DECAF ☕: DECOMPOSITION OF EVIDENCE, CONTRADICTION, AND FRAGILITY IN PERTURBATION RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Perturbation methods explain model decisions by measuring prediction changes under altered inputs, but response magnitude tells us only how much a model reacts, not what that reaction means. The same magnitude can support the final factual–counterfactual difference, oppose it, or arise strongly along the perturbation path yet vanish at the endpoint. We therefore track how the contrast develops as paired inputs are progressively revealed, using the final contrast to interpret the trajectory. We introduce DECAF *(*D*ecomposition of* E*vidence,* C*ontradiction,* A*nd* F*ragility)*, which routes aligned, opposed, and endpoint-null responses into evidence E, contradiction C, and fragility F. The decomposition preserves ordinary magnitude exactly, $\mathrm{Abs} = E + C + F$, and is unique under endpoint-relative axioms. Across controlled vision and tabular settings, the three components track independently measured behavior. In a 72-model ImageNet-9 audit, we compare cases with nearly identical response magnitude but different independently measured behaviors. The largest DECAF component agrees with an observed behavior in 96.4% of cases, compared with 35.0% for magnitude alone. Changing only the reveal path increases total response by nearly 80%, yet evidence barely changes while fragility grows by more than $4\times$. On FunnyBirds and ImageNet-1k, short forward-only DECAF trajectories outperform the tested general-purpose attribution baselines. On a 1B-scale DINOv2 model, a short trajectory matches a strong gradient-based baseline with $4.75\times$ lower wall time and $2.36\times$ lower peak memory.

Code: https://anonymous.4open.science/r/decaf-1457

1 INTRODUCTION

Post-hoc explanation asks what aspects of an input matter to a model's prediction. A natural way to answer this question is contrastively: compare the prediction with what would happen if a feature, region, concept, or context were changed. Perturbation-based and counterfactual methods instantiate this comparison by constructing alternative inputs and measuring the resulting change in model output (Fong & Vedaldi, 2017; Dhurandhar et al., 2018a; Goyal et al., 2019b; Covert et al., 2021a). This has the basic form of an effect measurement: change a factor, compare two responses, and summarize the difference. In practice, that response is often reduced to a scalar magnitude.

A magnitude answers an important question: *how much did the model react?* It leaves a second question open: *what did the reaction mean?* Consider an input and a paired counterfactual that changes one factor. Their prediction difference measures how strongly the model responds to that change. This endpoint comparison, however, shows only the net effect after inputs are fully observed. It does not show how the contrast emerges as information becomes available. We therefore consider a paired reveal: the paired inputs are progressively revealed along matched trajectories, so that their prediction difference can be observed at the same level of revealed information.

As the pair is revealed from an uninformative state, the contrast can evolve in different ways. It may align with the final effect, oppose it, or become large even when the fully observed pair differs little. These distinct behaviors can have identical absolute responses. Treating these responses as equivalent therefore hides behavior that matters for interpretation.

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.

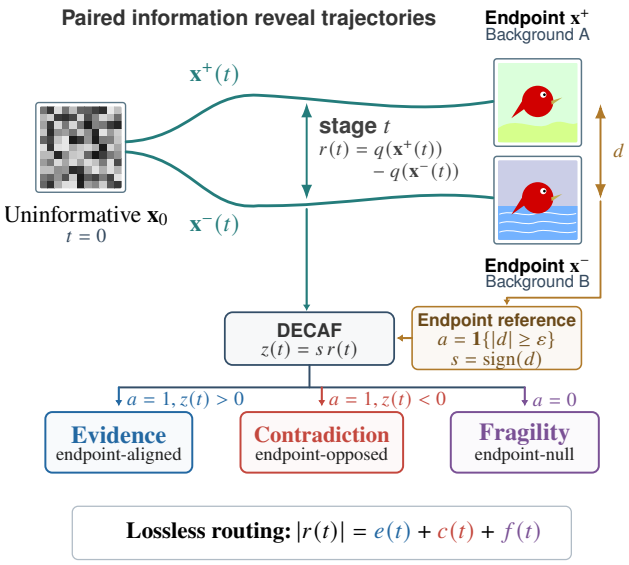


Figure 1: DECAF at a glance. The paired reveal measures a signed stage response, while the clean endpoint supplies an activity gate and orientation. DECAF routes the same response into endpoint-aligned evidence, endpoint-opposed contradiction, or endpoint-null fragility, with exact conservation. The compact summary reports the two headline consequences: mechanism recovery after ordinary magnitude is matched, and substantially more stable cross-protocol diagnostics.

We introduce DECAF in Figure 1, short for D*ecomposition of* E*vidence,* C*ontradiction,* A*nd* F*ragility*. DECAF resolves this ambiguity by using the final response between the fully observed pair as a semantic reference for the entire reveal process. Its magnitude determines whether the factor has a meaningful final effect, while its sign provides the reference direction. Intermediate responses that agree with this direction are routed to *evidence*, while responses that oppose it are routed to *contradiction*. If the fully observed pair shows little response, there is no meaningful final effect to align with, and intermediate responses are routed to *fragility*. We call this *semantic routing*: the observed responses are assigned different roles according to their relation to the final contrast, without changing the trajectories themselves. The routing is lossless, so evidence, contradiction, and fragility sum exactly to the ordinary response magnitude. **DECAF is model-agnostic**, therefore explaining any black-box model that returns a scalar score.

Relation to prior work. Gradient and path methods assign contributions to input coordinates or internal units (Simonyan et al., 2014; Selvaraju et al., 2017; Sundararajan et al., 2017; Shrikumar et al., 2017; Bach et al., 2015). Removal and sampling methods measure prediction changes under masking or replacement (Fong & Vedaldi, 2017; Fong et al., 2019; Petsiuk et al., 2018; Covert et al., 2021b). Counterfactual and causal approaches define meaningful contrasts (Goyal et al., 2019b; Chattopadhyay et al., 2019; Janzing et al., 2020). Evaluation work shows that baselines, removal operators, and off-manifold inputs can alter explanation behavior (Adebayo et al., 2018; Hooker et al., 2019; Rong et al., 2022; Jethani et al., 2023). DECAF begins after a paired response has been measured. It refines the response by its relation to the clean endpoint. Unlike positive–negative feature attribution, it also contains an endpoint-null branch. Unlike amortized explainers such as FastSHAP, it requires no learned explanation model (Jethani et al., 2022). A detailed comparison with feature attribution, counterfactual explanation, evaluation frameworks, and efficient black-box methods appears in Appendix A.

We make five contributions:

- *Problem.* We identify a semantic gap in perturbation explanation: response magnitude tells us how strongly a model reacts, but not what that reaction means.
- *Observation interface.* We introduce paired reveal, which tracks a factual–counterfactual contrast along matched trajectories as information becomes available.
- *Algorithm.* We introduce DECAF, which uses the final contrast to route intermediate responses into evidence, contradiction, and fragility. The decomposition is lossless and unique under simple axioms, and adds no model queries beyond the paired trajectories.
- *Behavioral meaning.* Controlled vision and tabular experiments show that the three components track learned reliance, effect reversal, and fragility across neural, linear, and tree-based models.
- *Practical consequence.* On ImageNet-9, DECAF recovers response semantics after magnitude is matched and diagnoses what changes across perturbation paths. It also outperforms the tested general-purpose attribution baselines on FunnyBirds and ImageNet-1k and scales efficiently to a large DINOv2 vision transformer.

2 PAIRED INFORMATION REVEALING: FORMAL SETUP

We formalize paired information revealing. Let $q : \mathcal{X} \to \mathbb{R}$ denote the model score. An input $\mathbf{x}^+$ and a factor-level counterfactual $\mathbf{x}^-$ define the final contrast

$$d = q(\mathbf{x}^+) - q(\mathbf{x}^-).$$

The pair need not be pre-specified: in feature attribution, $\mathbf{x}^-$ can be induced from $\mathbf{x}^+$ by a removal or replacement operator.

A paired reveal starts from a common uninformative state $\mathbf{x}_0$ and produces matched trajectories $(\mathbf{x}^+(t), \mathbf{x}^-(t))$ for $t \in \mathcal{T} = [0, 1]$, with

$$\mathbf{x}^+(0) = \mathbf{x}^-(0) = \mathbf{x}_0, \qquad \mathbf{x}^+(1) = \mathbf{x}^+, \qquad \mathbf{x}^-(1) = \mathbf{x}^-.$$

The signed response is $r(t) = q(\mathbf{x}^+(t)) - q(\mathbf{x}^-(t))$, with $d = r(1)$. Ordinary magnitude $|r(t)|$ preserves size but discards its relation to the final contrast; Section 3 uses it to assign semantic roles.

3 DECAF: A SEMANTIC DECOMPOSITION OF PAIRED RESPONSES

Section 2 gives us two quantities: the response $r(t)$ at each reveal level and the final contrast $d = r(1)$. DECAF uses the final contrast only as a semantic reference for interpreting the intermediate responses. If the final contrast is substantial, its direction tells us which responses agree with the model's eventual behavior and which oppose it. If the final contrast is negligible, intermediate responses cannot be interpreted as supporting or opposing a meaningful final effect.

Choose a practical threshold $\epsilon > 0$ and define

$$a = \mathbf{1}\{|d| \geq \epsilon\}, \qquad s = \operatorname{sign}(d).$$

The gate a determines whether the final contrast is active, while s orients the trajectory when it is active. Let

$$z(t) = s\, r(t),$$

and write $z^+ = \max\{z, 0\}$ and $z^- = \max\{-z, 0\}$. DECAF routes the response as

$$(e(t), c(t), f(t)) = \big(az^+(t),\ az^-(t),\ (1-a)|r(t)|\big).$$

An active response aligned with the final contrast becomes *evidence*; an active response pointing in the opposite direction becomes *contradiction*; and a response on a final-contrast-null pair becomes *fragility*. This logic is illustrated earlier in Figure 1.

Let μ be a normalized measure over stages and let the outer expectation cover paired examples and protocol randomness. We report

$$M = \mathbb{E}|d|, \qquad (E, C, F) = \mathbb{E}\int_{\mathcal{T}} (e(t), c(t), f(t))\, \mathrm{d}\mu(t), \qquad \mathrm{Abs} = E + C + F. \tag{1}$$

The finite-grid implementation replaces the integral by quadrature. No additional model query is needed once the signed paired trajectory is available. See Appendices B and C for finite-grid estimation, conditional summaries, threshold sensitivity, and streaming accumulation.

Forward-only implementation. The model enters only through evaluations of q. DECAF needs no gradients, parameters, or internal activations. It can run against a neural network, a tree ensemble, or a remote score API. Examples, stages, factors, paths, and models are independent batching dimensions.

Table 1: Model calls per image in ImageNet-9.

Method	Fwd.	Bwd.	Internal	Learned
DECAF-9	18	**0**	**No**	**No**
Input×Grad	1	1	Yes	No
IG	16	16	Yes	No
SmoothGrad	16	16	Yes	No
BlurIG	12	12	Yes	No
Occlusion	49	0	No	No
RISE	256	0	No	No

Algorithm 1: DECAF in four steps.

Require: score q, paired path $(\mathbf{x}^+(t), \mathbf{x}^-(t))$, ε

$d \leftarrow q(\mathbf{x}^+) - q(\mathbf{x}^-)$; $a \leftarrow \mathbf{1}\{|d| \geq \varepsilon\}$; $s \leftarrow \operatorname{sign}(d)$

for all $t \in \mathcal{T}$ **do** ▷ $r(t=1) = d$ is cached and reused

  $r(t) \leftarrow q(\mathbf{x}^+(t)) - q(\mathbf{x}^-(t))$

  $(e, c, f) \leftarrow \big(a(sr)^+, a(sr)^-, (1-a)|r|\big)$

end for

return M, E, C, F by averaging over stages and examples

Scope of the semantics. DECAF assigns observable, endpoint-relative roles to paired responses; it does not infer the latent cause of a trajectory response. These response roles are defined relative to the score q, the chosen factual–counterfactual pair, the reveal trajectory, and the practical threshold ϵ. In particular, *fragility* denotes response on a pair whose final contrast is negligible under this specification. It does not by itself distinguish off-manifold artifacts, boundary uncertainty, calibration effects, or other possible causes. An active pair may still be path-sensitive, but its response is described through aligned and opposed mass rather than F.

4 Theoretical Foundations of Semantic Routing

Section 3 defines a simple routing rule from a stage response $r(t)$ and final contrast d to evidence, contradiction, and fragility. We first ask whether this routing is arbitrary.

To formalize this routing, we establish three operational axioms that reflect practical explanation goals. First, *conservation* ensures that no response magnitude is lost or invented. Second, *endpoint gating* isolates sensitivity on pairs whose final contrast is negligible under the chosen threshold. Third, *directional support* distinguishes intermediate responses that move toward the model's final decision from those that oppose it.

Theorem 1 (Unique hard-gated semantic routing)**.** *Fix $a \in \{0, 1\}$, endpoint orientation $s \in \{-1, 1\}$ on the active branch, and response $r \in \mathbb{R}$. The unique nonnegative triple satisfying **conservation**, $e + c + f = |r|$; **endpoint gating**, $f = 0$ for $a = 1$ and $e = c = 0$ for $a = 0$; and **directional support**, $c = 0$ when $sr \geq 0$ and $e = 0$ when $sr \leq 0$, is*

$$(e, c, f) = \big(a(sr)^+,\ a(sr)^-,\ (1 - a)|r|\big).$$

Theorem 1 removes tunable mixtures once the endpoint gate and orientation are specified. It establishes uniqueness within the stated endpoint-relative semantics, not uniqueness of hard endpoint gating itself. Under these operational axioms, the routing is uniquely determined. *See Appendix B for the proof and the positive–negative (Jordan) decomposition interpretation.*

Theorem 1 shows that the routing is unique once the endpoint-relative semantics are fixed. We next ask whether retaining the three routed components provides information beyond ordinary magnitude.

Theorem 2 (DECAF strictly refines response magnitude)**.** *Ordinary response magnitude is a deterministic function of the* DECAF *profile,*

$$\mathrm{Abs} = E + C + F,$$

so every decision rule based on Abs *can also be implemented from (E, C, F). The reverse recovery fails in general: for every $m > 0$, the distinct profiles $(m, 0, 0)$, $(0, m, 0)$, and $(0, 0, m)$ all produce $\mathrm{Abs} = m$. Hence the refinement is strict for any decision problem that distinguishes two such profiles with positive probability.*

Thus DECAF is a lossless refinement of ordinary magnitude: any analysis based on Abs can be reproduced from (E, C, F), while the reverse recovery is impossible in general. The refinement is strict whenever a decision problem distinguishes response profiles that share the same magnitude. For example, if evidence, contradiction, and fragility are equally likely and all produce magnitude m, every magnitude-only classifier receives the same observation and cannot exceed $1/3$ accuracy, whereas the full profile distinguishes the three roles. Section 6 tests this consequence empirically after explicitly matching ordinary response magnitude. See Appendix B for the Bayes limit, strict information refinement, and unequal priors.

Theorem 2 shows that magnitude can collapse distinct response roles. A practically important case is the difference between an effect that disappears and one that reverses. Both can reduce aligned evidence, but only reversal should create contradiction.

Suppose the final contrast is active with magnitude $m > 0$. Let $\eta \in [0, 1]$ denote the probability that this effect is altered by context. With probability $1 - \eta$, the response remains aligned with magnitude m; with probability η, it is either suppressed to zero (attenuation) or reversed to $-m$ (inversion). We consider three simple response regimes: *preservation*, where the effect keeps its direction; *attenuation*, where it weakens toward zero; and *inversion*, where it reverses direction.

Proposition 1 (Contradiction separates attenuation from inversion)**.** *Under equal aligned and opposed magnitudes,*

Regime	E	C	Abs
Preservation	m	0	m
Attenuation to zero	$(1-\eta)m$	0	$(1-\eta)m$
Inversion	$(1-\eta)m$	ηm	m

In the inversion regime, $C/(E+C) = \eta$.

Attenuation and inversion can remove the same amount of aligned evidence, but only inversion creates contradiction. Thus C distinguishes loss of an effect from reversal of that effect. The controlled experiment in Section 5. *See Appendix B for the proof, unequal magnitudes, and continuous mixtures.*

When contradiction and fragility vanish, conservation gives Abs $= E$. Thus DECAF reduces exactly to ordinary magnitude when the response is already fully aligned; it introduces additional distinctions only when the observed behavior contains them.

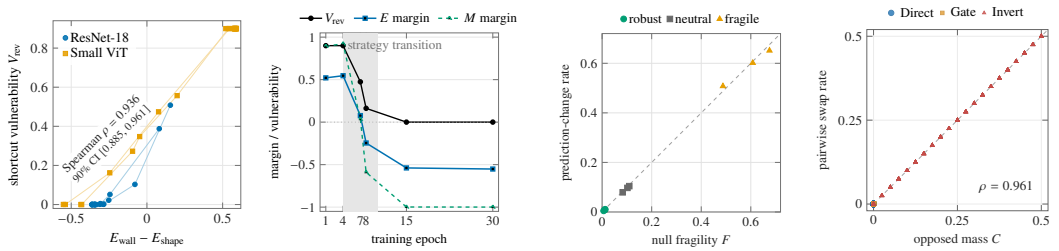
5 Controlled Diagnostics: Validating E, C, and F

The preceding theory defines endpoint-relative response roles. We now test whether these roles correspond to model behavior measured independently of DECAF. We begin with the 3D Shapes dataset, whose generative factors can be changed one at a time while the remaining factors are held fixed (Burgess et al., 2018). This gives exact factual–counterfactual pairs and lets us formulate three controlled learning problems in which reliance, off-path sensitivity, and effect reversal can be measured independently of DECAF. Across the controlled suite, we use *ResNet-18* and a small *ViT*. The base grid contains *30 trained models* and *180 model–factor units*, with additional dedicated checkpoints and training variants for the three behavioral tests. Comparing DECAF's decomposition with these independent behavioral measurements then tests whether each component carries the intended response semantics.

Evidence. We train object-shape classifiers in an environment where background wall color is correlated with the label, so models may rely on either this shortcut or the object itself. Reversing the wall–label correlation at test time gives an independent measure of shortcut reliance: a wall-dependent model loses accuracy, while a shape-dependent model remains stable. Across 52 training checkpoints, the evidence margin $E_{\text{wall}} - E_{\text{shape}}$ correlates 0.936 with this reversal vulnerability, with a 90% bootstrap interval of $[0.885, 0.961]$ (Figure 2(a)). Figure 2(b) shows the same relation during a within-model strategy transition; complete trajectories appear in Appendix E.

Fragility. We construct models for which floor color has little effect once the image is fully revealed, but different effects before full reveal. To create this difference, we expose models during training to partially revealed inputs while preserving the original shape task. In *fragile training*, the model is encouraged to react to changes in floor color at these partial states. In *robust training*, it is instead encouraged to remain unchanged, while *neutral training* uses only the original clean task without either objective. We then measure the resulting sensitivity independently using held-out floor-color interventions. If F captures endpoint-null path sensitivity, it should track this prediction-change rate. It does so closely (Figure 2(c)); intervention separation and cross-geometry checks appear in Appendix E.

Contradiction. We construct three tasks with the same object-color effect in the original wall context but different behavior when the wall context changes. In *Direct*, the effect is preserved; in *Gate*, it disappears; and in *Invert*, it reverses. The resulting label behavior—preservation, collapse, or swap—therefore provides an independent measure of what happened to the effect. If C captures opposition rather than mere weakening, it should remain near zero for Direct and Gate and increase only when the effect reverses. Across 30 models and all mismatch levels, C tracks the pairwise label-swap rate with $\rho = 0.961$, whereas ordinary magnitude does not ($\rho = -0.036$; Figure 2(d)). Detailed regime separation and calibration appear in Appendix E.



(a) Population-level reliance. **(b)** Within-model transition. **(c)** Endpoint-null sensitivity. **(d)** Label-level reversal.

Figure 2: The three response roles track independently measured model behavior. (a–b) Evidence follows shortcut reliance across checkpoints and through a within-model strategy transition. (c) Fragility follows off-path prediction change on endpoint-null pairs. (d) Contradiction follows pairwise label reversal. Additional intervention-separation, architecture-specific, and cross-geometry diagnostics appear in Appendix E.

6 Real-World Audit: Response Semantics on ImageNet-9

Section 5 established the meanings of E, C, and F in controlled learning problems. We now ask whether the same response semantics describe behavior on natural images, including models that were never trained for our audit. ImageNet-9 is useful for this purpose because its background variants keep the foreground object fixed while changing only the background (Xiao et al., 2021). This gives natural factual–counterfactual pairs in which we can ask whether a model's response to background change behaves as evidence, contradiction, or endpoint-null sensitivity.

Setup. We evaluate two complementary groups of models on these same background interventions. The first consists of 24 off-the-shelf ImageNet-1k classifiers. These models were trained independently of our experiment, so they test whether the response semantics established in Section 5 remain meaningful for existing natural-image models. Their 1,000-way predictions are aggregated into the nine ImageNet-9 superclasses before evaluation.

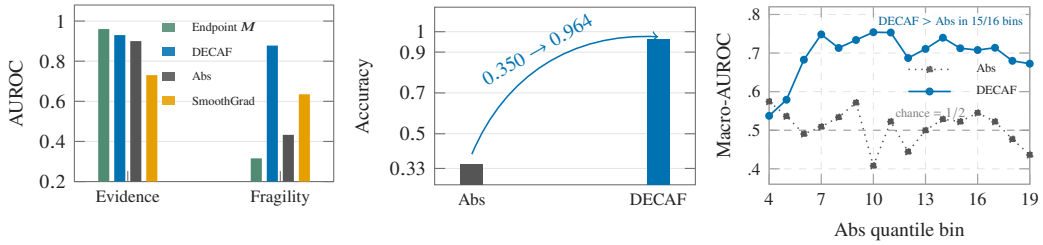
The second group contains 48 models trained directly on ImageNet-9. We fine-tune six backbones—*ResNet-50, ConvNeXt-Tiny, EfficientNet-B3, RegNetY-8GF, Swin-T, and ViT-B/16*—on four versions of the training data: the original images, images with randomly reassigned backgrounds, images with a fixed next-class background, and foreground-only images. These training conditions deliberately produce models with different degrees and directions of background dependence. This group therefore tests whether the same response semantics continue to track behavior as the model's learned use of background changes. Together, the two groups form a 72-model zoo: one provides models trained independently of our audit, while the other provides controlled diversity in the behavior being audited.

6.1 Response semantics survive natural images

The first question is whether the response roles validated in Section 5 remain connected to independently measured behavior on natural images. We consider two behaviors. Background reliance asks whether changing only the background substantially disrupts the model's prediction. Endpoint-null sensitivity asks whether a pair with little final background effect can nevertheless remain sensitive to held-out background corruptions. Both behaviors are defined independently of DECAF. If the response semantics transfer, E should identify the former and F the latter.

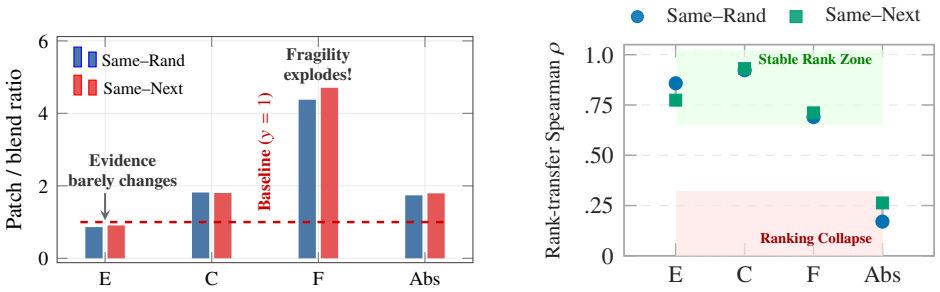
For background reliance, we mark a pair as positive when the model correctly classifies the image with a same-class background, but replacing that background with a random-class one either changes the predicted class or lowers the true-class probability by at least 0.20. We then ask whether these behavior-positive pairs tend to receive larger response scores. Using E directly as a ranking score gives an AUROC of 0.930: a value of 0.5 corresponds to random ranking and 1 to perfect separation. Ordinary response magnitude reaches 0.900, while endpoint magnitude reaches 0.960 (Figure 3(a)). The strong endpoint result is expected because this behavioral target is itself defined by a prediction change at the endpoint.

For endpoint-null sensitivity, we first restrict attention to pairs whose final background effect is negligible, $|d| < 0.02$. We then apply a separate set of background corruptions that are not used to construct the reveal path. A pair is marked positive if any of these held-out corruptions changes the



(a) Behavioral alignment. **(b)** Response-role accuracy. **(c)** Across magnitude bins.

Figure 3: Response semantics transfer to ImageNet-9. (a) Evidence and fragility align with independently measured natural-image behaviors. (b) Nearly identical response magnitudes can correspond to different response semantics, which DECAF preserves. (c) The advantage persists across magnitude bins.



(a) Response composition under patch reveal. **(b)** Model-rank transfer across reveal paths.

Figure 4: Reveal paths change response composition and ranking. (a) Patch reveal increases ordinary magnitude mainly through contradiction and fragility rather than evidence. (b) Ordinary-magnitude rankings transfer poorly across reveal paths, while evidence and contradiction remain substantially more stable.

predicted class or shifts the true-class probability by at least 0.20. We ask whether these independently identified sensitive pairs tend to receive larger F values than the remaining endpoint-null pairs. F reaches AUROC 0.878, compared with 0.433 for ordinary response magnitude, 0.316 for endpoint magnitude, and 0.635 for SmoothGrad (Figure 3(a)).

Together, the two tests show complementary behavior: E tracks consequential background use when the final effect is present, whereas F exposes sensitivity that remains when that final effect is negligible. Appendix F gives the full pair construction, held-out corruption set, behavioral-label definitions, and baseline details.

6.2 Same Response Magnitude, Different Response Semantics

We next ask whether the decomposition reveals information that ordinary response magnitude does not already contain. We reuse the independently defined evidence and fragility indicators from Section 6.1 and add a contradiction indicator: contradiction behavior is present when changing only the background makes the model switch from the foreground class to the class associated with the new background. These three behavioral indicators may overlap.

We then compare cases with different behavioral patterns but nearly identical ordinary response magnitudes. Requiring their Abs values to differ by at most 5% yields 8,289 matched comparisons. Because a case may exhibit more than one behavior, we do not force it into a single ground-truth class. Instead, we ask whether the largest of E, C, and F corresponds to a behavior that is actually present; overlap and ties are handled as detailed in Appendix F.

Under this test, a single magnitude provides no basis for preferring evidence, contradiction, or fragility. Abs reaches a role-agreement accuracy of 0.350, whereas DECAF reaches 0.964 (Figure 3(b)). Thus, nearly identical response magnitudes can accompany substantially different observed behaviors, while their decomposition preserves this distinction.

The conclusion also holds beyond the matched cases. Within narrow bins of ordinary response magnitude, macro-AUROC rises from 0.517 for Abs to 0.677 for DECAF, and DECAF exceeds Abs in 15 of the 16 bins that contain all three behavioral indicators (Figure 3(c)). Appendix F

Table 2: External attribution quality and measured ImageNet-1k compute. Quality is macro-averaged equally over the three architectures; time and memory are ImageNet-1k measurements. KernelSHAP is shown separately because it directly queries the same patch-deletion intervention used to define the ImageNet-1k evaluation target, giving it evaluation-specific information unavailable to the general baselines and DECAF.

Method	Access	FunnyBirds $\rho\uparrow$	ImageNet-1k $\rho\uparrow$	ms/img $\downarrow$	Peak GiB $\downarrow$
DECAF					
DECAF-3	Forward	0.372	0.359	22.1	11.2
DECAF-5	Forward	0.403	0.367	36.7	11.2
DECAF-9	Forward	**0.406**	**0.379**	65.9	11.2
General-purpose attribution					
DeepLIFT	Backward	0.197	0.341	3.1	6.6
IG-32	Backward	0.271	0.242	28.7	50.4
IG-U-32	Backward	0.200	0.295	28.7	50.4
RISE-512	Sampling	0.302	0.179	216.6	14.4
Endpoint-only reference					
Endpoint M	Endpoint	0.324	0.371	–	–
Benchmark-aligned reference (uses the evaluation intervention)					
KernelSHAP-512	Sampling	0.299	0.447	216.4	14.4

Table 3: DINOv2 ViT-g/14 on 238 PartImageNet strict-common-support images.

Method	Spearman $\uparrow$	sec/img $\downarrow$	Peak GiB $\downarrow$
DECAF-3	0.208	0.190	5.53
DECAF-5	0.215	0.300	5.66
DECAF-9	0.220	0.526	6.33
IG-16	**0.222**	0.710	9.01
IG-32	0.213	1.425	13.39
GradientSHAP	0.171	0.407	27.67
SmoothGrad-16	0.042	0.438	26.43
DeepLIFT	0.068	**0.084**	7.19

Table 4: Gain by the trajectory beyond endpoint-only M. Entries are paired $\Delta\rho = \rho(\mathrm{DECAF}) - \rho(M)$ with 95% confidence intervals. Larger values indicate better.

Contrast	FunnyBirds different intervention	ImageNet-1k same deletion
DECAF-3 $-M$	+.049 [.033, .064]	−.012 [−.014, −.010]
DECAF-5 $-M$	+.080 [.064, .096]	−.004 [−.007, −.002]
DECAF-9 $-M$	+.083 [.067, .099]	+.007 [.004, .010]

gives the exact matching rule, overlap and tie handling, per-bin support, and complete baseline comparisons.

6.3 What Changes When the Reveal Path Changes?

We finally ask whether the same endpoint information can produce different conclusions when it is revealed differently. We keep the model and both factual–counterfactual endpoints fixed. In one path, the whole image gradually changes from a common blurred image toward each endpoint; in the other, the same endpoint information is revealed region by region. We call them the blend and patch paths.

Ordinary response magnitude increases by about $1.8\times$ under the patch path, but the increase is not uniform across response roles (Figure 4(a)). Evidence remains close to its blend value, contradiction grows by about $1.8\times$, and fragility by more than $4\times$. This is consistent with their semantics: with the endpoints fixed, evidence remains oriented by the same final contrast, whereas fragility explicitly measures sensitivity along the chosen reveal path. The pattern is not guaranteed by the definition, but here the larger ordinary response comes primarily from fragility rather than additional evidence.

We also ask whether the reveal path changes conclusions about how models compare. For each model, we average E, C, F, and Abs over test images and rank the models separately by each quantity under the blend and patch paths. Spearman correlation measures the agreement between the two rankings. Ordinary-magnitude rankings are highly unstable ($\rho = 0.17/0.26$ for Same–Rand/Same–Next), whereas evidence rankings remain at $0.86/0.77$ and contradiction rankings near 0.93. Fragility is intermediate at $0.69/0.71$, consistent with its greater path dependence (Figure 4(b)). Patch-order and threshold checks are reported in Appendix G.

7 External Attribution Benchmarks and Large-Model Scaling

Sections 5–6 validate the response semantics of DECAF. We now ask whether it also works as ordinary feature attribution under targets defined without E, C, or F. Here the pair is constructed from the feature intervention: $\mathbf{x}^+$ is the original image and $\mathbf{x}^{-k}$ changes only part or patch k using a fixed removal or replacement rule; E_k ranks the features. FunnyBirds evaluates semantic parts under two separate replacement interventions, while ImageNet-1k IDSDS evaluates 16 fixed patches by single-patch deletion. Quality is the per-image Spearman correlation between the attribution ranking and this intervention-based ranking. Both datasets use *ResNet-50, VGG-16, and ViT-B/16*, with every method evaluated on the same eligible images within each model. DECAF-3/5/9 use three, five, or nine reveal stages.

All three DECAF trajectories exceed every listed general-purpose baseline on both datasets (Table 2). KernelSHAP reaches 0.447 on ImageNet-1k, but it directly queries the same deletion game used to define that target and is therefore shown separately. The full 50,000-image ImageNet check preserves

the ordering DECAF-5 $>$ IG-U-32 $>$ IG-32. Appendix H also reruns the official FunnyBirds RISE-6000 Single Deletion protocol and reproduces the published score within 0.005 on all three architectures (Hesse et al., 2023)

The compute advantage becomes clearer at larger model scale. On the *1B-parameter* model *DINOv2 ViT-g/14*, DECAF-5 and IG-32 have nearly identical quality (0.215 versus 0.213), while IG-32 requires $4.75\times$ the wall time and $2.36\times$ the peak allocated memory (Table 3).

Table 4 clarifies what the trajectory adds. On FunnyBirds, evaluation changes parts differently from the operation used to form $\mathbf{x}^{-k}$; the trajectory adds about 0.08 Spearman beyond M. ImageNet-1k instead evaluates the same patch deletion used to form $\mathbf{x}^{-k}$, so M is already aligned with the target; five stages nearly match it, while nine stages add only 0.007. Endpoint information therefore carries most of the attribution signal when evaluation repeats the same intervention. The trajectory adds its clearest value when the ranking must transfer to a different intervention. Complete intervals, per-architecture results, and boundary cases are in Appendix H.

8 Realized Mechanisms Across Models and Modalities

We test cross-modal transfer on 135 trained Covertype estimators from five model families under a strict common-start reveal. Both branches begin at the same legal row; shared clocks reveal the factor and endpoint context; and each realized response is routed before averaging. Thus $r(0) = 0$, $r(1) = d$, all binary inputs stay on support, and exact enumeration uses four queries. Appendix I gives the protocol and audits.

Table 5: Covertype under the strict common-start reveal. **(a)** Spearman correlation with realized behavior (95% intervals for DECAF). **(b)** Registered worker time per model. † Tree models only.

(a) Behavioral alignment

Held-out behavior	Var.	DECAF ρ	Abs	M	Native SHAP	SHAP inter.
Preservation	E	**0.747** [0.647,0.836]	0.694	0.657	0.781	$-0.251^{\dagger}$
Actual inversion	C	**0.989** [0.976,0.998]	-0.079	-0.001	-0.205	$0.093^{\dagger}$
Endpoint-null change	F	**0.973** [0.943,0.987]	0.499	0.148	0.481	$0.069^{\dagger}$

(b) Measured method cost

Method	Models	Worker-s/model	Rel.
DECAF	135	**1.42**	$1.0\times$
PDP/ALE inter.	135	2.22	$1.6\times$
KernelSHAP	45	45.30	$32.0\times$
Native SHAP	135	387.65	$273.8\times$
SHAP interaction	54	15097.63	$10662\times$

Across five model families, C tracks actual inversion with $\rho = 0.989$, F tracks endpoint-null change with $\rho = 0.973$, and E tracks preservation with $\rho = 0.747$. The nominally *Fragile* logistic model instead becomes endpoint-active ($M = 0.299$, $E = 0.150$, $F < 10^{-4}$), showing that DECAF follows the learned mechanism rather than the treatment label. Disjoint test halves retain correlations of 0.748, 0.989, and 0.974.

The label-free (E, C, F) readout reaches macro-AUROC 0.818, and 0.950 after conditioning on magnitude; M, Abs, and native SHAP remain at 0.464–0.467. The audit needs no gradients or internal activations. Native SHAP uses $274\times$ the worker time, while TreeSHAP interaction uses $10{,}662\times$ and reaches only $\rho = 0.093$ on inversion. The strict reveal therefore retains the mechanism signal at low black-box cost.

9 Discussion and Conclusion

Perturbation magnitude records how strongly a model reacts, but equal magnitudes can reflect different behavior. DECAF uses the final factual–counterfactual contrast to route intermediate responses into evidence, contradiction, and fragility while preserving $\mathrm{Abs} = E + C + F$.

Across controlled vision, natural images, external attribution benchmarks, and a strict tabular input reveal, these components track held-out behavior, recover mechanisms after magnitude is controlled, and diagnose whether a new path adds evidence or merely amplifies fragility. The method remains forward-only and model-agnostic.

These semantics are relative to the chosen pair, score, and path rather than hidden causal mechanisms. The endpoint describes its own intervention; the trajectory matters when behavior must transfer beyond it. Magnitude says how much; DECAF reveals what kind.

USE OF AI STATEMENT

Large language models were used as assistive tools during this work. They were used to improve the clarity and language of author-written text, to support literature retrieval and discovery of related work, to assist with research ideation and experimental execution (including discussing experimental designs, generating and debugging research code, and analyzing intermediate results), and to assist in developing and checking mathematical arguments and proofs.

All research questions, methodological choices, experimental protocols, interpretations, mathematical claims, and conclusions were determined and critically reviewed by the authors. All AI-assisted code, references, calculations, proofs, and factual claims were independently inspected and verified by the authors. The authors take full responsibility for the scientific content and the final manuscript.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Jock Blackard. Covertype [dataset]. UCI Machine Learning Repository, 1998.

David Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, pp. 265–272, 1953.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. 4 2018.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2019.

Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: a causal perspective. In *International Conference on Machine Learning*, pp. 981–990. PMLR, 2019.

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, 2021a.

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, 2021b.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, volume 31, 2018a.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, volume 31, pp. 590–601, 2018b.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Advances in neural information processing systems*, 32, 2019.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2950–2958. IEEE, 2019.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 3449–3457. IEEE, 2017.

Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In *International Conference on Learning Representations*, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3681–3688, 2019.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019a.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International conference on machine learning*, pp. 2376–2384. PMLR, 2019b.

Yikai Gu, Lele Cao, Bo Zhao, Lei Lei, and Lei You. Discover: A solver for distributional counterfactual explanations. In Michelangelo Ceci, Dragi Kocev, Anna Monreale, Celine Vens, Peter Flach, Giuseppe Manco, Myra Spiliopoulou, and Elio Masciari (eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track*, pp. 362–379, Cham, 2027. Springer Nature Switzerland.

Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3981–3991, 2023.

Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Benchmarking the attribution quality of vision models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916, 2020.

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

Neil Jethani, Adriel Saporta, and Rajesh Ranganath. Don't be fooled: Label leakage in explanation methods and the importance of their quantitative evaluation. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8925–8953, 2023.

Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5050–5058, 2021.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2668–2677, 2018.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280. Springer, 2019.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5338–5348, 2020.

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

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference*, 2018.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In John DeNero, Mark Finlayson, and Sravana Reddy (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 97–101, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-3020.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 18770–18795, 2022.

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

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

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *International Conference on Learning Representations Workshop*, 2014.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020.

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

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021.

Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9680–9689, 2020.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Lei You, Lele Cao, Mattias Nilsson, Bo Zhao, and Lei Lei. Distributional counterfactual explanations with optimal transport. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1135–1143. PMLR, 03–05 May 2025.

Lei You, Yijun Bian, and Lele Cao. Joint distribution–informed shapley values for sparse counterfactual explanations. In *International Conference on Learning Representations*, volume 2026, pp. 100183–100205, 2026.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833, 2014.

APPENDICES

A Detailed Related Work **14**

B Theoretical Foundations and Proofs **15**

C Estimators, Complexity, and Implementation **17**

D Controlled Experimental Setup **18**

E Selected Controlled Results **20**

F ImageNet-9 Experimental Details **23**

G ImageNet-9 Protocol Robustness Checks **25**

H Forward-Only Attribution: Complete Results and Boundaries **26**

I Covertype Experimental Details and Complete Results **29**

A DETAILED RELATED WORK

The closest literatures all study model response, but they attach different objects and semantics to that response. We organize the comparison around the object being explained and the question it answers.

A.1 GRADIENT, PATH, AND SIGNED ATTRIBUTION

Gradient explanations measure local sensitivity, while path methods accumulate sensitivity from a baseline to the input (Simonyan et al., 2014; Selvaraju et al., 2017; Sundararajan et al., 2017; Smilkov et al., 2017; Xu et al., 2020; Kapishnikov et al., 2021; You et al., 2026). Layer-wise relevance propagation, DeepLIFT, FullGrad, and SHAP instead decompose a prediction into additive feature contributions under conservation, reference, or game-theoretic principles (Bach et al., 2015; Shrikumar et al., 2017; Srinivas & Fleuret, 2019; Lundberg & Lee, 2017). Some of these methods preserve positive and negative contributions (Dhurandhar et al., 2018b).

DECAF studies a different object. It does not distribute one prediction across input coordinates. It decomposes the scalar response of an already specified paired intervention. The clean endpoint supplies the orientation, and the endpoint-null branch separates fragility from positive or negative evidence. Spatial saliency and DECAF are therefore complementary: saliency indicates where a response is localized, whereas DECAF indicates what semantic role the paired response plays.

A.2 PERTURBATION, COUNTERFACTUAL, AND CAUSAL EXPLANATION

Occlusion, Meaningful Perturbations, Extremal Perturbations, RISE, LIME, and removal-based Shapley methods construct interventions and summarize the resulting model changes (Zeiler & Fergus, 2014; Fong & Vedaldi, 2017; Fong et al., 2019; Petsiuk et al., 2018; Ribeiro et al., 2016; Covert et al., 2021b). Counterfactual and contrastive methods retrieve or synthesize changes that alter a decision or support a contrast (Goyal et al., 2019b; Chang et al., 2019; Dhurandhar et al., 2018b; You et al., 2025; Gu et al., 2027). Causal attribution and concept-intervention methods interpret relevance through explicit interventions on inputs, representations, or human-interpretable concepts (Chattopadhyay et al., 2019; Janzing et al., 2020; Goyal et al., 2019a; Kim et al., 2018; Koh et al., 2020).

DECAF takes the paired intervention as input rather than searching for it. Counterfactual construction defines the factor and identification assumptions; the paired trajectory measures the response; endpoint orientation then decomposes that response into evidence, contradiction, and fragility. The same decomposition can therefore sit on top of different removal operators or counterfactual generators.

A.3 BASELINES, OFF-MANIFOLD EFFECTS, AND FAITHFULNESS EVALUATION

Baseline and path choices can substantially change attribution (Sturmfels et al., 2020; Kindermans et al., 2019), and perturbations can move samples away from the data manifold (Frye et al., 2021).

Sanity checks, retraining-based evaluations, infidelity measures, and debiased removal metrics test whether explanations reflect the model rather than artifacts of the evaluation protocol (Adebayo et al., 2018; Hooker et al., 2019; Yeh et al., 2019; Rong et al., 2022). Other work documents label leakage, explanation fragility, and adversarial manipulation (Jethani et al., 2023; Ghorbani et al., 2019; Dombrowski et al., 2019; Slack et al., 2020).

DECAF does not claim path invariance. It keeps the intervention protocol explicit and asks which component changes when the protocol changes. This distinction matters in ImageNet-9: nested-patch reveal raises total response by roughly 80%, but the increase comes primarily from contradiction and fragility rather than evidence. DECAF addresses a semantic ambiguity that remains even when a paired response has been measured faithfully.

A.4 Efficiency, black-box access, and benchmark context

Black-box perturbation methods trade model access for repeated queries; RISE and dense occlusion can require many forward evaluations, while amortized methods such as FastSHAP add a separate explainer-training problem (Petsiuk et al., 2018; Jethani et al., 2022). DECAF instead reuses the signed scores already computed by a paired perturbation curve. The routing step adds no model queries, requires no gradients or internal activations, and can be batched across examples, stages, factors, and models.

ImageNet-9 is a standard setting for studying background reliance and distribution shift (Xiao et al., 2021; Geirhos et al., 2020). DECAF asks a finer question than whether a model responds to background: is that response endpoint-aligned evidence, endpoint-opposed contradiction, or sensitivity that appears only when the endpoint effect is null?

B Theoretical Foundations and Proofs

This appendix develops the compact theory used in the main text. We begin with the pointwise decomposition, then lift it to a trajectory, a population, and a decision problem. The proofs require only integrability of the signed response and do not assume differentiability, smoothness, or a particular model class.

B.1 Positive and negative parts

For a real number u, define $u^+ = \max\{u, 0\}$ and $u^- = \max\{-u, 0\}$. Then

$$u = u^+ - u^-, \qquad |u| = u^+ + u^-, \qquad u^+u^- = 0. \tag{2}$$

For an integrable function $z : \mathcal{T} \to \mathbb{R}$, the signed measure $\nu(A) = \int_A z(t)\,\mathrm{d}\mu(t)$ admits the Jordan decomposition $\nu = \nu^+ - \nu^-$, where $\mathrm{d}\nu^+ = z^+\mathrm{d}\mu$ and $\mathrm{d}\nu^- = z^-\mathrm{d}\mu$. DECAF applies this decomposition to the endpoint-oriented response on the active branch and reserves a separate branch for endpoint-null pairs.

B.2 Canonicality

Proof of Theorem 1. Fix a, s, and r. If $a = 0$, endpoint gating requires $e = c = 0$. Conservation then forces $f = |r|$. Hence the triple is unique.

Now suppose $a = 1$. Endpoint gating gives $f = 0$. Let $z = sr$. If $z \geq 0$, directional support requires $c = 0$ and conservation gives $e = |r| = z = z^+$. If $z \leq 0$, directional support requires $e = 0$ and conservation gives $c = |r| = -z = z^-$. Thus

$$(e, c, f) = \big(a(sr)^+, a(sr)^-, (1 - a)|r|\big). \tag{3}$$

The construction is nonnegative, conserves $|r|$, obeys the gate, and has the stated support. Therefore it is the unique triple satisfying the axioms. □

The theorem also yields a projection interpretation. On an active pair, e is the magnitude of the Euclidean projection of r onto the endpoint-aligned ray $\{s\lambda : \lambda \geq 0\}$, while c is the magnitude of the projection onto the opposite ray. The null branch is not a projection. It records that the endpoint does not provide a stable orientation at the chosen threshold.

B.3 MAGNITUDE NON-IDENTIFIABILITY AND STRICT REFINEMENT

Proof of Theorem 2. Fix $m > 0$. Consider three responses. First, let the endpoint be active with $s = 1$ and let $r = m$. Then $(e, c, f) = (m, 0, 0)$. Second, keep the same endpoint and let $r = -m$. Then $(e, c, f) = (0, m, 0)$. Third, let the endpoint be null and let $r = m$. Then $(e, c, f) = (0, 0, m)$. In every case, the observed magnitude is $|r| = m$. A statistic that observes only $|r|$ therefore cannot distinguish the three mechanisms. □

Corollary 1 (Strict response refinement)**.** *The DECAF profile is at least as informative as ordinary magnitude for every decision problem, because* $\mathrm{Abs} = E + C + F$ *is a deterministic function of the profile. It is strictly more informative for any mechanism-identification problem that assigns positive probability to two distinct profiles with the same magnitude.*

Proof. Any rule that uses Abs can be composed with the map $(E, C, F) \mapsto E + C + F$. Theorem 2 gives distinct profiles that share one magnitude. A decision problem that rewards correct mechanism identification separates those profiles, so no magnitude-only rule can match the best rule that observes the full profile. This is the standard logic of strict Blackwell refinement (Blackwell, 1953). □

A useful finite case makes the gap concrete. Suppose evidence, contradiction, and fragility are equally likely and all produce magnitude m. Every magnitude-only classifier receives the same observation, so its best accuracy is $1/3$. Under unequal priors, the best magnitude-only accuracy is the largest prior. The DECAF profile identifies the mechanism exactly in this idealized construction.

B.4 PRESERVATION, ATTENUATION, AND INVERSION

Proof of Proposition 1. Let the endpoint orientation be s. Under preservation, the stage response is sm, so $E = m$, $C = 0$, and $\mathrm{Abs} = m$.

Under attenuation, the response is sm with probability $1 - \eta$ and zero with probability η. Taking expectations gives $E = (1 - \eta)m$, $C = 0$, and $\mathrm{Abs} = (1 - \eta)m$.

Under inversion, the response is sm with probability $1 - \eta$ and $-sm$ with probability η. The first branch contributes m to evidence, while the second contributes m to contradiction. Therefore $E = (1 - \eta)m$, $C = \eta m$, and $\mathrm{Abs} = m$. The ratio is $C/(E + C) = \eta$. □

The equal-magnitude assumption isolates the inversion probability. If the aligned and opposed branches have magnitudes m_+ and m_-, then

$$E = (1 - \eta)m_+, \qquad C = \eta m_-, \qquad \frac{C}{E + C} = \frac{\eta m_-}{(1 - \eta)m_+ + \eta m_-}. \tag{4}$$

Thus the opposed fraction measures probability mass only when the two branches have comparable effect size. In general, C measures opposed causal mass.

B.5 POPULATION FACTORIZATION

Let $\pi = \Pr(|d| \geq \varepsilon)$. Define branch-conditional summaries whenever the corresponding branch has positive probability. Since evidence and contradiction vanish on the null branch, while fragility vanishes on the active branch,

$$E = \pi E^{\mathrm{active}}, \qquad C = \pi C^{\mathrm{active}}, \qquad F = (1 - \pi)F^{\mathrm{null}}. \tag{5}$$

Unconditional quantities combine branch prevalence and conditional intensity. Conditional quantities answer a different question and become unstable when their branch is rare. The factorization is therefore useful when an external outcome is defined only within the active or null branch, but the primary experiments report unconditional population summaries.

B.6 INVARIANCES

Endpoint swap. Swapping $\mathbf{x}^+$ and $\mathbf{x}^-$ multiplies both d and $r(t)$ by -1. The oriented response $sr(t)$ is unchanged. Hence M, E, C, F, and Abs are invariant.

Score translation. Replacing q by $q + b$ leaves every difference unchanged.

Positive affine score reparameterization. Let $q' = \lambda q + b$ with $\lambda > 0$. Then $d' = \lambda d$, $r'(t) = \lambda r(t)$, and $s' = s$. If the endpoint threshold is transformed consistently as $\varepsilon' = \lambda\varepsilon$, the gate is unchanged and

$$(M', E', C', F', \mathrm{Abs}') = \lambda(M, E, C, F, \mathrm{Abs}).$$

Hence all normalized component proportions are invariant.

With a numerically fixed threshold ε, however, the gate obeys

$$a'_\varepsilon = \mathbf{1}\{|d| \geq \varepsilon/\lambda\}.$$

Consequently,

$$E'_\varepsilon = \lambda E_{\varepsilon/\lambda}, \qquad C'_\varepsilon = \lambda C_{\varepsilon/\lambda}, \qquad F'_\varepsilon = \lambda F_{\varepsilon/\lambda}.$$

Thus positive scaling preserves component proportions only when gate membership is preserved. Otherwise, the discrepancy is controlled by the threshold-mass bound in Appendix B.8.

Negative score scaling. A negative scaling reverses the semantic meaning of the target score. The endpoint orientation still makes the component magnitudes invariant after multiplying by $|\lambda|$, but the analyst has changed the behavior being explained. We therefore define the score direction before analysis.

B.7 Protocol reparameterization

Let $h : \widetilde{\mathcal{T}} \to \mathcal{T}$ be an increasing bijection and define the reparameterized path $\widetilde{r}(u) = r(h(u))$. Pointwise components are reindexed:

$$\widetilde{e}(u) = e(h(u)), \quad \widetilde{c}(u) = c(h(u)), \quad \widetilde{f}(u) = f(h(u)). \tag{6}$$

If the integration measure is pushed forward consistently, $\widetilde{\mu} = h^{-1}_{\#}\mu$, then the integrated profile is invariant. If both paths are instead integrated with a uniform coordinate measure, their AUCs may differ. Dynamic DECAF summaries are therefore protocol-relative, and every experiment reports the path and stage measure.

B.8 Threshold stability

Let $0 \leq \varepsilon_1 < \varepsilon_2$ and suppose the per-example path summary is bounded by B. The two gates differ only on pairs satisfying $\varepsilon_1 \leq |d| < \varepsilon_2$. Consequently, for each component $G \in \{E, C, F\}$,

$$|G_{\varepsilon_2} - G_{\varepsilon_1}| \leq B \, \Pr(\varepsilon_1 \leq |d| < \varepsilon_2)\,. \tag{7}$$

The threshold is stable when little endpoint mass lies near the boundary. We report sensitivity over multiple thresholds rather than treating one numerical cutoff as universal.

C Estimators, Complexity, and Implementation

DECAF reuses the signed scores of a paired perturbation curve. Once those scores are available, the decomposition is elementwise.

C.1 Finite-grid estimators

Let $t_1, \ldots, t_T$ be stage points with nonnegative quadrature weights $w_1, \ldots, w_T$ that sum to one. For pair i, define

$$E_i = \sum_{j=1}^{T} w_j e_i(t_j), \quad C_i = \sum_{j=1}^{T} w_j c_i(t_j), \quad F_i = \sum_{j=1}^{T} w_j f_i(t_j). \tag{8}$$

For N pairs, the empirical estimates are

$$\widehat{E} = \frac{1}{N}\sum_{i=1}^{N} E_i, \quad \widehat{C} = \frac{1}{N}\sum_{i=1}^{N} C_i, \quad \widehat{F} = \frac{1}{N}\sum_{i=1}^{N} F_i, \quad \widehat{M} = \frac{1}{N}\sum_{i=1}^{N} |d_i|. \tag{9}$$

Every implementation checks the exact sample-level identity $|r_i(t_j)| = e_i(t_j) + c_i(t_j) + f_i(t_j)$. We use trapezoidal weights on ordered grids unless stated otherwise. When the protocol contains randomness, factual and counterfactual branches share that randomness; repeated paths are averaged or retained as clusters for uncertainty estimation.

C.2 Query complexity

Suppose we evaluate K factors, T stages, J counterfactual maps per factor, and R protocol repetitions. If factual branch evaluations are shared across factors, the number of model evaluations per example is

$$RT(1 + KJ). \tag{10}$$

For one factor and one map this is $2RT$. DECAF adds no model evaluations to the ordinary signed paired trajectory; its arithmetic cost is linear in the number of scalar responses.

The ImageNet-9 main path uses $T = 9$, so it requires 18 forward evaluations per pair. Under our configurations, Integrated Gradients and SmoothGrad each use 16 forward–backward evaluations, BlurIG uses 12, Occlusion uses 49 forward evaluations, and RISE uses 256. These counts describe access and query structure rather than hardware-independent wall time.

C.3 Memory, batching, and black-box models

A streaming implementation stores the endpoint sign, current stage scores, and three accumulators. Its additional memory is $O(BK)$ scalars for batch size B and K factors, beyond the model and input batch. Examples, branches, stages, factors, counterfactual maps, repetitions, and models provide independent batching dimensions.

DECAF requires no differentiability. It applies to neural networks, tree ensembles, simulators, and remote APIs whenever a stable real-valued score is available. For stochastic models, branches should share random seeds or use repeated queries. Hard labels are formally sufficient, but probabilities, logits, margins, regression values, or action values produce more informative endpoint effects.

D Controlled Experimental Setup

The controlled suite gives DECAF an identifiable target before we move to natural images. Every experiment uses exact factor interventions in 3D Shapes, shared protocol randomness across branches, and two architectures. The experiments differ in the behavior they are designed to activate.

D.1 Dataset and factors

3D Shapes contains $480{,}000$ rendered scenes generated from six factors: floor color, wall color, object color, object size, object shape, and object orientation (Burgess et al., 2018). The Cartesian product is fully enumerated. This structure gives us exact counterfactuals: to intervene on one factor, we change its index and hold the other five indices fixed.

We use binary prediction tasks derived from the factors. The base benchmark contains five tasks:

- **Object color:** a binary split of the object-color values.
- **Wall color:** a binary split of the wall-color values.
- **Object shape:** the two shapes that are common to all model-response supports.
- **Color–shape XOR:** the exclusive-or of binary object color and shape.
- **Context gate:** a color decision whose active factor depends on wall context.

For each task we train ResNet-18 and a small ViT with three random seeds. The resulting 30 models are crossed with all six factors, producing 180 model–factor units.

D.2 CLEAN ENDPOINT PAIRS

For a factual scene $\mathbf{x}^+$ and target factor k, the counterfactual $\mathbf{x}^-$ changes only factor k. Binary factors are flipped. Multivalued color, size, and orientation factors use fixed involutive maps so that applying the map twice returns the original value. We use two independent maps for multivalued factors and average their measurements.

The clean endpoint effect is computed on 8,192 held-out factual–counterfactual pairs. The main threshold classifies a pair as active when the absolute target-score difference exceeds the experiment-specific ε. The endpoint classification is a property of the model and pair, not of the task label.

D.3 PRIMARY REVEAL PROTOCOL

The main controlled protocol is covariance-matched Gaussian reveal. Let $\boldsymbol{\mu}$ and $\widehat{\boldsymbol{\Sigma}}$ be the empirical mean and covariance of the image distribution. For $\alpha \in [0, 1]$,

$$\mathbf{x}_\alpha = \boldsymbol{\mu} + \sqrt{\alpha}(\mathbf{x} - \boldsymbol{\mu}) + \sqrt{1-\alpha}\,\boldsymbol{\eta}, \qquad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}). \tag{11}$$

The factual and counterfactual branch share the same $\boldsymbol{\eta}$. At $\alpha = 1$ the path reaches the clean input; at $\alpha = 0$ the state is independent of the individual sample under the fitted Gaussian reference. The covariance match makes the stimulus second-order neutral in whitened data coordinates. DECAF does not depend on this particular path; it only consumes the paired scores that the path produces.

The base benchmark uses 4,096 dynamic factual pairs, three noise seeds, two counterfactual maps, and 21 points along a trace-matched covariance family. The primary CMMR endpoint is accompanied by a trace-matched pixel-Gaussian endpoint. A diagonal covariance and a trace-normalized covariance-power path provide held-out geometries.

D.4 ALTERNATIVE PROTOCOLS

The covariance interpolation is

$$\boldsymbol{\Gamma}_\lambda = (1-\lambda)\widehat{\boldsymbol{\Sigma}} + \lambda\tau\mathbf{I}, \qquad \tau = \frac{\operatorname{tr}(\widehat{\boldsymbol{\Sigma}})}{D},$$

where D is the pixel dimension. The two endpoints use data covariance and trace-matched isotropic pixel covariance. The diagonal held-out path uses $\operatorname{diag}(\widehat{\boldsymbol{\Sigma}})$. The power path raises covariance eigenvalues to a power and renormalizes the trace. These paths test whether a semantic conclusion depends on one second-order geometry.

We use the protocol family as an audit, not as a learned worst-case score. The main paper reports CMMR-based DECAF and uses the alternatives to characterize transfer.

D.5 EVIDENCE VALIDATION THROUGH TRAINING TRAJECTORIES

The evidence experiment trains shape classifiers in an environment where wall color is strongly correlated with the label. We use two training correlations, 0.95 and 0.99, two architectures, two seeds, and save checkpoints throughout training. This creates eight trajectories and 52 selected checkpoints.

The external target is shortcut-reversal vulnerability,

$$V_{\mathrm{rev}} = \mathrm{Acc}(p_{\mathrm{test}} = 0.95) - \mathrm{Acc}(p_{\mathrm{test}} = 0.05).$$

A shortcut-dominant checkpoint performs well when wall color remains aligned and fails when the correlation reverses. A shape-dominant checkpoint is stable. We compare V_{rev} with the evidence margin $E_{\mathrm{wall}} - E_{\mathrm{shape}}$.

D.6 FRAGILITY VALIDATION

The fragility experiment uses a binary object-shape task and treats floor color as the candidate endpoint-null factor. We train robust, neutral, and fragile variants for each architecture and three seeds, producing 18 models.

The neutral variant uses clean training only. The robust variant is encouraged to keep its output stable under floor-color counterfactuals at intermediate states. The fragile variant is encouraged to respond to floor color away from the endpoint while preserving the clean shape task. The intended comparison is restricted to models for which clean accuracy remains high and floor color remains endpoint-null.

The behavioral target is the prediction-change rate under held-out floor-color interventions at intermediate states. We also measure randomized-floor stability and transfer across covariance geometries.

D.7 Contradiction validation

The contradiction benchmark uses two binary variables: object color A and wall context G. The factual endpoint is evaluated in the context $G = 1$, where all three tasks share the clean rule $Y = A$. Under the swapped context $G = 0$:

$$\text{Direct: } Y = A, \qquad \text{Gate: } Y = H, \qquad \text{Invert: } Y = 1 - A,$$

where H is object shape. Direct preserves the object-color effect. Gate removes it. Invert reverses it.

We train 30 models: three tasks, two architectures, and five seeds. Balanced validation accuracy is at least 0.9987. A binary symmetric context channel swaps the wall context with probability $\eta \in [0, 0.5]$. All intermediate images remain in the support of the generated dataset. Two independent wall-color involutions test map transfer.

The label-level outcomes are preserve, collapse, and swap. The key contrast is between Gate and Invert: both reduce aligned evidence, but only Invert should create endpoint-opposed mass.

D.8 Statistical summaries

The base benchmark uses stratified bootstrap over task, architecture, and seed. The evidence experiment bootstraps complete trajectories. The fragility and contradiction experiments report seed-level variation and architecture-stratified summaries. Bootstrap intervals use 500 repetitions unless stated otherwise. We treat the model or training trajectory as the main statistical unit rather than counting every image as independent evidence.

E Selected Controlled Results

This appendix expands the behavioral validation in Section 5 without repeating its main tests. We report five positive extensions: a matched-magnitude example, the complete response atlas, all evidence-training trajectories, fragility checks across interventions and reveal geometries, and contradiction checks across regimes, architectures, seeds, and counterfactual maps.

E.1 Matched-magnitude response roles

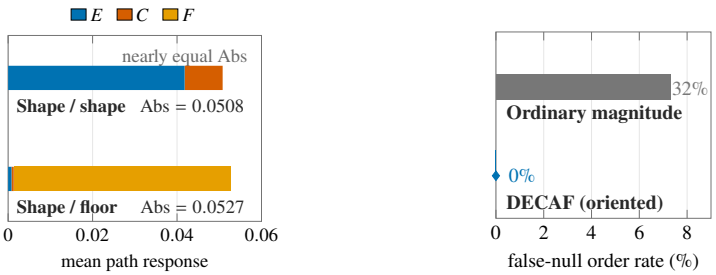
Figure 5 gives a concrete example of the scalar ambiguity discussed in the main text. In an object-shape model, the true shape factor and an endpoint-null floor-color factor produce nearly equal ordinary response magnitude, but their response roles differ sharply. The same distinction matters across the controlled benchmark: ordinary magnitude can rank an endpoint-null factor above a supported factor, whereas endpoint-oriented routing removes these false-null reversals.

E.2 Complete decomposition atlas

Figure 6 displays every task–factor combination. Endpoint-null factors concentrate near the fragility corner, direct color tasks concentrate in evidence, and interaction tasks carry more contradiction and mixed mass. Across 125 endpoint-null units, 96.8% of ordinary response is fragility on average. Across 48 endpoint-supported units, the mean composition is 64.1% evidence, 12.5% contradiction, and 23.4% null fragility.

E.3 Evidence across training trajectories

Figure 7 places exact shortcut vulnerability and the DECAF evidence margin on the same eight checkpoint trajectories. Six trajectories exhibit nonconstant vulnerability, and all six show positive



(a) Nearly equal magnitude, different response roles.

(b) Orientation removes false-null reversals.

Figure 5: Equal response magnitude can hide different behavior. (a) The true shape factor and an endpoint-null floor-color factor have nearly identical ordinary magnitude, yet DECAF separates evidence from fragility. (b) Across the benchmark, ordinary magnitude ranks an endpoint-null factor above a supported factor in 7.32% of comparable pairs, whereas endpoint-oriented routing reduces this rate to zero.

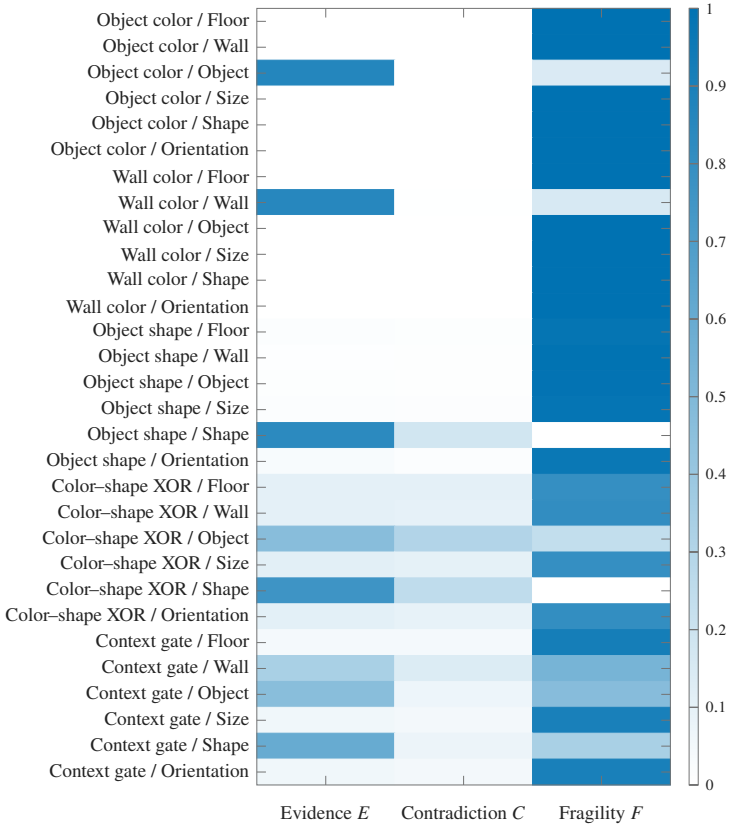
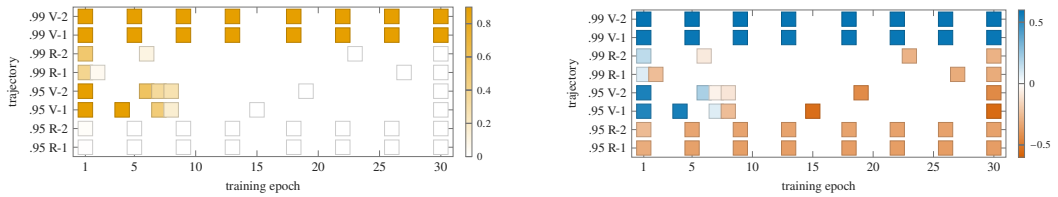


Figure 6: Complete controlled decomposition atlas. Results cover five tasks, two architectures, three seeds, and six factors. Endpoint-null factors are dominated by fragility, while task-relevant and interaction factors contain evidence and contradiction in different proportions.

within-trajectory association between evidence margin and shortcut vulnerability. The two Small-ViT trajectories trained at the strongest shortcut correlation remain shortcut-dominant, and DECAF correspondingly keeps their evidence margin high.

E.4 Fragility: intervention separation and reveal geometry

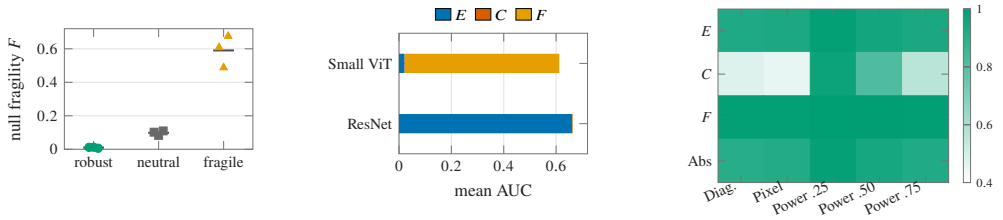
The main text validates F against independently measured off-path prediction change. Figure 8 provides three complementary checks. First, F cleanly orders robust, neutral, and fragile interventions. Second, similar total response can correspond to different endpoint-relative roles across architectures. Third, the robust–neutral–fragile ordering transfers from the primary covariance-matched reveal to



(a) Exact shortcut vulnerability. **(b)** Evidence margin.

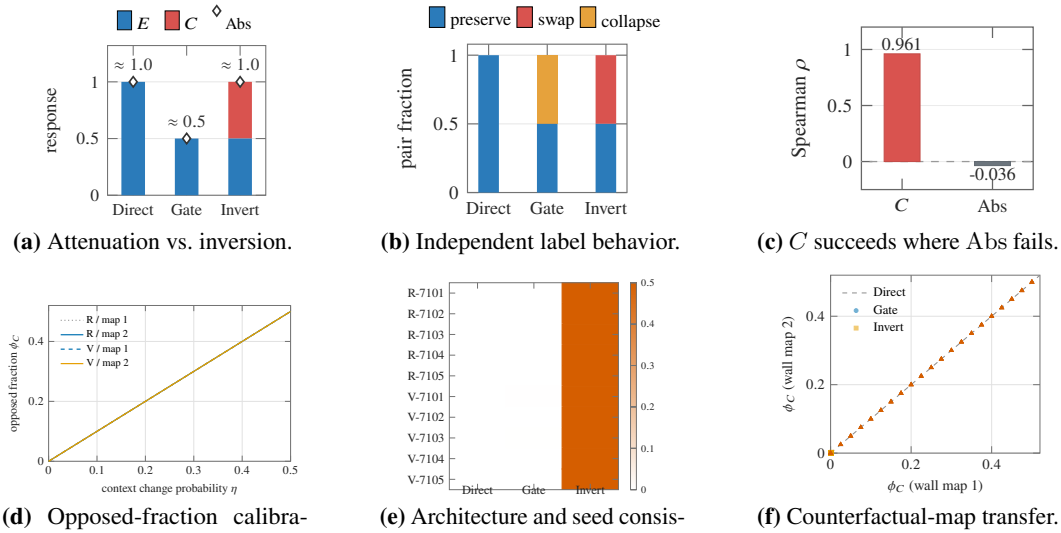
Figure 7: Evidence follows shortcut reliance across complete training trajectories. The two panels show the same eight checkpoint sequences through an external behavioral measure and the DECAF evidence margin.

diagonal covariance, trace-matched pixel Gaussian, and three covariance-power geometries. The controlled fragility result is therefore not tied to one second-order reveal geometry, although the numerical summaries remain protocol-relative.



(a) Intervention separation. **(b)** Similar magnitude, different roles. **(c)** Cross-geometry rank transfer.

Figure 8: Additional fragility checks. (a) F separates robust, neutral, and fragile training. (b) Similar ordinary response can be endpoint evidence in one model and endpoint-null fragility in another. (c) The controlled ordering transfers across held-out reveal geometries.



(a) Attenuation vs. inversion. **(b)** Independent label behavior. **(c)** C succeeds where Abs fails.

(d) Opposed-fraction calibration. **(e)** Architecture and seed consistency. **(f)** Counterfactual-map transfer.

Figure 9: Additional contradiction checks. Top: contradiction separates attenuation from inversion, agrees with independently measured label behavior, and succeeds where ordinary magnitude does not. Bottom: the opposed fraction is calibrated and stable across architectures, seeds, and alternative wall-color counterfactual maps.

E.5 Contradiction: Regime Separation and Calibration

The main text shows that C tracks independently measured label reversal. Figure 9 expands this result. Direct preserves the effect, Gate suppresses it, and Invert reverses it; only inversion creates substantial opposed mass. The induced label behavior independently separates into preservation, collapse, and swap. Across architectures, seeds, and wall-color maps, the opposed fraction remains well calibrated. Raw sign changes can be frequent when the stage effect is numerically close to zero,

but the corresponding opposed mass remains small, separating magnitude-aware contradiction from sign counting.

F IMAGENET-9 EXPERIMENTAL DETAILS

This appendix gives the construction and secondary diagnostics behind Section 6. The main text keeps only the tests needed for the central argument. Here we specify the data splits, model zoo, paired background interventions, independent behavioral indicators, magnitude-controlled evaluation, spatial-attribution baselines, and cross-architecture comparison.

F.1 DATA VARIANTS AND DISJOINT SPLITS

We use the official ImageNet-9 Backgrounds Challenge variants (Xiao et al., 2021): Original, Mixed-Same, Mixed-Rand, Mixed-Next, Only-FG, and the challenge backgrounds. The variants preserve foreground identity while changing how the background is constructed. We match variants by foreground identifier and retain 4,050 foregrounds for which the required variants are available.

A deterministic split keeps the roles of the data separate. We reserve 1,644 paired foregrounds for the broad model-level response scan and an 820-foreground deep pool for the sample-level benchmark; 768 foregrounds from this pool are used for the expensive baseline comparison. The remaining foregrounds are not used in that sample-level benchmark. No foreground appears in more than one split.

F.2 MODEL ZOO AND COMMON SCORE SPACE

The 72-model zoo contains two complementary groups. The first consists of 24 off-the-shelf ImageNet-1k classifiers from torchvision and timm. These models were trained independently of our audit and span ResNet, ResNeXt, WideResNet, DenseNet, EfficientNet, RegNet, ConvNeXt, DeiT, ViT, BEiT, Swin, MaxViT, and CoAtNet architectures.

The second group contains 48 models trained directly on ImageNet-9. We fine-tune six backbones—ResNet-50, ConvNeXt-Tiny, EfficientNet-B3, RegNetY-8GF, Swin-T, and ViT-B/16—on four versions of the training data: Original, Mixed-Rand, Mixed-Next, and Only-FG, using two random seeds. This $6 \times 4 \times 2$ design broadens the range of learned background dependence while retaining the same ImageNet-9 prediction task.

We train the fine-tuned models for eight epochs with AdamW, cosine decay, one warm-up epoch, BF16 computation, random resized crops, and horizontal flips. ResNet-50, EfficientNet-B3, and RegNetY-8GF use batch size 256 and learning rate 3×10^{-4}. ConvNeXt-Tiny, Swin-T, and ViT-B/16 use batch size 128 and learning rate 10^{-4}.

All models are evaluated in the same nine-class score space. For an ImageNet-1k model, we compute its 1,000-way softmax and sum the probabilities belonging to each official ImageNet-9 superclass. We do not sum logits or apply a second softmax. The ImageNet-9 fine-tuned models output nine logits directly. In both groups, the scalar analyzed by DECAF is the probability assigned to the true ImageNet-9 superclass.

F.3 PAIRED BACKGROUND INTERVENTIONS AND REVEAL PATHS

For each foreground, we construct two paired background interventions. Same–Rand compares Mixed-Same with Mixed-Rand: the foreground is unchanged, while a same-class background is replaced by a random-class background. This pair is used for the evidence and fragility analyses. Same–Next compares Mixed-Same with Mixed-Next, where the replacement background comes from superclass $(y + 1) \bmod 9$. This pair creates a directional background change for the contradiction analysis.

The primary reveal begins from a shared blurred midpoint. For endpoints $\mathbf{x}^+$ and $\mathbf{x}^-$,

$$\mathbf{x}^0 = \mathrm{Blur}\left(\frac{\mathbf{x}^+ + \mathbf{x}^-}{2}\right), \tag{12}$$

and each branch is linearly revealed from $\mathbf{x}^0$ to its endpoint. We evaluate nine stages, $t \in \{0, 0.125, \ldots, 1\}$. The main endpoint threshold is $\varepsilon = 0.02$ in true-class probability.

To test whether conclusions depend on the reveal path, we also use nested-patch reveal on an 8×8 grid. Patches are ordered by endpoint-difference energy, with a fixed random tie-break. Two independently tie-broken orders are evaluated. Both branches use the same patch order and the same neutral starting image. Appendix G reports the corresponding path-order and threshold checks.

F.4 Behavioral indicators defined without DECAF

The sample-level benchmark defines three observable behavioral indicators directly from model predictions. None uses E, C, or F, and the indicators are allowed to overlap.

Evidence behavior. Set $Y_E = 1$ when Mixed-Same is classified correctly and replacing its background with Mixed-Rand either changes the predicted class or lowers the true-class probability by at least 0.20. This records a consequential change caused by replacing the background.

Contradiction behavior. Set $Y_C = 1$ when changing only the background makes the model switch from the foreground class to the class associated with the new Mixed-Next background. This is an independently observed directional reversal in the model's prediction.

Fragility behavior. We first restrict attention to Same–Rand pairs whose fully revealed endpoints produce little difference, $|d| < 0.02$. The endpoint pair alone cannot tell us whether such a case is otherwise sensitive, so we test it separately with additional background changes that are not part of the DECAF reveal. These changes use Gaussian blur, Gaussian noise, pixelation, color shift, and patch shuffle, each at two severities. Set $Y_F = 1$ if any of them changes the predicted class or shifts the true-class probability by at least 0.20.

These indicators are external behavioral checks, not definitions of the DECAF components. Their purpose is to ask whether the response roles obtained from the paired trajectory agree with behavior measured independently of that trajectory decomposition.

F.5 Magnitude-controlled evaluation

The deep benchmark uses 32 models and 768 foregrounds, giving 24,576 model–image units. We use two complementary ways to control ordinary response magnitude.

First, we divide the units into 20 quantile bins of Abs. Within a bin, a behavioral AUROC is computed only when both positive and negative examples are present for that indicator. The main text reports both the valid-bin summary and the stricter common-support comparison in which all three indicators are represented.

Second, we construct one-to-one matched comparisons between cases with different behavioral indicator patterns but nearly identical ordinary response magnitudes. We require the relative difference in Abs to be at most 5%, yielding 8,289 matched comparisons.

Because Y_E, Y_C, and Y_F may overlap, the matched evaluation does not force every case into one ground-truth class. For a case with at least one active indicator, we examine the largest score among the three response-role coordinates. The case receives full credit when the largest coordinate corresponds to an active behavioral indicator. If several coordinates tie for the maximum, credit is divided across the tied coordinates. The reported matched-pair accuracy averages this role-agreement score over the matched cases. Ordinary magnitude has only one scalar value and therefore cannot prefer evidence, contradiction, or fragility.

The implementation also records a separate pairwise-ranking diagnostic: when an independent behavioral indicator differs across a matched pair, it asks whether the corresponding response-role score changes in the same direction. This ranking diagnostic is distinct from the role-agreement accuracy reported in Section 6.2.

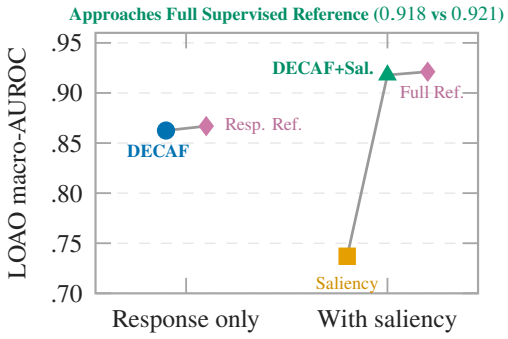


Figure 10: Cross-architecture transfer and spatial complementarity. All learned results hold out one architecture family during fitting. Response features from DECAF nearly match the supervised response-statistics reference, while adding spatial attribution provides a further gain and approaches the full supervised reference.

F.6 Spatial-attribution baselines

The deep benchmark includes six standard spatial-attribution methods: Input×Gradient, Integrated Gradients with 16 steps, SmoothGrad with 16 noise samples, BlurIG with 12 blur levels, a 7×7 Occlusion grid, and RISE with 256 masks. For each method, we sum absolute attribution inside the foreground–background difference region. These scores describe where the prediction is sensitive; they are not used to define the three behavioral indicators above.

F.7 Cross-architecture comparison and spatial complementarity

For the learned comparison, every representation is evaluated with the same supervised protocol: train on all but one architecture family and evaluate on the held-out family. The DECAF representation uses (M, E, C, F). A response-statistics reference uses $(M, \mathrm{Abs}, \mathrm{Net}, \mathrm{SignFlip}, \mathrm{active\ rate})$. The spatial representation uses the six attribution scores above. We also evaluate DECAF combined with the spatial scores and a full reference combining response statistics with spatial attribution.

Figure 10 contains the panel moved from the main text. Using response information alone, DECAF reaches leave-one-architecture-family-out macro-AUROC 0.862, close to the supervised response-statistics reference at 0.867. Spatial attribution alone reaches 0.737. Combining DECAF with spatial attribution reaches 0.918, close to the full supervised reference at 0.921. This is best read as a complementarity check: the response decomposition and spatial attribution retain different information.

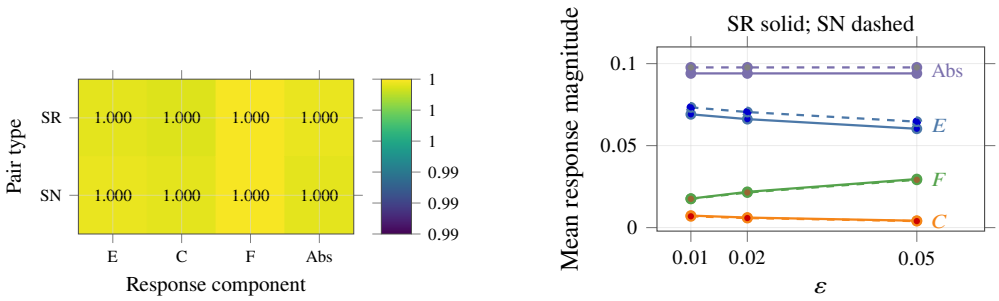
F.8 Statistical reporting

For direct behavioral comparisons, we report AUROC and AUPRC. For learned comparisons, we report leave-one-architecture-family-out results so that the evaluated architecture family is not used to fit the probe. Undefined AUROCs remain missing rather than being replaced by chance. Where bootstrap intervals are reported, models—rather than individual model–image rows—are the resampling unit.

G ImageNet-9 Protocol Robustness Checks

The protocol audit in Section 6 changes the reveal path while keeping the factual–counterfactual endpoints fixed. We retain two checks that test whether the reported effect is caused by an arbitrary implementation choice.

First, the nested-patch reveal is evaluated with two independently tie-broken patch orders. Their response summaries agree almost perfectly, showing that the blend-to-patch contrast is not an accident of one patch sequence. Second, we recompute the decomposition at $\varepsilon \in \{0.01, 0.02, 0.05\}$. Changing



(a) Patch-order stability. **(b)** Endpoint-threshold sensitivity.

Figure 11: Robustness of the ImageNet-9 protocol audit. (a) The two independently tie-broken nested-patch orders agree almost perfectly for every response component. (b) Varying the endpoint threshold changes the allocation between active and endpoint-null response, but not the qualitative protocol conclusion.

the threshold reallocates response between the active and endpoint-null branches as expected, but preserves the qualitative conclusion that patch reveal increases total response without increasing evidence.

H Forward-Only Attribution: Complete Results and Boundaries

This appendix expands the compact comparison in Section 7. It records the strict common-support protocol, complete baseline results, an official-protocol FunnyBirds reproduction check, the endpoint-versus-trajectory ablation, full-scale ImageNet validation, measured compute, large-model scaling, and the PartImageNet boundary case.

H.1 Protocol and strict common support

FunnyBirds and ImageNet-1k use the same three architectures: ResNet-50, VGG-16, and ViT-B/16. FunnyBirds evaluates semantic-part rankings against two held-out operators, Telea inpainting and background-texture replacement. ImageNet-1k IDSDS partitions every image into a fixed 4×4 grid and evaluates the ranking of 16 patch scores against in-domain single-deletion effects. We first filter to images classified correctly by each model. We then intersect image IDs across the methods included in the strict comparison. The resulting support contains 499, 497, and 488 FunnyBirds images, and 7,663, 7,189, and 8,285 ImageNet images, respectively.

For each image, the primary metric is Spearman correlation between feature attribution and the benchmark target. We average within each model and then macro-average equally across architectures. All intervals use 1,000 paired image-cluster bootstrap replicates. We never average the FunnyBirds and ImageNet columns into one leaderboard.

Endpoint $M_k = |d_k|$ is an endpoint-only reference. On ImageNet IDSDS, the endpoint audit gives $\max_k |d_k - g_k| = 0$ and confirms that every stored endpoint effect uses the same deletion pair as the evaluation target. The direct target-derived $|g_k|$ and independently persisted DECAF M_k differ only by a mean numerical discrepancy of 6.0×10^{-6}.

H.2 Complete cross-dataset method comparison

Published-benchmark sanity check. On the original full-scale IDSDS ResNet-50 setting, our IG and IG-U scores are 0.194 and 0.252, closely matching the published values of 0.196 and 0.255 (Hesse et al., 2024). This agreement provides an external check on our ImageNet implementation. Our FunnyBirds evaluation instead uses held-out intervention operators and strict common support, so it is not intended to reproduce the dataset's native part-based evaluation protocol (Hesse et al., 2023). We therefore compare its baseline values only within our registered protocol.

The stronger IG baseline changes across datasets. IG-32 exceeds IG-U-32 on FunnyBirds, while IG-U-32 is stronger on IDSDS. DECAF-5 exceeds both variants on both datasets. KernelSHAP exhibits the opposite transfer pattern: it is strongest on IDSDS, whose target is the deletion game it queries, but is weaker than endpoint M and DECAF under the held-out FunnyBirds operators.

Table 6: Complete strict-common-support attribution results. Entries are equal-architecture macro-average Spearman correlations with 95% paired image-cluster bootstrap intervals. Endpoint M is an endpoint-only reference. KernelSHAP-512 is a deletion-game reference.

Method	Access	FunnyBirds ρ [95% CI]	ImageNet-1k IDSDS ρ [95% CI]
DECAF			
DECAF-3	Forward only	0.372 [0.353, 0.392]	0.359 [0.355, 0.364]
DECAF-5	Forward only	0.403 [0.385, 0.422]	0.367 [0.362, 0.371]
DECAF-9	Forward only	0.406 [0.388, 0.425]	0.379 [0.375, 0.383]
General-purpose attribution			
Input×Gradient	Backward	0.019 [−0.002, 0.042]	0.098 [0.094, 0.101]
IG-16	Backward	0.266 [0.246, 0.285]	0.238 [0.235, 0.242]
IG-32	Backward	0.271 [0.251, 0.290]	0.242 [0.238, 0.245]
IG-U-32	Backward	0.200 [0.178, 0.222]	0.295 [0.292, 0.299]
DeepLIFT	Backward	0.197 [0.176, 0.217]	0.341 [0.338, 0.344]
GradientSHAP	Backward	0.226 [0.204, 0.246]	0.236 [0.233, 0.240]
SmoothGrad-16	Backward	−0.045 [−0.065, −0.023]	−0.015 [−0.018, −0.011]
RISE-512	Sampling	0.302 [0.283, 0.321]	0.179 [0.175, 0.183]
RISE-U-512	Sampling	0.294 [0.274, 0.315]	0.111 [0.107, 0.115]
Endpoint-only reference			
Endpoint M	Endpoint only	0.324 [0.303, 0.344]	0.371 [0.366, 0.377]
Deletion-game reference			
KernelSHAP-512	Sampling	0.299 [0.278, 0.319]	0.447 [0.443, 0.451]

Table 7: FunnyBirds native Single Deletion reproduction and native-target audit. **Top:** the official RISE-6000 setting reproduces the published Single Deletion scores within 0.005 on all three architectures. **Bottom:** raw Spearman correlation with the same native part-removal target. Endpoint M is included to show the strong target alignment of this evaluation. These native-target results are a reproduction and boundary check, not a replacement for the held-out FunnyBirds comparison in Table 2.

Architecture	Published RISE SD	Reproduced RISE SD	$\lvert\Delta\rvert$
ResNet-50	0.560	0.558	0.002
VGG16	0.730	0.726	0.004
ViT-B/16	0.790	0.788	0.002

Native-target raw Spearman ρ

Method	ResNet-50	VGG16	ViT-B/16
Endpoint M	0.757	1.000	0.983
DECAF-3	0.484	0.926	0.909
DECAF-5	0.565	0.898	0.864
DECAF-9	0.616	0.901	0.850
RISE-6000	0.115	0.452	0.576

Native-protocol reproduction check. The FunnyBirds column above evaluates attribution transfer to two held-out intervention operators, rather than the dataset's native Single Deletion protocol. We therefore ran a separate implementation check using the official FunnyBirds setting: the released checkpoints, native semantic part removals, and RISE with 6,000 masks, an 8×8 coarse grid, and mask probability $p = 0.1$ (Hesse et al., 2023). As shown in Table 7, the reproduced RISE scores differ from the published values by less than 0.005 for every architecture. This check supports the implementation fidelity of the RISE baseline used in our benchmark; it does not imply that the held-out score of 0.302 should numerically match the native Single Deletion scores, because the two evaluations use different intervention targets and RISE configurations.

The native and held-out FunnyBirds experiments answer different questions. The native target is defined by the same semantic part-removal contrast used to construct the endpoint, whereas the main benchmark evaluates transfer to held-out inpainting and texture-replacement interventions. The native audit therefore verifies baseline implementation and exposes the target-aligned boundary; the held-out benchmark remains the test of attribution transfer.

Table 8: Paired endpoint-versus-trajectory tests. Differences are DECAF minus endpoint M. Positive values indicate ordinary-attribution information beyond the endpoint.

Dataset	Contrast	Mean Δ	95% CI	Model wins
FunnyBirds	DECAF-3 $-M$	+0.0488	[0.0334, 0.0642]	2/3
FunnyBirds	DECAF-5 $-M$	+0.0797	[0.0637, 0.0962]	2/3
FunnyBirds	DECAF-9 $-M$	+0.0828	[0.0671, 0.0993]	2/3
ImageNet-1k	DECAF-3 $-M$	-0.0119	$[-0.0142, -0.0096]$	1/3
ImageNet-1k	DECAF-5 $-M$	-0.0045	$[-0.0072, -0.0017]$	1/3
ImageNet-1k	DECAF-9 $-M$	+0.0073	[0.0045, 0.0101]	2/3

Table 9: Endpoint-versus-trajectory ablation by architecture. Best Δ is the largest DECAF-3/5/9 Spearman minus endpoint M.

Dataset	Model	Endpoint M	DECAF-3	DECAF-5	DECAF-9	Best Δ
FunnyBirds	ResNet-50	0.388	0.425	0.447	0.451	+0.064
FunnyBirds	VGG-16	0.306	0.283	0.298	0.293	-0.008
FunnyBirds	ViT-B/16	0.277	0.409	0.465	0.475	+0.197
ImageNet-1k	ResNet-50	0.384	0.362	0.370	0.387	+0.002
ImageNet-1k	VGG-16	0.671	0.640	0.640	0.644	-0.027
ImageNet-1k	ViT-B/16	0.058	0.077	0.090	0.105	+0.047

H.3 Endpoint versus trajectory

FunnyBirds is the clean trajectory-value test because its evaluation operators differ from the explanation endpoint. All three DECAF grids significantly exceed M. Five stages capture 96.3% of the nine-stage gain over the endpoint. IDSDS evaluates the same deletion contrast that defines d_k. Endpoint M is therefore already highly aligned with its target. Three and five stages preserve most of that signal, while nine stages add a small but stable gain.

The trajectory gain is architecture-dependent in the tested models. It is largest for ViT-B/16 on both datasets, modest for ResNet-50, and negative for VGG-16. This pattern is empirical rather than a general architecture law, but it shows that the value of intermediate responses is not uniform across model classes.

The separate native FunnyBirds audit in Appendix H.2 provides the complementary endpoint-aligned case: when the evaluation target is the exact native part-removal contrast, endpoint M reaches $\rho = 0.913$ and exceeds all trajectory summaries.

Together, the two FunnyBirds evaluations isolate the role of the trajectory: the endpoint best recovers its own intervention effect, while intermediate responses add value when attribution must transfer to a different intervention.

H.4 Full-scale ImageNet validation

The 10,000-image conclusions are not a favorable-subset artifact. The full validation set preserves the ordering $M >$ DECAF-5 $>$ IG-U-32 $>$ IG-32. The full-scale paired difference between DECAF-5 and M is -0.0075 with a 95% interval of $[-0.0086, -0.0062]$.

H.5 Measured compute and large-model scaling

DECAF-3 is the speed-oriented Pareto point. Relative to IG-32, it has higher quality on both datasets, is $1.30\times$ faster, and uses $4.50\times$ less peak memory. DECAF-5 trades additional latency for higher quality while keeping the same inference-scale memory. KernelSHAP-512 is $5.89\times$ slower than DECAF-5 and uses six times as many forward rows.

H.6 PartImageNet as a task-aligned boundary case

PartImageNet supplies semantic part masks and evaluates held-out part-removal effects. Direct part removal and coalition methods are therefore unusually aligned with the evaluation target. We retain the benchmark as a boundary case rather than using it to define the main general-purpose comparison.

Table 10: Full 50,000-image ImageNet validation scale check. Values are equal-architecture macro-average IDSDS over 115,876 correctly classified model–image units.

Method	Macro ρ [95% CI]	ResNet-50	VGG-16	ViT-B/16
Endpoint M	0.3708 [0.3685, 0.3731]	0.3778	0.6731	0.0615
DECAF-5	0.3633 [0.3613, 0.3653]	0.3609	0.6395	0.0894
IG-U-32	0.2942 [0.2926, 0.2958]	0.2522	0.4939	0.1366
IG-32	0.2397 [0.2380, 0.2411]	0.1941	0.4002	0.1247

Table 11: Measured ImageNet-1k compute. Rows per image count model-forward input rows after batching. Timing is compute-only and macro-averaged across ResNet-50, VGG-16, and ViT-B/16.

Method	Access	Rows/image	Backward?	ms/image ↓	Peak GiB ↓
DECAF-3	Forward only	51	No	22.1	11.2
DECAF-5	Forward only	85	No	36.7	11.2
DECAF-9	Forward only	153	No	65.9	11.2
Input×Gradient	Backward	1	Yes	1.4	2.3
DeepLIFT	Backward	2	Yes	3.1	6.6
GradientSHAP	Backward	16	Yes	14.4	25.9
IG-32	Backward	32	Yes	28.7	50.4
IG-U-32	Backward	32	Yes	28.7	50.4
SmoothGrad-16	Backward	16	Yes	16.1	25.9
RISE-512	Sampling	512	No	216.6	14.4
KernelSHAP-512	Sampling	512	No	216.4	14.4

Table 12: PartImageNet strict-common-support Spearman. Part-removal and coalition methods exploit the supplied semantic part groups and are directly aligned with the held-out part-removal target.

Method	Spearman [95% CI]
Part-LIME-1000	0.478 [0.458, 0.497]
Part Occlusion	0.477 [0.458, 0.496]
Exact Part-Shapley	0.433 [0.412, 0.454]
KernelSHAP-512	0.426 [0.405, 0.447]
Endpoint M	0.364 [0.342, 0.385]
DECAF-9	0.363 [0.341, 0.384]
DECAF-5	0.358 [0.337, 0.379]
DECAF-3	0.350 [0.329, 0.370]
RISE-512	0.299 [0.276, 0.320]
IG-32	0.289 [0.270, 0.311]

This result marks the method boundary clearly. When perfect semantic parts are already supplied and the target is itself a part-removal effect, direct part interventions can be stronger. Even in that setting, the DECAF trajectories remain above standard gradient-path attribution and preserve their forward-only memory advantage.

I Covertype Experimental Details and Complete Results

This appendix reports the complete Covertype audit behind Section 8. The experiment uses a natural tabular base and controlled context–factor channels. Its purpose is not to assume that every treatment succeeds. It measures which response mechanism each trained model actually realizes under a strict common-start input-space reveal.

I.1 Setup and Operational Outcomes

We balance classes 1 and 2 from Covertype (Blackard, 1998) and retain 240,000 examples. The 54 natural features are augmented by one binary context and one binary candidate factor. The direction module uses context G and factor Z; the fragility module uses context H and factor U. Natural features, data splits, 135 trained estimators, endpoint definitions, and behavioral outcomes are frozen across the original and corrected analyses. No primary model is retrained.

Table 13: Covertype benchmark design. Each model family uses three seeds. Direction experiments additionally use shortcut strengths $p \in \{0.75, 0.95\}$. Both modules use the strict common-start stochastic input reveal.

Module	Regimes	Held-out operational behavior	Families	Models	Path
Direction	Direct, Gate, Invert	preserve / collapse / invert	5	90	common-start input reveal
Fragility	Robust, Mild, Fragile	endpoint-null prediction change	5	45	common-start input reveal
Total				135	

For the direction module, the endpoint context is $G = +1$, the alternate context is $G = -1$, and the candidate factor is $Z \in \{-1, +1\}$. Both branches begin at the same legal baseline $(X, -1, Z_0)$ with a shared $Z_0 \in \{-1, +1\}$. Shared monotone clocks reveal the factor and endpoint context. For each realization, the paired response R_t takes the values

$$R_t \in \{0, d_{\text{alt}}, d_{\text{end}}\}, \qquad \Pr(R_t = 0) = 1-t, \quad \Pr(R_t = d_{\text{alt}}) = t(1-t), \quad \Pr(R_t = d_{\text{end}}) = t^2.$$

The fragility module replaces (G, Z) with (H, U). Every nonzero state is obtained from one of four legal binary support queries. We route each realized response before taking its expectation, rather than routing the mean signed response. This construction guarantees $r(0) = 0$ and $r(1) = d_{\text{end}}$ while keeping all intermediate inputs on the binary support. The formal audit covers all 135 estimators and verifies common start, endpoint agreement, legal inputs, route-before-expectation, and conservation to machine precision.

For the direction module, preservation, collapse, and inversion are defined from held-out signed effects under the endpoint and alternate contexts. For the fragility module, the primary behavior is the prediction-change rate under $H = -1$ among examples satisfying the endpoint-null gate. The main uncertainty estimates use 500 joint model-family/seed cluster-bootstrap repetitions.

I.2 Complete Behavior Alignment

Table 14 reports the complete rank comparison. Evidence uses preservation, contradiction uses actual inversion, and fragility uses endpoint-null alternate-context prediction change. The first two columns use 90 direction models; the last uses 45 fragility models. Parenthetical counts mark methods with partial coverage.

Table 14: Spearman correlation with realized behavior under the corrected reveal. SHAP interaction is available for random forests and XGBoost; KernelSHAP and LIME use their fixed formal subsets. F uses the endpoint-null behavior target.

Method	Preservation ρ	Actual inversion ρ	Endpoint-null change ρ
DECAF component	0.747	**0.989**	**0.973**
Endpoint M	0.657	-0.001	0.148
Abs	0.694	-0.079	0.499
Signed net	0.804	-0.162	0.038
SignFlip	-0.436	0.972	0.867
OppMass	-0.430	**0.989**	0.941
Native SHAP	0.781	-0.205	0.481
SHAP interaction	-0.251 (54)	0.093 (54)	0.069 (27)
KernelSHAP	0.811 (30)	-0.219 (30)	0.441 (15)
Global PFI	0.664	-0.029	0.676
Context-conditioned PFI	0.298	0.311	0.739
PDP/ALE interaction	-0.100	0.827	0.937
LIME	0.845 (30)	-0.384 (30)	0.506 (15)

Joint family/seed cluster bootstrap gives

$$\rho(E, V_{\text{keep}}) = 0.7473, \qquad 95\%\ \text{CI} = [0.6466, 0.8363],$$

$$\rho(C, V_{\text{inv}}) = 0.9886, \qquad 95\%\ \text{CI} = [0.9764, 0.9983],$$

and

$$\rho(F, V_{\text{null}}) = 0.9732, \qquad 95\%\ \text{CI} = [0.9427, 0.9872].$$

A hash-defined cross-fitted audit computes DECAF on one half of each model's test rows and behavior on the other, then swaps the halves. The resulting correlations are 0.748 for evidence,

0.989 for contradiction, and 0.974 for fragility, with 95% intervals $[0.645, 0.836]$, $[0.977, 0.997]$, and $[0.949, 0.983]$. Thus all three associations persist without sharing observations between the response summary and behavioral outcome.

I.3 The Same Treatment Produces Different Realized Mechanisms

Table 15 shows why treatment labels cannot serve as mechanism ground truth. The Invert generator produces genuine reversal in random forests and multilayer perceptrons, but mostly collapse in the remaining families. The Fragile generator becomes endpoint evidence in logistic regression, while nonlinear families retain endpoint-null mass.

Table 15: Model-family audit. Direction columns average the Invert regime across two strengths and three seeds. Fragility columns average the Fragile regime across three seeds. "Active" is the endpoint-active fraction; "Null change" is the alternate-context prediction-change rate on endpoint-null examples.

	Invert treatment			Fragile treatment				
Model family	C	Invert rate	Collapse rate	M	E	F	Active	Null change
HistGradientBoosting	0.0000	0.000	1.000	0.0198	0.0269	0.0607	0.340	0.623
Logistic regression	0.0000	0.000	1.000	0.2991	0.1495	0.00007	0.989	0.000
MLP	0.0559	0.743	0.257	0.0279	0.0337	0.0272	0.350	0.219
Random forest	0.0152	0.816	0.169	0.0565	0.0664	0.0213	0.673	0.327
XGBoost	0.00001	0.002	0.998	0.0201	0.0333	0.0551	0.346	0.552

Contradiction is largest in the Invert regime for the two families that realize substantial reversal: random forests and multilayer perceptrons. Histogram gradient boosting and logistic regression collapse the candidate effect, and XGBoost almost entirely collapses it. In the fragility module, both corrected F and endpoint-null behavior increase from Robust to Fragile in four of the five families. Logistic regression is the exception because it learns the candidate factor as endpoint evidence rather than endpoint-null sensitivity.

I.4 Fixed Semantic Readout and Magnitude Conditioning

The fixed semantic benchmark assigns E, C, and F to Evidence, Contradiction, and Fragility without fitting a meta-classifier. Table 16 reports both the unconditional result and the within-Abs-bin result. The latter controls response magnitude through 12 quantile bins.

Table 16: Mechanism identification with fixed semantics and after conditioning on ordinary magnitude. Learned references use mechanism labels and are not access-matched to fixed DECAF.

Method	Fixed macro-AUROC	Fixed macro-AUPRC	Within-Abs macro-AUROC	Within-Abs macro-AUPRC
DECAF (fixed $E/C/F$)	**0.818**	**0.781**	**0.950**	**0.956**
Endpoint M	0.487	0.437	0.465	0.633
Abs	0.504	0.438	0.464	0.608
Native SHAP	0.523	0.437	0.467	0.618
SHAP interaction	0.540	0.440	–	–
Context-conditioned PFI	0.556	0.545	0.534	0.647
PDP/ALE interaction	0.539	0.438	0.575	0.700
Strong tabular reference (supervised)	0.968	0.937	1.000	1.000
DECAF probe (supervised)	0.903	0.860	0.964	0.969
Combined empirical reference (supervised)	0.975	0.948	1.000	1.000

The fixed DECAF mechanism accuracy is 0.507. The supervised DECAF probe reaches accuracy 0.880, and the combined empirical reference reaches 0.907. These learned references use leave-one-model-family-out calibration, mechanism labels, and multiple input summaries. They bound cross-family decodability but do not replace the label-free comparison. The corrected within-Abs result remains strong: fixed DECAF reaches macro-AUROC 0.950 and macro-AUPRC 0.956, whereas

endpoint magnitude, ordinary magnitude, and native SHAP reach macro-AUROC 0.465, 0.464, and 0.467.

I.5 Measured Cost and SHAP-Interaction Audit

Table 17 reports the historical registered cumulative worker time. The corrected protocol preserves the same four model queries per natural row, so the query and predicted-row accounting is unchanged. We retain the historical timing contract rather than mixing it with the resumed correction run on different hardware.

Table 17: Complete registered method cost. SHAP interaction uses 128 stratified examples per tree model, split into four 32-example shards.

Method	Models	Worker-s/model	Relative to DECAF	Predicted rows
DECAF / M / Abs / Net / OppMass / SignFlip	135	1.42	1.0×	25,920,000
PDP/ALE interaction	135	2.22	1.57×	38,880,000
LIME	45	8.38	5.92×	11,796,480
PFI / context PFI	135	13.32	9.41×	207,360,000
Retraining reference	135	19.70	13.91×	–
KernelSHAP	45	45.30	31.99×	377,501,760
Native SHAP	135	387.65	273.8×	–
SHAP interaction	54	15097.63	10662×	–

The formal SHAP-interaction run completed 216 shards over 54 tree models. It consumed at least 811,875 CPU-seconds (225.5 CPU-hours), used up to 32 concurrent shards, and required 17.71 hours of elapsed stage time. Its inversion correlation was only 0.093 with a 95% interval spanning zero. The corrected reveal retains four legal support queries per natural row and preserves endpoint M exactly for all 135 estimators.

I.6 Matched-Pair Audit and Its Limits

The corrected deterministic matcher produces only 10 protocol pairs, covering 26.7% of the 75 primary semantic units. Nine pairs are within the same model family, but the median relative endpoint-magnitude difference remains 0.401 and the maximum is 0.817. Fixed DECAF reaches matched-pair mechanism accuracy 0.450. This set therefore remains too small and too poorly matched in endpoint magnitude for a headline claim. We use the much better populated within-Abs-bin analysis in Table 16 and retain the pair result only as a support audit.