (R_\Delta, t_\Delta)$. Put $\omega = \log R_\Delta$ and $v = R_\Delta^{-1} t_\Delta$; both are functions of Δ alone, and $\exp(\omega, 0)\exp(0, v) = (R_\Delta, R_\Delta v) = \Delta$. Define

$$\gamma^{(2)}_{R',R}(\alpha) = R'\, q_\Delta(\alpha), \qquad q_\Delta(\alpha) = \begin{cases} \exp\big(2\alpha\,(\omega, 0)\big), & \alpha \le \tfrac{1}{2}, \\ \exp(\omega, 0)\,\exp\big((2\alpha - 1)(0, v)\big), & \alpha > \tfrac{1}{2}, \end{cases}$$

so $q_\Delta(0) = e$ and $q_\Delta(1) = \Delta$. Because q depends on R', R only through Δ, and $(gR')^{-1}(gR) = \Delta$, we get $\gamma^{(2)}_{gR',gR} = g\,\gamma^{(2)}_{R',R}$: the family is left-covariant, so by (a) the induced attribution is natural for all f, and completeness holds by the chain rule. Its body velocity is $(2\omega, 0)$ then $(0, 2v)$ — not constant — so by (b) it is not the screw path. (Smoothing the junction changes nothing: the two velocities differ whenever $\omega \neq 0 \neq v$.)

(d) Closing the existence claim with an explicit witness. That the two *paths* differ does not by itself make the two *attributions* differ. The two paths do induce distinct per-axis signed trace measures $\mu_k^{(1)} \neq \mu_k^{(2)}$, and $\psi_k(f) = \int g_k \,\mathrm{d}\mu_k$ is linear in the body derivative g; but distinct signed trace measures alone do not establish distinct attributions, because the body derivative must be the exact differential of a scalar field, so separating the measures by an arbitrary smooth g need not produce an admissible f. The claim is therefore closed by an explicit witness, in the abelian case where everything is computable.

Take $G = \mathbb{R}^2$ acting by translation, baseline $(0, 0)$, target $(1, 1)$. To make the model symmetry explicit, introduce a context $c \in \mathbb{R}^2$ and let $F(c, q) = (q_1 - c_1)(q_2 - c_2)$. Then $F(c + g, q + g) = F(c, q)$, so F is a genuinely G-invariant model in the sense of the proposition; fixing $c = 0$ gives the scalar field $f(x, y) = xy$ used below, and the naturality shift f^g is exactly the context shift $F(g, \cdot)$. Both families below depend on the endpoints only through the displacement Δ, so each is left-covariant and hence natural by (a), and each is complete by the chain rule: the components of each sum to $f(1, 1) - f(0, 0) = 1$. Naturality here is the condition of (a) — the function travels with the endpoints, $\Psi[f^g;\, g{+}b,\, g{+}p] = \Psi[f;\, b, p]$ — and this example shows why that is the right definition: $f(x, y) = xy$ is emphatically *not* translation-invariant, and demanding $f(g\cdot) = f(\cdot)$ would leave only constants, for which every attribution rule agrees trivially. We check both readings numerically: naturality holds to 2×10^{-16}, while the naive reading fails by $O(1)$, as it must.

- **Straight path** $\gamma^{(1)}(\alpha) = (\alpha, \alpha)$, velocity $(1, 1)$. Along it $\partial_x f = y = \alpha$ and $\partial_y f = x = \alpha$, so $A^{(1)} = \big(\int_0^1 \alpha \,\mathrm{d}\alpha,\ \int_0^1 \alpha \,\mathrm{d}\alpha\big) = (\tfrac{1}{2}, \tfrac{1}{2})$.
- **Axis-ordered path** (x first, then y). On the first leg $\partial_x f = y = 0$, so it contributes nothing; on the second $\partial_y f = x = 1$. Hence $A^{(2)} = (0, 1)$.

$A^{(1)} = (\tfrac{1}{2}, \tfrac{1}{2}) \neq (0, 1) = A^{(2)}$, with both rules natural and both complete. This proves Proposition 1. The $SE(3)$ pair of (b)–(c) — screw versus rotate-then-translate — is the geometrically relevant realization of the same phenomenon on the group of interest here; that pair is separated numerically rather than analytically, and the numerical check below finds the two rules disagree on the top-ranked axis in 60.4% of cases, so the separation is not delicate in practice.

The chart-straight path is not left-covariant. It is tempting to take A_2 to be the straight line in *chart* coordinates, $\xi(\alpha) = (1 - \alpha)\log R' + \alpha \log R$: its body velocity is $\mathrm{dexp}_{-\xi(\alpha)}(\xi')$, which is α-dependent off the abelian case, so it is certainly not the screw path. But that family is **not left-covariant**, because $\log(gR) \neq \log g + \log R$ in general, and hence

$$\exp\big((1 - \alpha)\log(gR') + \alpha \log(gR)\big) \;\neq\; g \exp\big((1 - \alpha)\log R' + \alpha \log R\big).$$

Such a rule is not equivariant, so it cannot witness Proposition 1. The distinction is practical as well as formal: canonicalize-then-attribute is frame-dependent unless the canonicalizer itself is equivariant. The arm of Section 5 therefore uses an *equivariantly derived* chart (principal axes of the placed ligand, with the axis sign set by a third moment), which makes it frame-exact; the chart-straight path is not equivariant on its own.

Numerical check of the witness. On the docking scorer of Section 5 (12 complexes $\times$ 4 draws), the screw and rotate-then-translate rules have equivariance defects $\max|\phi - \phi_g| = 3.4 \times 10^{-12}$

and 4.3×10^{-12} under a random world frame change, while their relative difference has median $\|\phi_1 - \phi_2\|/\|\phi_1\| = 0.86$ and their top-ranked degree of freedom differs on 60.4% of inputs.

Equivariance constrains the *family* of admissible paths; it does not select within that family. A second axiom is needed for that, and the one commonly adopted — constant body velocity — is a further path-selection criterion, logically independent of equivariance and completeness. The raw chart-straight construction is *not* left-covariant and hence is not a member of the admissible family at all. The witness for Proposition 1 is the explicit abelian pair of (d); the rotate-then-translate family of (c) supplies the geometrically relevant $SE(3)$ realization, separated numerically rather than analytically.

B First-order regime: force field

The first of the three smaller controlled regimes: an $SE(3)$-invariant pair potential, where the native object is the force $-\nabla E$ and the intervention is a small atomic displacement. Figure 2 reports all four axes.

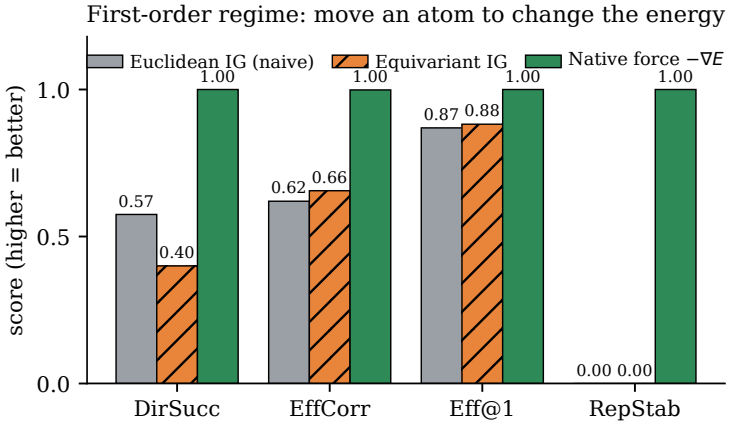


Figure 2: First-order regime. In this regime the native force $-\nabla E$ leads both attribution methods on all four axes. The recommended-move angle under rotation (not shown) is $0°$ for the native force vs. $96°/100°$ for naive/equivariant IG — both are directionally frame-unstable.

C Steelman: discrete point-ablation

Figure 3 shows the adversarial steelman summarized in Section 5: point ablation on an invariant DeepSets network, where removing a point is non-local and the native object is not the ground truth by construction.

D Third domination: policy shortcut detection

Figure 4 shows the study summarized in the main text: whether the frame-instability of attribution detects a defect that a plain prediction-level rotation probe misses. It does not.

E Proof and scope of Proposition 2

Proof. By (H1) the map $x \mapsto h(x)$ separates G-orbits, so every G-invariant function of x factors through h. A is G-invariant by assumption, hence $A = \Lambda(h)$ for some Λ, measurable because h and A are Borel. If in addition $\sigma(n) = \sigma(h)$, then $\Lambda(h)$ is $\sigma(n)$-measurable, giving $\sigma(A) \subseteq \sigma(n)$. □

This is the standard fact that an invariant function of a complete invariant factors through it, and we claim no novelty for it. The content is in the hypotheses.

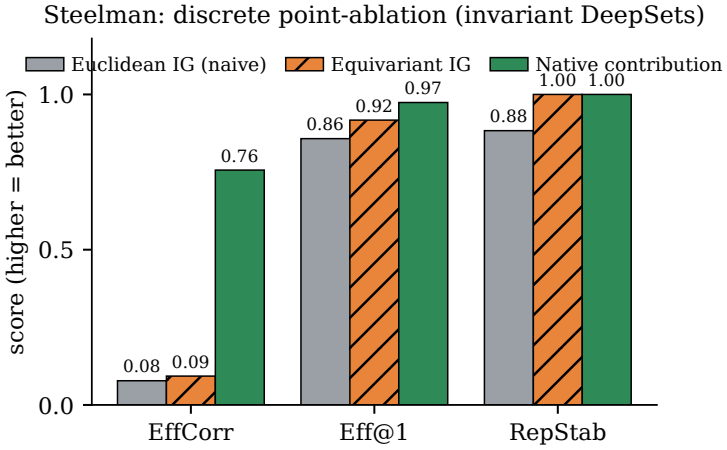


Figure 3: Steelman (discrete point-ablation, invariant DeepSets, SO(3)). Both IG-family scores have low correlation with the discrete effect (EffCorr ≈ 0.08), compared with 0.76 for the native decomposition.

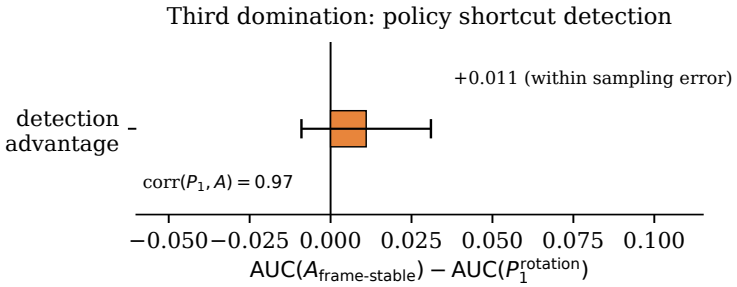


Figure 4: Third, independent domination (policy shortcut detection). The advantage of attribution's frame-instability over a plain prediction-level rotation probe is $+0.011$ AUC — indistinguishable from zero — while the two signals correlate at 0.97. Frame-stable attribution is a repackaging of the model's own derivative.

Neither hypothesis is superfluous. For (H1), take $f(x) = \int_G u(gx)\,\mathrm{d}g$ for an arbitrary u: the model is exactly G-invariant, but nothing forces a per-element factorization, so no native per-element readout need exist and the conclusion is unavailable. For (H2), whenever n is a strictly coarser function of h than h itself, A may separate inputs that n does not; the proposition then bounds A by h alone.

Where the hypotheses hold in this paper. (H1) holds in all three domains examined here, and (H2) fails in all three — but for two different reasons, which the Table 4 separates them: in one domain the attribution refines the native readout, and in the other two the two objects are incomparable. Each entry shows that A is not a function of n; combined with $\sigma(A) \subseteq \sigma(h)$ from the first part of the proposition, that forces $\sigma(n) \neq \sigma(h)$.

(H2) therefore fails throughout, and in two of the three cases the failure is not even a matter of n being coarser — the two objects answer different questions about the same descriptors. Choosing a readout that satisfies (H2) after seeing A would make the structural claim unfalsifiable, so the readout is held fixed. The consequence is stated in Section 6: the structural claim is the bound by h, and the comparison against n is empirical.

Table 4: Where the hypotheses of Proposition 2 hold in the three domains examined here. (H1) holds throughout; (H2) fails throughout, for two distinct reasons.

Domain	**(H1) complete local invariants**	**(H2) $\sigma(n) = \sigma(h)$, and the relation that does hold**
MD force fields	yes — invariant local descriptors	**no**; *incomparable*: E_i and a per-atom attribution are separate functions of the same h_i, with no reduction either way
Reward models	yes — identifiability fixes the learnable content	**no**; *attribution refines native*: summing a complete per-token attribution returns the per-pair difference, which does not determine it
Neural operators	yes — resolution-invariance is definitional	**no**; *incomparable*: the invariant feature and a per-location saliency are separate functions of h

F Inside the corner: which object survives

The 75% figure of Section 4 is a *generic* rate, not a hard-corner one: all 200 sampled ModelNet40 test shapes fall below the sign-stability threshold $\min|\text{skewness}| \leq 0.3$, so on this data near-symmetry is a property of the sample rather than a filter applied to it.

Separately, on an $SE(3)$-invariant model under a *physical* conjugation, the per-DOF attribution moves an order of magnitude more than the invariant subspace-energy summary. That experiment retrains its scorer per run and discards poses with vanishing attribution, and its spread is correspondingly wide: over 8 runs the ratio has median $11.3\times$ with range 8.6–$22.1\times$ (7–12 surviving poses). We therefore claim only the direction and not the magnitude. If it holds, the frame-robust deliverable in the one place equivariant attribution has an edge over canonicalization is again a coarse readout of the kind the model already exposes.

G Relation to automorphism-induced non-canonicity

Xu et al. (2026) prove that when a graph carries a nontrivial automorphism and no minimal explanation is fixed by it, no rule can be single-valued, minimal and symmetry-respecting at once; they mechanise the criterion and find it binding for 24% of molecules with interchangeable substituents. Their obstruction and ours point the same way — symmetry correctness does not hand you a unique explanation — but neither subsumes the other.

Theirs concerns which *report* to surface when the input's stabilizer forces ties among equally-scored elements; the obstruction requires a nontrivial automorphism of the input. Ours concerns which *rule* to use and holds away from such ties: the witness in Proposition 1 requires no nontrivial input automorphism and no score tie — it is built on an abelian group. We therefore make no priority claim for the general observation that symmetry fails to determine an explanation. To our knowledge the rule-level non-identification statement has not been made within this naturality-constrained path-attribution setting; given how quickly work on symmetry and explanation appeared through 2026, we state that as the scope of the distinction rather than as a claim about the literature as a whole.

One of their findings bears directly on the boundary above: they report that attribution can separate inputs the forward representation does not. That is one reason we state the boundary of Section 6 conditionally, as a bound by the descriptors h under (H1), rather than claiming the native readout is always sufficient.

H Refinement stress tests on a trained reward model

The smaller controlled regimes of Appendices B–D structurally favour the native arm: the first-order intervention is aligned with the native gradient, and the point-ablation model is additive. We there-

fore evaluate decision value in a preregistered setting with neither property — a trained transformer reward model, a discrete intervention, and an outcome external to the model.

Setting. A trained transformer reward model scores solutions to mathematics problems. The native readout is the scalar $n = r(y) - r(b)$ for a fixed anchor b, and the attribution is a step-level decomposition whose components sum to n, so it is a refinement of the native scalar rather than a quantity that merely differs from it. Candidate interventions are discrete (regenerate the solution from step j onward), identical across arms, and constructed without reference to either arm's scores. The outcome is external to the reward model: whether the regenerated solution reaches the ground-truth answer. Items come from public benchmarks of model-generated solutions with human-annotated first-error steps.

Validity of the setting. Two checks were fixed before any arm was computed. The intervention operator is meaningful: regenerating from the annotated first error succeeds far more often than regenerating one step later ($+55.6$ percentage points, 95% CI $[+38.9, +72.2]$, paired over problems), while the unedited solutions succeed 0% of the time. And the reward model has a strong native readout, which only a negative control reveals: judged on genuinely improved answers alone, a general preference model and a mathematics-specific model look alike (85.9% vs. 88.0%), but adding the control separates them ($\Delta = +11.4$ vs. $+58.7$ points; $\mathrm{AUC} = 0.629$ vs. 0.897).

Choice of exact Shapley. Treating the attribution as a refinement of the native scalar requires completeness to hold numerically, not only analytically. We use exact Shapley values because their efficiency identity gives $\sum_{t=1}^{T} \phi_t = v(N) - v(\varnothing) = n$ algebraically, with no quadrature and therefore no integration error. They are also invariant under the reward gauge $r'(x, y) = r(x, y) + \delta(x)$, since δ cancels in every coalition marginal, which matches the identifiability structure we cite as the native object in this domain (Skalse et al., 2023).

Design. We use the mathematics-specific 7B reward model, which exhibited the stronger native readout in the validity check above. Coalition value $v(S) = r(x, y_S)$ with y_S the solution restricted to steps in S in original order; $T \leq 8$ so 2^T is exact. Items are benchmark problems with human-annotated first-error steps. Development split: 189 problems / $10{,}492$ coalitions. Confirmatory split, held out until protocol finalization: 377 problems / $26{,}424$ coalitions. The intervention operator and the external verifier are those described above.

Validity gates, all fixed before any outcome. (G1) completeness $|\sum_t \phi_t - n|/|n| = 1.3 \times 10^{-14}$; (G2) gauge invariance $\max |\phi' - \phi| = 0$ over three shifts δ; (G3) actionable problems $n = 154 \geq 120$ on the confirmatory split; (G4) first-error position spread $32/38/30\%$ across thirds. Two further gates cover the external verifier for this split: a 1:1 join to the source benchmark ($377/377$) and blind agreement of our symbolic checker with the benchmark's own correctness labels. The comparison covers all 1000 items, correct ones included — scoring only incorrect items would let a checker that always answers "wrong" score perfectly. Agreement is 98.9% ($901/911$) on the items from which a final answer could be extracted, and extraction succeeded on $911/1000$ (91.1%). The preregistration set the threshold at $\geq 98\%$ agreement without stating whether extraction failures enter the denominator, so this is a conditional rate: the 89 extraction failures were not constrained by the gate. Two handling rules were fixed before any outcome was computed. An extraction failure counts as a failed intervention ($R = 0$) for every policy alike; and we additionally report the comparison restricted to the problems on which the checker could judge every candidate ($n = 72$), which gives $\Delta = -33.3$ pp (CI $[-47.3, -18.1]$), the same direction as the headline.

Result. Table 5 gives the four axes. The attribution carries genuine localization signal, replicating a development-split observation that we specified prospectively before opening the confirmatory split. It does not improve on the native counterfactual policy at localization, and it is decisively worse at action while costing $11.2\times$ more scoring forwards.

Effect of generation truncation. Regeneration is capped at 512 new tokens; a truncated generation essentially never succeeds (1.2% versus 44.1%). Because earlier restart points must write more, truncation hits the native arms hardest — selected-j truncation rates are 6.5% for Shapley, 20.8%

Table 5: Trained-transformer refinement stress test, confirmatory split; full version of Table 3. ΔSuccess@1 (Shapley minus native) $= -19.5$ pp, CI $[-29.9, -9.1]$; Shapley vs. random on Loc@1 $+14.6$ pp, CI $[+8.5, +20.7]$; Shapley vs. native on Loc@1 $+2.7$ pp, CI $[-3.7, +8.8]$.

	Loc@1 (n=377)	**Succ@1** (n=154)	**CostReg.**	**sel. compute**
Exact Shapley $\arg\min_t \phi_t$	32.1%	37.7%	0	$2^T = 70.1$
Native counterfactual policy	29.4%	**57.1**%	0	$T+1 = 6.3$
Preregistered Shapley rule s_A	28.4%	48.7%	—	$2^T = 70.1$
Trivial restart ($j = 0$)	—	56.5%	125	0
Random	17.5%	37.0%	—	0
chance ($\overline{1/T}$)	21.2%			

for the native policy and 28.6% for trivial restart — so the bias runs *in favour of* the attribution. Restricting to non-truncated candidates widens the gap rather than closing it: 40.3% / 71.3% / 79.1%, i.e. ΔSuccess@1 moves from -19.5 to roughly -31 pp.

Interpretation and limitations. CostRegret is descriptive only: the success sets differ across arms, so conditioning on success makes the comparison a selection effect rather than an efficiency result. We also do not claim the two arms have equal information access — the native policy issues a query per candidate while the attribution summarizes the original response once — only that under the stated protocol the attribution is both the more expensive arm and the less useful one. Finally, the preregistered rule $s_A(j) = -\sum_{t \geq j} \phi_t$ was at chance on the development split (20.6% against 21.2%) but above chance on the confirmatory split (28.4%, $+10.9$ pp over random). We report that divergence rather than resolving it; it is the reason we report both the preregistered rule and the prospectively specified $\arg\min_t \phi_t$ rather than only the better one.

I AXIOM VERDICTS

Table 6 separates, for each construction that delivers frame-invariance, which of its stated axioms is free (a consequence of the construction), which is a substantive requirement, and which is a convention that selects one admissible rule among many. The third column is where the non-identification of Section 4 enters: in every row it is non-empty.

Table 6: Axiom verdicts across three constructions that deliver frame-invariance. None of these frame-invariance axioms, by itself, establishes explanatory fidelity or downstream utility, so in no case does satisfying them justify the representative they force.

Construction	**Axioms offered for frame-invariance**	**Verdict**
Group-path IG	completeness; left-equivariance; constant body velocity	free (chain rule); requirement; convention
Canonicalization	invariance via a canonical section of $\mathcal{X} \to \mathcal{X}/G$	requirement; the section is a convention — and no continuous canonical choice exists in general (Dym et al., 2024), which is why weighted frames were introduced
Manifold / Riemannian IG	path is a geodesic of a chosen metric; invariance under that metric's isometries	convention (which metric); requirement (isometry invariance)

J REAL-COMPLEX GEOMETRIC ACTIONABILITY

Preregistered before any arm was computed. 47 real protein–ligand complexes (crystal ligand + PDB pocket) — the same library as Section 3.1; poses are parameterised as $SE(3) \times \mathbb{T}^m$ relative to the crystal conformer, so the crystal is $(T, \theta) = (I, 0)$ and RMSD to it is exact. Complexes are split

by sorted index into a development half (24) and a pre-designated held-out half (23); the held-out half was opened once.

Design. A decoy perturbs *one* uniformly drawn DOF, with the step size normalised per DOF so that each produces the same RMSD displacement δ_0 (mean 0.85 Å to the crystal). Perturbing several DOFs at once makes the task degenerate — a policy that always moves a translation DOF, using no explanation at all, then recovers 70–88% of the oracle improvement — whereas under the single-DOF decoy trivial policies recover a *negative* fraction of the oracle. Every arm names one DOF; the sign of the move is then fixed for all arms alike by two score queries, so the comparison is only about *which* DOF. Statistics are clustered bootstrap over complexes (5000), MIE 8 pp.

Validity checks, fixed before the held-out half was opened. These are computed over all 282 gate interventions across the 47 complexes, development and held-out together. Naive IG flips its which-DOF conclusion under a world-frame rotation on 57.8% of interventions, while Repaired IG and canonicalize-then-attribute both flip on 0.00%; the held-out half supplies $n = 138 \geq 120$ interventions; and no trivial policy reaches the oracle ceiling. On the held-out half, under the exact-DOF Loc@1 rule used for the arms in Table 7, the always-translate policy scores 4.3% against a 5.9% chance rate ($\overline{1/d}$ over the $6 + m$ degrees of freedom).

Held-out results. Table 7 gives the five preregistered arms, the steelman arm described under Robustness below, and two trivial policies, on the 138 held-out interventions. The native counterfactual policy is the only arm that improves the pose. Repairing the frame defect leaves the outcome where it was: Repaired IG against naive IG is $\Delta\mathrm{Loc@1} = -1.4$ pp, CI $[-8.0, +5.1]$, and $\Delta(\Delta\mathrm{RMSD}) = -0.003$ Å, CI $[-0.090, +0.081]$, under an identical scorer, baseline pose, step budget and sign rule.

Table 7: Pre-designated held-out half: 138 interventions over 23 complexes, mean decoy RMSD 0.85 Å; full version of Table 2. Only the native counterfactual policy improves the pose.

arm	Loc@1	ΔRMSD (Å)	% oracle	Success@1	selection compute
Native counterfactual	**72.5%**	**+0.590**	69.5%	**74.6%**	38 forward
Native gradient	21.0%	−0.104	−12.2%	32.6%	**1 backward**
Naive IG	10.1%	−0.261	−30.7%	20.3%	32 steps
Repaired IG	8.7%	−0.264	−31.0%	21.7%	32 steps
Canonicalize-then-IG	8.0%	−0.279	−32.9%	20.3%	32 steps
Repaired IG, 8 baselines	2.9%	−0.325	−38.3%	20.3%	8 × 32 steps
Random	4.3%	−0.269	−31.7%	14.5%	0
Always translate	4.3%	−0.332	−39.1%	14.5%	0
chance	5.9%				

Robustness. The ordering holds at all three step sizes ($0.5/1.0/1.5\,\delta_0$): the native policy is the only arm with a positive mean ΔRMSD at each ($+0.278/ + 0.590/ + 0.161$ Å, against $-0.042/ - 0.264/ - 0.661$ for Repaired IG). Integrated gradients need a baseline pose that the task does not supply, whereas the native arms need none; the crystal is excluded as a baseline because it is the answer. A steelman arm averaging eight random baselines, fixed *before* the held-out half was opened, rules out a single unlucky baseline as the cause: at 2.9% it is the weakest arm in the table, below chance, at eight times the cost. The conclusion also does not rest on the mix of injected fault types. Success@1(Repaired IG) − Success@1(native policy) is -45.0 pp on translation faults (n=20), -66.7 pp on rotation (n=30) and -50.0 pp on torsion (n=88), the same sign and a comparable size in every stratum; the native policy's Loc@1 is $90.0/83.3/64.8\%$ across the three. One stratum-level detail explains the modest aggregate for the native *gradient* arm: it localises translation faults on 80.0% of cases but rotation faults on 3.3%.

Limitations. A prespecified fault-type balance criterion was narrowly missed — translation faults are 14.9% of the 282 gate interventions against a threshold of $\geq 15\%$ — so we report this study as supporting rather than confirmatory evidence; the stratified analysis above gives the same qualitative conclusion in every fault type. The scorer is a published functional, not a learned model, so we make

no claim here about learned geometric models. The single-fault decoy is simpler than a real docking error, and RMSD is not symmetry-corrected. Finally, the native counterfactual policy issues one query per candidate while the attribution decomposes the pose once, so those two arms do not have the same query structure — though the attribution is the more expensive one. The repaired-versus-naive comparison, which is the test this appendix is built around, is matched on every column.

K STRUCTURAL SCOPE ACROSS THREE DOMAINS

Table 8 records, for each of the three domains this paper touches, the invariant object an equivariant attribution is built to produce and the invariant object the model already exposes. The two need not have the same granularity, which is why Proposition 2 separates the bound by a complete invariant descriptor from sufficiency of the scalar native readout, and why the comparison in Section 5 is empirical rather than structural.

Table 8: In each domain the invariant substrate equivariant attribution targets already exists natively — though the native object and the attribution need not have the same granularity. What an equivariant construction contributes in these domains is therefore a characterization of invariant information the domain already has, rather than a new source of it.

Domain	**Proposed equivariant object**	**Native object that already exists**
MD force fields (rotation)	rotation-invariant per-atom importance	atomic energy decomposition $E = \sum_i E_i$ — the core ansatz of essentially every NNP; per-atom/pairwise/monomer decompositions are standard tools
Reward models (reward gauge)	per-token attribution robust to the reward's arbitrary zero	known partial-identifiability: reward is fixed only up to a prompt-shift $\delta(x)$; the only learnable information is the reward *difference* of two responses to the same prompt; Contrastive-IG supplies the machinery
Neural operators (grid)	discretization-invariant saliency	resolution/discretization-invariance is the *defining* property of neural operators; resolution-invariant features are native

L CHANCE-CALIBRATED FRAME INSTABILITY

A flip rate has no interpretation on its own, because a rule that names one of $6 + m$ degrees of freedom disagrees with itself most of the time by construction. The reference line is the flip rate of a rule that re-draws its conclusion independently from the same per-complex marginal over degrees of freedom. Table 9 reports the observed and chance rates together, with $\kappa = (p_{\text{chance}} - p_{\text{obs}})/p_{\text{chance}}$ as the resulting frame-consistency coefficient, at five magnitudes of world-frame rotation.

Table 9: The wedge, calibrated. Naive Euclidean IG's which-DOF conclusion does degrade under a world-frame change, but never approaches chance: even under a complete reorientation it retains substantial frame-consistent signal ($\kappa = 0.23$, bootstrap CI over complexes excludes 0). The equivariant construction is exactly 0% at every magnitude — but the defect it removes leaves substantial above-chance frame-consistent signal in place. 564 decisions per row.

world-frame change	observed flip	chance flip	κ (frame-consistency)
0.10π	23.6%	85.2%	0.72 [0.69, 0.76]
0.25π	41.1%	87.1%	0.53 [0.48, 0.58]
0.50π	56.6%	88.2%	0.36 [0.32, 0.40]
0.75π	65.2%	88.5%	0.26 [0.22, 0.30]
1.00π	68.6%	89.0%	0.23 [0.19, 0.27]

M MATCHED-ACCESS STUDY: DESIGN AND CONFIRMATORY RESULTS

Matched-access motivation. The reward-model comparison of Section 5.2 differs in the action rule as well as the features, so it cannot isolate the marginal value of attribution. This study fixes everything but the features.

Attribution feature block and confound control. Exact Shapley satisfies $\sum_t \phi_t = n$ to machine precision (largest deviation 5.3×10^{-15} over 377 problems), so a suffix-sum feature equals the native scalar at $j = 0$. Handing the attribution arm $\sum_{t \geq j} \phi_t$ while withholding n from the native arm would therefore smuggle a native readout into the attribution block, and the effect is large enough to decide the experiment on its own: with n withheld, the pooled candidate-level AUC gain is $+0.0543$ (95% CI $[+0.0298, +0.0831]$); with n given to the native arm, where it belongs, the same quantity is $+0.0036$ ($[-0.0042, +0.0106]$). Centring does not fix it either: $\tilde{S}_j = \sum_{t \geq j} \phi_t - \frac{T-j}{T} n$ removes the mean but not the scale, leaving $\mathrm{corr}(\tilde{S}, n) = +0.50$ and $R^2(n \mid A) = 0.31$. Both arms therefore receive n, and the attribution block is standardised within the problem: $z_j = (\phi_j - n/T)/s_\phi$ with s_ϕ the standard deviation of $\phi_1, \ldots, \phi_T$, and use $A^{\mathrm{shape}} = (z_j,\ \mathrm{rank}(-\phi)_j,\ \sum_{t \geq j} z_t,\ D_j)$ with $D_0 = 0$ and $D_{j \geq 1} = (\phi_j - \phi_{j-1})/s_\phi$; s_ϕ itself is withheld because it correlates with $|n|$.

Gate G-S1. Every feature is invariant to $\phi \mapsto \alpha\phi + \beta\mathbf{1}$ for $\alpha > 0$, the group that moves n while fixing the shape. The gate draws 64 random (α, β) per problem and requires $\|A^{\mathrm{shape}}(\alpha\phi + \beta\mathbf{1}) - A^{\mathrm{shape}}(\phi)\|_\infty < 10^{-9}$ while those draws actually move n. Measured: block change 2.4×10^{-13}, median $|n|$ displacement 18.6, at most 8.3×10^{-5} of draws leaving n fixed. The gate is not vacuous: the alternative definition $D_0 = \phi_0$, which puts a *level* into a shape block, violates it by 2.0×10^{1}. Invariance does not make the block independent of n in the empirical distribution; shape and level may covary, and we do not residualise that away, because doing so would bias toward the null we predict.

Shared feasible action set. Both policies select from $J_i^{\mathrm{valid}} = \{j : \neg\text{truncated} \wedge \neg\text{gen_fail}\}$, which excludes truncated or failed generations using metadata available before either policy chooses. This prevents restart-position differences from inducing unequal token-budget exposure across arms: unconstrained, the two policies pick systematically different restart points and the budget converts that difference into an outcome difference (mean restart index 3.68 versus 1.97, with 7.0% versus 37.4% of selected candidates truncated), so a negative Δ could measure a budget artefact rather than a feature effect. Under the shared set the arms are balanced — mean restart index 2.45, 2.47 and 2.48, with 0% truncated selections by construction. The mask is *analysis-defined* rather than a set of impossible actions, and it does not destroy the task: on development, truncated candidates were still correct 1 time in 398 (0.25%) against 19.4% for the rest, and the number of problems with at least one correct candidate is 154 of 377 with or without it.

Development result. Out-of-fold (problem-wise 5-fold $\times$ 5 seeds, denominator 377, problems with no valid candidate counted as failures for every arm): Success@1 is 0.3634 (native), 0.3607 (native + attribution), 0.3395 (native + four random features), against 0.3714 for the parameter-free $\arg\max_j c_{ij}$. Paired bootstrap over problems (10,000): $\Delta_{\mathrm{attr}} = -0.27$ pp $[-1.86, +1.33]$ and $\Delta_{\mathrm{noise}} = -2.39$ pp $[-4.24, -0.80]$. The McNemar off-diagonals are small (b=4, c=5 and b=1, c=10), so this is a thin result and we do not treat it as a single strong piece of evidence.

Confirmatory evaluation. The preregistration, the selections and the inputs were hashed before any outcome was computed. Of 904 problems: 3 have no valid candidate, 21 have exactly one and are forced to the same choice, 672 are concordant with identical candidate-text hashes, 203 are discordant and adjudicable, and 5 are discordant with no usable gold answer and enter worst-case bounds rather than being deleted. The deterministic bound $|\Delta| \leq D/N$ is 23.0 pp, and the minimum detectable effect at 80% power was 4.5 pp before the outcomes were opened.

Adjudication as executed. Two annotators judged all 406 selected candidates of the 203 adjudicable discordant problems independently, after each passed a qualification gate of 30 development items with verified labels (93.3% and 96.7%; the adjudicator scored 100%). They agreed on 385 items (94.8%, $\kappa = 0.90$); the 21 remaining — 10 disagreements and 11 abstentions — went to a

blinded third adjudicator who saw neither prior verdict. Twenty items were shown twice to each annotator as a blinding-integrity check and were answered consistently $19/20$ and $20/20$. Of 202 resolved discordant pairs, 171 had the same outcome in both arms, 14 favoured the attribution arm and 17 the native arm. Six problems remain unresolved and are carried through worst-case bounds of $[-1.00, +0.33]$ pp.

Automatic verifiers are not a substitute here. On a preregistered single run they reach sensitivity 80.7% on OlympiadBench and 91.3% with specificity 94.0% on Omni-MATH, the sources of 830 of the 904 problems. Those errors are arm-independent and would attenuate Δ toward zero, which is the direction of the conclusion this paper reports, so using them as the primary instrument would make the upper bound on a gain look tighter than the evidence supports.

Outcome definition. Following the rest of the paper, an item counts as a success when the candidate's final answer is mathematically equivalent to the benchmark's gold answer; the annotators were instructed not to check the reasoning and not to second-guess the gold. A wrong gold, or a wrong solution that lands on the right value, therefore counts the same way for both arms. Such errors are arm-independent and add noise to Δ without giving it a direction.