LOST IN RANKING: WHEN ATTRIBUTION SCORES HIDE CLASS CONTRAST

Anonymous authors
Paper under double-blind review

ABSTRACT

Attribution methods assign spatial importance for a requested class, but scoring each class map separately can conceal distinctions already present in their difference. We show that this loss changes model comparisons and derive an exact decomposition of its two sources: spatial content common to both maps alters their individual rankings but cancels on subtraction; maps highlighting different objects supply complementary evidence when the evaluated regions include background. In vision-transformer experiments on structured and randomly sampled image pairs, contrast scoring raises gradient–attention rollout's mean source AUC from 0.647 to 0.814 and from 0.713 to 0.891, respectively. A supported advantage of one model under separate-map scoring becomes unresolved under contrast scoring in both populations. Common-component removal accounts for 69–78% of the aggregate scoring gaps for rollout and Transformer Attribution. We also prove that any nonconstant contrast-only score on unrestricted maps must depend on the maps' relative amplitude. Mass, sign, and spatial-support controls establish where the conclusions transfer. The resulting evaluation principle is concrete: retain the maps, declare their amplitude convention, and form the class contrast before reducing it to a spatial score.

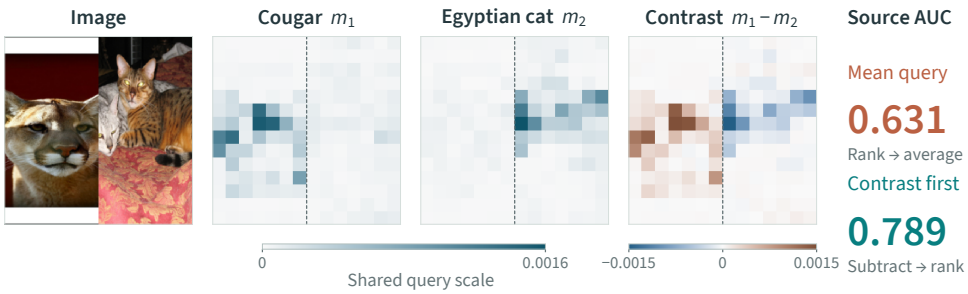

Figure 1: **The same maps contain more class separation than their separate rankings reveal.** Transformer Attribution on a cougar/Egyptian-cat composite scores 0.631 when the two source-localization AUCs are averaged, and 0.789 when their difference is ranked. Each map assigns importance for one requested class; warm/cool contrast favors cougar/Egyptian cat. Native 14×14 fields are unsmoothed, the query maps share a display scale, and dashed lines mark the source boundary. This ViT-B example is selected near the median evaluator gain, not as a prevalence estimate (Appendix A.2.5).

1 INTRODUCTION

An attribution benchmark can miss the distinction it is meant to measure. Given an image containing two classes, a method returns one importance map for each requested class. Ranking each map against its own region measures separate localization. Ranking the difference between the maps measures relative evidence for one class over the other. The distinction matters: the maps in Figure 1 separate their sources substantially better than their individual rankings suggest.

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.

It also changes comparative conclusions. Across two separately sampled populations, gradient–attention rollout on ViT-B AugReg-1k has a supported advantage over AugReg-21k under separate-map AUC. Scoring the difference between each model's class maps leaves that advantage unresolved. The maps, images, and checkpoints are unchanged; only the evaluator's order of operations differs. Separate AUC values cannot recover the contrast score once the maps have been discarded.

We identify two mechanisms. A spatial field common to both maps can alter their rankings while cancelling in the difference. Separately, even disjoint maps that perfectly highlight their respective objects leave background tied with the other object in each individual map. Subtraction combines their complementary evidence. A low separate-map score cannot distinguish these cases. An exact three-score decomposition does: score the original maps, remove their common minimum and rescore, then rank their unchanged difference.

The paper makes three contributions. First, we characterize the information retained by contrast scoring, decompose its disagreement with separate ranking, and calibrate the terms on ideal object maps. Second, we prove that a nonconstant contrast-only statistic cannot also ignore independent map scales, making an amplitude convention necessary. Third, experiments on structured and random class pairs connect the decomposition to changed model comparisons, with controls for attribution mass, sign, and spatial support. The practical outcome is a rescoreable evaluation protocol for a specified class-contrast question.

Localization and class sensitivity. Grid Pointing Game (GridPG) and Focus score positive attribution mass in a requested class's grid cells or mosaic regions (Böhle et al., 2021; Rao et al., 2022; Arias-Duart et al., 2022); negative attribution can also be evaluated (Fresz et al., 2024). Our main statistic instead ranks locations on a two-source partition. These evaluators share separate-query assessment, but have different ideal-map baselines. Source membership is also distinct from a causal attribution target: the disconnected-input controls of Rao et al. (2022) address whether another source can influence a prediction. Relatedly, modified backpropagation can lose class and layer sensitivity (Sixt et al., 2020), and common additions to logits can alter gradients while preserving softmax predictions (Srinivas & Fleuret, 2021). Our decomposition identifies how common numerical content affects a spatial score; its semantic origin requires a separate analysis.

Contrastive explanations and evaluation. Contrastive excitation backpropagation (Zhang et al., 2016), contrastive LRP (Gu et al., 2019), weighted class contrasts (Wang & Wang, 2022), two-stage explanations (Xie et al., 2023), and class-discriminative attention maps (Brocki et al., 2024) already distinguish alternatives. We analyze the evaluator applied to fixed maps, including its exact equivalence to scoring an antisymmetric explanation pair. This complements prediction-gap evaluation under input masking (Elisha et al., 2026) and comparisons of original and adversarial maps (Lenhard et al., 2026). The question here is which class distinction survives the reduction from maps to scores.

2 Contrast before ranking

2.1 Two questions, two scores

Let $m_1, m_2 \in \mathbb{R}^\Omega$ be maps for two requested classes on a common finite grid. The binary target y marks source region 1; both regions are nonempty. ROC AUC is the fraction of cross-region token pairs ranked correctly, with half credit for ties (Mann & Whitney, 1947; Fawcett, 2006). Define

$$B_{\mathrm{Q}}(y; m_1, m_2) = \tfrac{1}{2}\{\mathrm{AUC}(y, m_1) + \mathrm{AUC}(1-y, m_2)\}, \tag{1}$$

$$B_{\mathrm{C}}(y; m_1, m_2) = \mathrm{AUC}(y, m_1 - m_2). \tag{2}$$

Mean-query AUC, B_{Q}, asks whether each map ranks its own source higher. *Contrast-first AUC*, B_{C}, asks whether relative evidence for query 1 ranks its source higher. Since complementing the target complements AUC,

$$B_{\mathrm{Q}} = \tfrac{1}{2} + \tfrac{1}{2}\{\mathrm{AUC}(y, m_1) - \mathrm{AUC}(y, m_2)\}. \tag{3}$$

The distinction is a difference of AUCs versus the AUC of a difference. Both give $1/2$ to every equal-map pair, so that null check cannot distinguish them. Nor can the two individual AUCs determine B_{C}: with $y = (1, 0)$, the pairs $((2, 0), (1, 0))$ and $((1, 0), (2, 0))$ have the same individual-query scores $(1, 0)$, but contrast scores 1 and 0.

Write the symmetric field and half-difference as

$$S = \frac{m_1 + m_2}{2}, \qquad D = \frac{m_1 - m_2}{2}, \qquad (m_1, m_2) = (S + D, S - D). \tag{4}$$

Then

$$B_{\mathrm{C}}(y; m_1, m_2) = B_{\mathrm{Q}}(y; D, -D). \tag{5}$$

Contrast-first scoring is exactly the original evaluator applied to an antisymmetric explanation pair. It preserves a distinction that separate scalar scores can lose.

2.2 Two mechanisms of evaluator disagreement

For nonnegative maps, their pointwise minimum $C = \min(m_1, m_2)$ is common numerical content. Removing it leaves opposite nonnegative parts:

$$S = C + |D|, \qquad (m_1 - C, m_2 - C) = ((m_1 - m_2)_+, (m_2 - m_1)_+), \tag{6}$$

where $(t)_+ = \max(t, 0)$. The residuals have disjoint support and the original difference. Let $B_{\mathrm{H}} = B_{\mathrm{Q}}(y; (m_1 - m_2)_+, (m_2 - m_1)_+)$ be their mean-query score. The exact decomposition

$$B_{\mathrm{C}} - B_{\mathrm{Q}} = \underbrace{B_{\mathrm{H}} - B_{\mathrm{Q}}}_{\text{common-component removal}} + \underbrace{B_{\mathrm{C}} - B_{\mathrm{H}}}_{\text{combining one-sided evidence}} \tag{7}$$

separates two operations that preserve the contrast. The first removes the common minimum; the second combines the residuals before ranking. These are signed score changes, not causal effects or fractions of semantic information. The identity also holds for signed maps, whose minimum need not represent common nonnegative content. Notice that S itself is not spatial overlap: it includes $|D|$ even when the maps are disjoint.

Proposition 2.1 (Calibration for ideal object maps)**.** *Suppose m_k is the indicator of an object entirely within source k, occupying fraction f_k of that source. On the whole-source target,*

$$B_{\mathrm{Q}} = B_{\mathrm{H}} = \tfrac{1}{2} + \tfrac{1}{4}(f_1 + f_2), \tag{8}$$

$$B_{\mathrm{C}} = \tfrac{1}{2} + \tfrac{1}{2}(f_1 + f_2 - f_1 f_2), \qquad B_{\mathrm{C}} - B_{\mathrm{Q}} = \tfrac{1}{4}(f_1 + f_2) - \tfrac{1}{2} f_1 f_2. \tag{9}$$

At $f_1 = f_2 = 0.3$, these perfectly class-specific maps score 0.65 under B_{Q} and 0.755 under B_{C}, although $C = 0$. In either individual map, its own background ties with the other object; the other map resolves that comparison. Restricting this ideal target to object tokens gives both scores 1. Positive-attribution mass also gives both maps full credit for any positive object area. Thus the one-sided term calibrates a property of the spatial rank target, not a failure of class specificity. Appendix A.1.4 proves the formulas by counting object/background pairs.

Rank separation is also distinct from contrast magnitude. If $m_2 = W$ and $m_1 = W + \epsilon y$, where W has distinct cross-source values and $\epsilon > 0$ is smaller than the smallest absolute cross-source difference, then $B_{\mathrm{Q}} = 1/2$ and $B_{\mathrm{C}} = 1$ even as $\|m_1 - m_2\|_\infty = \epsilon \to 0$. We therefore report common-mass diagnostics alongside rank diagnostics (Section 4.4). Neither alone measures behavioral faithfulness.

2.3 Contrast specificity requires an amplitude convention

Proposition 2.2 (Contrast specificity)**.** *For fixed y, a statistic $T(m_1, m_2)$ depends only on D if and only if*

$$T(m_1 + E, m_2 + E) = T(m_1, m_2) \quad \textit{for every spatial field } E. \tag{10}$$

A common field cancels in the difference; conversely, choosing $E = -S$ leaves $(D, -D)$. Contrast AUC satisfies the criterion. Mean-query AUC does not: for $y = (1, 0)$ and $D = (1, 0)$, changing S from $(0, 0)$ to $(2, 0)$ changes B_{Q} from 1 to $1/2$, while B_{C} stays 1 (Figure 6).

Theorem 2.3 (Contrast specificity and independent scale invariance are incompatible)**.** *Fix y. Let $T : \mathbb{R}^\Omega \times \mathbb{R}^\Omega \to \mathbb{R}$ be invariant to every common-field addition. If $T(\alpha m_1, m_2) = T(m_1, m_2)$ for every map pair and every $\alpha > 0$, then T is constant.*

Proof. By Proposition 2.2, write $T(m_1, m_2) = G(m_1 - m_2)$. For arbitrary F, F', take $m_1 = F' - F$ and $m_2 = F' - 2F$. Their difference is F, whereas $2m_1 - m_2 = F'$. Scale invariance gives $G(F) = G(F')$. □

The theorem makes relative amplitude part of the contrast question. Independent positive scaling preserves each query's AUC, yet

$$\alpha m_1 - \beta m_2 = (\alpha - \beta)S + (\alpha + \beta)D \tag{11}$$

mixes the symmetric field into the difference when $\alpha \neq \beta$. Per-query L1, max-absolute, or min–max normalization therefore defines a different contrast. Invariance to a scale-erasing normalizer would imply the theorem's scaling hypothesis when its degenerate-input convention is also scale invariant. A single positive scaling after subtraction preserves contrast AUC, as do independent spatially constant offsets. The theorem concerns unrestricted real fields; opposite nonnegative parts are a restricted scale-invariant family (Appendix A.1.5).

2.4 How common content masks token comparisons

For a token p in source 1 and n in source 2, write $a = S_p - S_n$ and $b = D_p - D_n$. Mean-query scoring averages the ranking credits of $a + b$ and $b - a$. When $|a| > |b|$, these signs oppose: it gives half credit regardless of the contrast's orientation. Contrast scoring retains the sign of b.

With no cross-source query ties, let π be this strictly masked fraction and $a_{\text{in}}, a_{\text{out}}$ the tie-aware contrast accuracies inside and outside it. Then

$$B_{\text{Q}} = \tfrac{\pi}{2} + (1 - \pi)a_{\text{out}}, \qquad B_{\text{C}} - B_{\text{Q}} = \pi(a_{\text{in}} - \tfrac{1}{2}), \qquad B_{\text{Q}} \leq 1 - \tfrac{\pi}{2}. \tag{12}$$

An empty subset contributes zero to its weighted term. The identity explains both gains and losses: masking conceals incorrect comparisons too. The full tie-aware partition and a strict-loss example appear in Theorem A.1. Appendix A.1.8 connects contrast to operators linear in the explained function; Appendix A.1.9 extends the invariance criterion to multiple queries.

3 Empirical design

3.1 Fixed maps on two source populations

We compose two ImageNet validation photographs into a 224×224 image (Deng et al., 2009; Russakovsky et al., 2015). Each source is square-resized with bilinear interpolation, normalized for its checkpoint, and centrally half-cropped. The 14×14 patch grid has 98 tokens per source. Its target is source membership, including background. Each checkpoint must predict distinct classes on the full sources and retain both among its top ten composite predictions; all six must agree on the ordered queries. This recognition filter precedes attribution capture and does not test object visibility after cropping. Queries are model predictions and need not match the source annotations.

Structured pairs. After excluding recorded development sources, 2,000 sampled photographs are sorted by their class-organized file index and paired adjacently. Of 1,000 candidates, 170 qualify side by side and 180 top/bottom, with 151 shared pairs. We retain every qualifier. The side-by-side pairs differ by at most four class-list positions; 21 share an annotated class but have distinct agreed predictions. These are the structured primary and secondary populations.

Random distinct-class pairs. A separate bank draws 2,000 ordered class pairs uniformly from distinct annotated classes, then unused photographs within each class; source order is independently randomized. The eligible frame contains 45,969 photographs across all 1,000 classes. Of 517 joint qualifiers, a prespecified cap retains the first 200 in saved order, covering 322 annotated classes. Sampling, exclusions, cap, and analysis were fixed before qualification. This study changes both photographs and pairing, testing transfer to a different qualified population (Appendix B.6).

3.2 Model panel and attribution constructions

The six fixed classifiers span architecture size and training history (Dosovitskiy et al., 2021; Steiner et al., 2022; Touvron et al., 2021; 2022): ViT-B with original ImageNet-21k pretraining; ViT-B

AugReg trained on ImageNet-1k or pretrained on ImageNet-21k; ViT-L AugReg with ImageNet-21k pretraining; DeiT-B; and DeiT III-B. Every 21k checkpoint is fine-tuned to the same ImageNet-1k output task. All use 224×224 inputs and 16-pixel patches. The base models have 12 blocks and 768-dimensional states; ViT-L has 24 blocks and width 1024. The central AugReg comparison holds the ViT-B architecture fixed across two training histories; it is a comparison of these checkpoints, not an isolated causal effect of pretraining data. Exact loader tags are in Appendix C.1.

Gradient–attention rollout composes row-normalized $I + \mathrm{mean}_h \, \mathrm{ReLU}(A^h \odot \nabla_{A^h} z_c)$ across layers from block 0, where A^h is head h's attention and z_c the requested class logit. Its class-row patch field is sum-normalized. This variant of the gradient–attention product (Chefer et al., 2021a) has explicit row and return normalization. *Transformer Attribution* (TA) uses the official implementation of Chefer et al. (2021b) on its supported ViT-B original and DeiT-B checkpoints. Contrasts use the declared returns; rollout's saved fields have unit mass by construction.

Controlled sign and scale. A final-layer projection isolates return conventions:

$$C_j^{\mathrm{val}} = \sum_a \widetilde{W}_O^a v_j^a, \qquad f_c(j) = \langle \nabla_h \log p_c, C_j^{\mathrm{val}} \rangle. \tag{13}$$

Here v_j^a is token j's value vector in head a, $\widetilde{W}_O^a$ its output projection including LayerScale, and h the final class-token state before the last normalization. The derivative passes through that normalization and the classifier; p_c is the softmax probability. The field measures value content aligned with a local class direction. It omits attention weights and differs from CDAM's token-gradient construction (Brocki et al., 2024).

We cross signed/rectified returns with native/per-query L1 amplitude, rectifying before division. The native difference $f_c - f_d$ equals $\langle \nabla_h (z_c - z_d), C_j^{\mathrm{val}} \rangle$ because the common log-softmax derivative cancels. Independently computing both forms checks this identity (Appendix A.1.7).

3.3 Paired analysis

Locally frozen plans specify four primary effects: $B_{\mathrm{C}} - B_{\mathrm{Q}}$ for rollout and TA, signed minus rectified native B_{Q}, and native minus independently L1-normalized signed B_{C}. We average fixed checkpoints equally within each composite and bootstrap composites, preserving all paired fields. The structured study uses 5,000 draws and the random study 50,000. We report 95% percentile intervals and 98.75% Bonferroni marginal intervals for each four-effect family. The latter were prespecified for random pairs and computed as a 50,000-draw sensitivity analysis for structured pairs. Coverage is approximate and conditional on the fixed models and qualified populations; the populations are analyzed separately.

Decompositions and matched comparisons use pointwise 95% intervals from 5,000 draws. We call an effect *supported* when its stated interval excludes zero, and *unresolved* when it includes zero. Structured mass, box-support, scale-sweep, and development controls are exploratory; the random study prespecified its supporting decomposition and mass analyses. All scores use native token support and exact ties, checked independently. Appendices A, B, and C collect supporting science, methods, and execution/provenance, respectively.

4 Results

4.1 Class separation survives in maps that separate scoring understates

The scoring gap recurs in both populations (Table 1). Rollout's mean-query and contrast AUCs are 0.6465 and 0.8144 on structured pairs, and 0.7126 and 0.8911 on random pairs. TA's corresponding scores are 0.6683 and 0.8181, and 0.7252 and 0.8917. All four primary effects have positive lower bounds under the 98.75% marginal intervals in each population (Appendices A.4.4 and A.3.1).

The decomposition identifies different mechanisms across constructions (Figure 2). For rollout, removing the common minimum contributes 0.1289 on structured pairs and 0.1398 on random pairs: 76.7% and 78.3% of the aggregate gaps. One-sided evidence supplies the remaining 0.0390 and 0.0387. TA's common-removal shares are 69.5% and 68.8%. Both terms have positive pointwise intervals in both populations.

Table 1: **All four planned effects persist across source populations.** AUC differences with paired 95% percentile intervals: 5,000 draws for structured pairs, 50,000 for random pairs. Rollout and projection effects average six fixed checkpoints; TA averages its two supported checkpoints.

Effect	Structured ($n = 170$)	Random ($n = 200$)
Rollout: $B_C - B_Q$	$0.1679\ [0.1611, 0.1747]$	$0.1786\ [0.1739, 0.1832]$
TA: $B_C - B_Q$	$0.1498\ [0.1412, 0.1580]$	$0.1665\ [0.1611, 0.1719]$
Signed − rectified B_Q	$0.0415\ [0.0392, 0.0436]$	$0.0453\ [0.0436, 0.0471]$
Native − L1 signed B_C	$0.0064\ [0.0046, 0.0084]$	$0.0062\ [0.0046, 0.0080]$

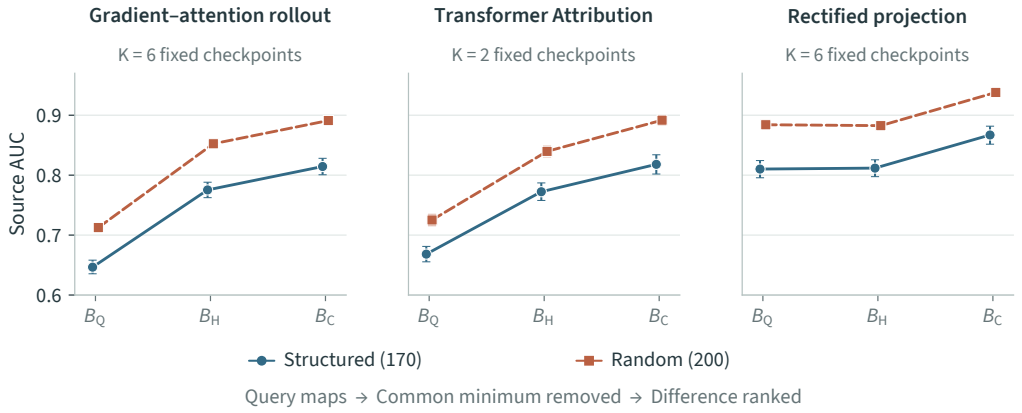

Figure 2: **Common removal dominates for attention maps; complementary evidence dominates for the rectified projection.** Each path follows $B_Q \to B_H \to B_C$: score query maps, remove their common minimum and rescore, then rank their difference. Common removal accounts for 69–78% of rollout/TA's aggregate gaps across the two populations, but has little effect for the projection. Points and bars show means and pointwise 95% composite-bootstrap intervals. K is checkpoint support; comparisons between constructions require matched checkpoints (Figure 3).

Rectified projection behaves differently. Its common-removal term is $0.0016\ [-0.0001, 0.0034]$ on structured pairs and $-0.0015\ [-0.0024, -0.0006]$ on random pairs, while its one-sided terms are 0.0553 and 0.0554. Common removal can lower a score even when the full contrast raises it. The triplet separates that behavior from the ideal-map effect.

The difference between constructions persists when checkpoint support is held fixed (Figure 3). On both ViT-B original and DeiT-B, rollout and TA gain mainly through common removal, while rectified projection gains mainly from combining one-sided evidence. This comparison holds images, queries, and checkpoint fixed; it does not depend on pooling six checkpoints for one construction and two for another.

The token-pair identity explains the lost rankings. Rollout strictly masks 64.3% of cross-source comparisons on structured pairs and 53.7% on random pairs. Their contrast accuracies are 76.1% and 83.3%, whereas mean-query scoring gives each half credit. Rollout and TA have no query ties in either population. Gains are positive at every checkpoint on average, but individual losses remain: rollout improves in 1,009/1,020 structured and 1,198/1,200 random checkpoint/composite observations; TA improves in 332/340 and 400/400. These are descriptive counts within fixed panels.

4.2 Sign, amplitude, and target determine what transfers

Retaining sign in the controlled projection raises mean-query AUC by 0.0415 on structured pairs and 0.0453 on random pairs. On structured pairs, contrast AUC rises by only $0.0059\ [0.0042, 0.0076]$: subtraction recovers much of the ranking information lost by rectification. Native log-probability contrast and the independently computed pairwise projection have identical cross-source rank relations in both populations.



Figure 3: **The two mechanisms persist on matched checkpoint support.** Each panel holds the checkpoint, 170 composites, and queries fixed. Points show $B_{\mathrm{Q}}, B_{\mathrm{H}}, B_{\mathrm{C}}$ with pointwise 95% composite-bootstrap intervals and small horizontal offsets for legibility. Common removal contributes little for rectified projection on either checkpoint, consistent with its low common-mass share; mass and rank changes remain distinct quantities.

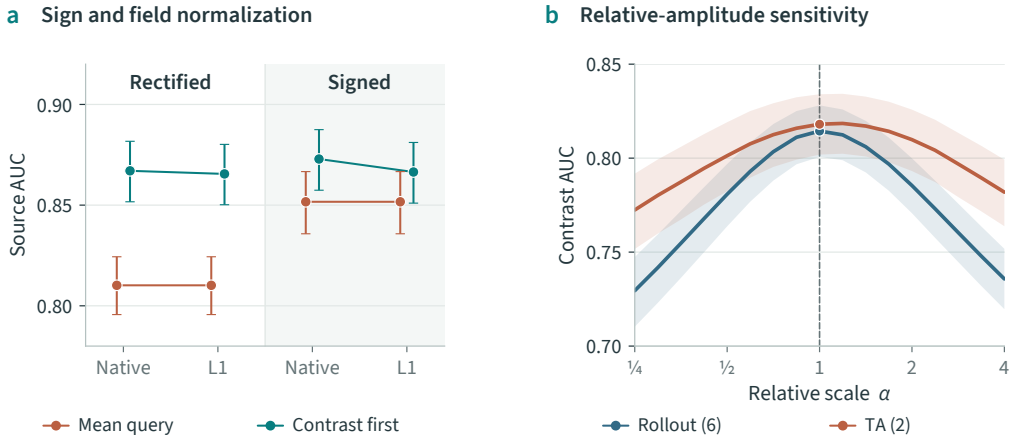

Figure 4: **A contrast score requires a declared amplitude convention.** Structured primary population. **a**, L1 scaling leaves query ranks unchanged but alters contrast AUC for the same signed or rectified projection. **b**, Contrast AUC for $\alpha m_1 - m_2$ varies with relative scale in rollout and TA. The dashed guide is the declared return, $\alpha = 1$; all 17 scales are shown. Bars and shading are pointwise 95% composite-bootstrap intervals.

Independent L1 normalization preserves each signed query's AUC but changes its contrast, as predicted by Theorem 2.3. Rollout and TA are also sensitive to relative amplitude across the 17-scale sweep (Figure 4). Primary scores use the declared returns at $\alpha = 1$, near the sampled maxima but not selected by target score.

Two controls delimit these results. On 32 separate DeiT-B development composites, attention Grad-CAM (Selvaraju et al., 2017; Chefer et al., 2021b) has a sign-by-evaluator interaction of -0.0077 $[-0.0162, -0.0004]$, opposite to the projection's; partial LRP has an unresolved evaluator gap (Appendix A.2.6). Sign sensitivity therefore depends on the attribution construction.

For spatial support, annotated boxes retain some area from both sources in 168/170 structured composites, and at least half their area in 154. On 134 query-matched composites with box-intersecting tokens from both sources, restricting both evaluators to those tokens reduces the projection's one-sided term by 0.0070, while rollout/TA one-sided-term changes remain unresolved. The comparison

holds images and checkpoints fixed. Boxes include background and do not certify object visibility; the crop example in Figure 10 shows why source qualification alone is insufficient (Appendix A.4.3). All four planned directions also persist in the top/bottom layout and on its 151 shared pairs (Tables 2 and 4).

4.3 THE EVALUATOR CHANGES WHICH ADVANTAGES ARE SUPPORTED

The practical consequence is a changed model comparison (Figure 5). Rollout on ViT-B AugReg-1k exceeds AugReg-21k under mean-query AUC in both populations; neither contrast interval resolves an advantage. The paired query-minus-contrast change is 0.0625 $[0.0522, 0.0728]$ on structured pairs and 0.0526 $[0.0456, 0.0596]$ on random pairs. These interaction intervals establish evaluator dependence; intervals spanning zero under contrast do not establish equivalence or reversal.

AugReg-1k masks fewer token comparisons. For the tie-free checkpoints, let Δ denote the first-minus-second difference and bars their arithmetic averages. Applying Equation (12) to both checkpoints gives

$$\Delta B_Q = -\Delta\pi\,(\bar{a}_{\text{out}} - \tfrac{1}{2}) + (1 - \bar{\pi})\,\Delta a_{\text{out}}. \tag{14}$$

The first term changes the masked share at the average outside-mask accuracy; the second changes that accuracy at the average unmasked share. This exact symmetric accounting has no replacement order. It assigns 0.0550 of the structured 0.0621 query advantage, and 0.0428 of the random 0.0495 advantage, to the masking-share term. Outside-mask accuracy supplies the rest. Thus the model comparison is tied directly to the ranking information that separate evaluation suppresses.

Method comparisons remain specific. TA's mean-query advantage over rollout on DeiT-B is unresolved under contrast AUC in both populations, but TA retains a contrast advantage on ViT-B original and in the matched mean on random pairs.

Positive-attribution mass provides a direct test beyond ranks. We compare the fraction of each map's positive mass in its own source with the same score applied to the positive and negative parts of their difference. TA retains a contrast-mass advantage on both supported checkpoints in both populations: the matched means are 0.0349 $[0.0275, 0.0426]$ and 0.0342 $[0.0283, 0.0403]$ for structured and random pairs. Neither mass score resolves a random-pair AugReg advantage (Appendix A.4.1). These outcomes differ from the AUC comparisons: the evaluator's metric, as well as its order of operations, determines which advantages are supported.

4.4 RANK GAINS AND COMMON MASS ANSWER DIFFERENT QUESTIONS

Common-mass share supplies amplitude information that the AUC decomposition discards. For nonnegative maps with positive combined mass, define

$$\omega = \frac{2\sum_j \min(m_{1j}, m_{2j})}{\sum_j (m_{1j} + m_{2j})}, \qquad 1 - \omega = \frac{\sum_j |m_{1j} - m_{2j}|}{\sum_j (m_{1j} + m_{2j})}. \tag{15}$$

The complement is the absolute contrast mass relative to the two maps' total mass. On random pairs, rollout's mean common-mass share is 60.5%, while common removal accounts for 78.3% of its aggregate AUC gap. TA's corresponding quantities are 28.8% and 68.8% (Table 10). A modest common-mass fraction can therefore have a large ranking effect; spatial variation and placement matter.

These are distinct diagnostics under the declared return convention. The score-gap share is signed accounting and can lie outside $[0, 1]$, whereas ω is a nonnegative overlap fraction. Neither identifies the common field's semantics. Conversely, the small-contrast example in Section 2.2 shows that nearly identical maps can yield perfect contrast ranks. Reporting both diagnostics prevents a ranking gain from being mistaken for strong class-specific attribution. Appendix A.4.2 gives checkpoint-level values and degenerate-case conventions.

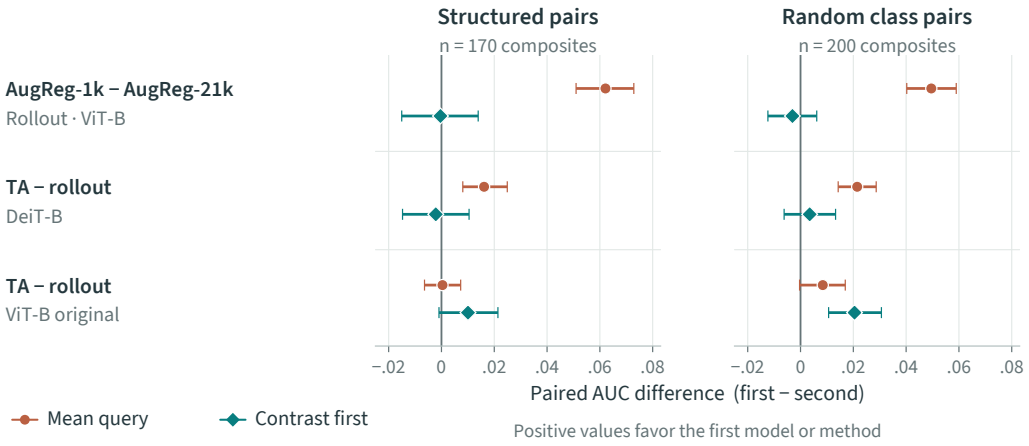

Figure 5: **A supported model advantage depends on whether the evaluator preserves class contrast.** Paired AUC differences on structured ($n = 170$) and random ($n = 200$) composites; bars are pointwise 95% composite-bootstrap intervals. Positive values favor the first named model or method. AugReg-1k's rollout advantage and TA's DeiT-B advantage are supported by separate-query scoring but unresolved under contrast scoring in both populations. TA retains a contrast advantage on ViT-B original in the random population. Each comparison holds checkpoint support, images, and query maps fixed across evaluators; Table 11 gives exact values and the matched mean.

5 Implications and Limits

A class-contrast evaluation should preserve the information needed to distinguish its queries. Three choices make that requirement operational:

1. **Retain maps on common support.** Separate scalar scores cannot reconstruct contrast rankings. Fix the input, queries, and spatial target before comparing methods.
2. **Declare relative amplitude.** Native outputs in common units, equal-mass returns, and equal-maximum returns define different contrasts. Fix the convention before consulting target scores; report sensitivity when its choice is not uniquely justified.
3. **Report the score triplet and matched comparisons.** $(B_{\mathrm{Q}}, B_{\mathrm{H}}, B_{\mathrm{C}})$ separates common removal from complementary evidence. Add a magnitude or common-mass diagnostic because ranking alone discards contrast strength.

This protocol complements work separating models, maps, metrics, and evaluation criteria (Kümmerer et al., 2018; Boggust et al., 2023; Hedström et al., 2023; Yeh et al., 2019). Appendix B.10 gives the reporting card, and Appendix A.5 shows why fixing query identity also matters when inputs change.

The theory is architecture-independent; the experiments use fixed transformer checkpoints. Random sampling broadens the source population, while qualification conditions it on six-model agreement. The target includes context and background; common content may matter to the classifier. Establishing behavioral faithfulness requires specified input interventions (Hooker et al., 2019). Our results establish how spatial evaluators use attribution information, and which comparisons change when that information is retained.

6 Conclusion

Separate localization rankings can hide class distinctions already present in attribution maps. An exact decomposition separates common-content masking from complementary one-sided evidence, and two source populations connect these mechanisms to a supported model advantage that becomes unresolved under contrast scoring. The scale theorem and metric/support controls specify the conditions of that conclusion. **To evaluate class contrast, form the contrast before reducing the maps to scores.**

AI USE STATEMENT

We used ChatGPT 6 Pro, Codex Astra-6, Claude Code Fable 5.1 and Fable 5.1 Max, and Gemini 3.1 Pro. The authors originated the research concept, central mathematical constructions, experimental design, figure concepts, and initial prose, and made the final scientific and editorial decisions.

AI tools assisted literature discovery, drafting and substantive rewriting of sections from author-written material, language editing, manuscript organization, abstract refinement, and simulated peer review. Critique of experiments and interpretations prompted refinements, additional experiments, and qualifications on claims. Mathematical assistance included checking proofs and assumptions, improving notation, and identifying corrections, including those leading to the authors' substantial revision of Proposition A.2.

Codex assisted extensively with implementation and debugging; AI tools also supported correctness tests, data reformatting, algorithmic figure production, and organization and verification of the reproducibility package. Execution was gated on correctness checks, including numerical tests on simulated cases. Experimental source images and annotations came from established datasets; no generative-AI images, labels, or observations were used as experimental evidence. Figures were produced by code with a traceable path to the underlying data.

Verification included AI review, exact-arithmetic tests, and AI-assisted Lean formalization of the finite results specified in Appendix C.4. The authors inspected all final mathematics and checked its correspondence with the code and formalization. The authors assessed the relevance of retrieved literature and checked all citations, supplemented by two AI tools following official source links. The authors reviewed the final text, claims, data provenance, code, figures, and artifacts, and take full responsibility for the submission.

REPRODUCIBILITY STATEMENT

The supplementary PDF contains complete proofs and counterexamples (Appendix A.1), supporting results (Appendix A), methods and numerical conventions (Appendix B), and checkpoint identities and verification coverage (Appendix C). These appendices are submitted separately from the main paper.

Upon acceptance, we will release the full reproducibility package and editable manuscript source with the final version. An anonymous link to the package can also be shared with reviewers and area chairs during the rebuttal period. The release will include all 50,600 principal native map pairs, spatial targets and annotation provenance, the random candidate bank and qualification records, development and segmentation fields, both locally frozen analysis plans, checkpoint and input identities with SHA-256 hashes, capture libraries, pinned upstream code, environment records, and scoped Lean sources with build/axiom receipts. A documented CPU workflow regenerates statistics, paired intervals, tables, and figures without GPU inference or model weights.

Completed independent checks compare three AUC implementations, replay random sampling and all 12,000 qualification decisions, verify the masking, common-minimum, mass, and box analyses, reconstruct segmentation metrics from all 9,266 distinct fields and masks, and compare TA with saved upstream field and forward fixtures. Exact-rational and numerical tests complement the scoped formal checks. These checks validate saved computations and implementation consistency; new inference additionally requires separately obtained checkpoint weights and original photographs. The code, native-data archives, editable source, and Lean project are deferred to the release described above.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems 31*. Curran

Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Anna Arias-Duart, Ferran Parés, Dario Garcia-Gasulla, and Víctor Giménez-Ábalos. Focus! rating XAI methods and finding biases. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8, 2022. doi: 10.1109/FUZZ-IEEE55066.2022.9882821.

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. doi: 10.1371/journal.pone.0130140. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140.

Angie Boggust, Harini Suresh, Hendrik Strobelt, John V. Guttag, and Arvind Satyanarayan. Saliency cards: A framework to characterize and compare saliency methods. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2023. doi: 10.1145/3593013.3593997. URL https://vis.csail.mit.edu/pubs/saliency-cards/.

Moritz Böhle, Mario Fritz, and Bernt Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10029–10038, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Bohle_Convolutional_Dynamic_Alignment_Networks_for_Interpretable_Classifications_CVPR_2021_paper.html.

Lennart Brocki, Jakub Binda, and Neo Christopher Chung. Class-discriminative attention maps for vision transformers. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=MH7xfUWHfP.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 397–406, 2021a. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chefer_Generic_Attention-Model_Explainability_for_Interpreting_Bi-Modal_and_Encoder-Decoder_Transformers_ICCV_2021_paper.html.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791, 2021b. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chefer_Transformer_Interpretability_Beyond_Attention_Visualization_CVPR_2021_paper.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. URL https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Yehonatan Elisha, Seffi Cohen, Oren Barkan, and Noam Koenigstein. Rethinking saliency maps: A cognitive human aligned taxonomy and evaluation framework for explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(5):3750–3758, 2026. doi: 10.1609/aaai.v40i5.37375. URL https://ojs.aaai.org/index.php/AAAI/article/view/37375.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

Benjamin Fresz, Lena Lörcher, and Marco F. Huber. Classification metrics for image explanations: Towards building reliable XAI-evaluations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1–19, 2024. doi: 10.1145/3630106.3658537.

Jindong Gu, Yinchong Yang, and Volker Tresp. Understanding individual decisions of CNNs via contrastive backpropagation. In *Computer Vision – ACCV 2018*, volume 11363 of *Lecture Notes in Computer Science*, pp. 119–134. Springer, 2019. doi: 10.1007/978-3-030-20893-6_8.

Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. ImageNet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110(3):328–348, 2014. doi: 10.1007/s11263-014-0713-9. URL https://link.springer.com/article/10.1007/s11263-014-0713-9.

Anna Hedström, Leander Weber, Dilyara Bareeva, Daniel Krakowczyk, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. URL https://jmlr.org/papers/v24/22-0142.html.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

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 Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pp. 267–280. Springer, Cham, 2019. doi: 10.1007/978-3-030-28954-6_14.

Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In *Computer Vision – ECCV 2018*, volume 11220 of *Lecture Notes in Computer Science*, pp. 798–814. Springer, 2018. doi: 10.1007/978-3-030-01270-0_47.

Tamara Lenhard, Nicole Wagner, and Horst Zisgen. A rank-based Wasserstein distance for comparing the contrastive power of post-hoc XAI techniques. In *IEEE International Conference on Image Processing (ICIP)*, 2026. doi: 10.1109/ICIP61757.2026.11630346. URL https://elib.dlr.de/224294/.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi: 10.1214/aoms/1177730491.

Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards better understanding attribution methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10223–10232, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rao_Towards_Better_Understanding_Attribution_Methods_CVPR_2022_paper.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified BP attributions fail. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9046–9057. PMLR, 2020. URL https://proceedings.mlr.press/v119/sixt20a.html.

Suraj Srinivas and François Fleuret. Rethinking the role of gradient-based attribution methods for model interpretability. In *International Conference on Learning Representations*, 2021. URL https://fleuret.org/papers/srinivas-fleuret-iclr2021.pdf.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=4nPswr1KcP.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In *Computer Vision – ECCV 2022*, volume 13684 of *Lecture Notes in Computer Science*, pp. 516–533. Springer, 2022. doi: 10.1007/978-3-031-20053-3_30.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. URL https://openaccess.thecvf.com/content_CVPRW_2020/html/w1/Wang_Score-CAM_Score-Weighted_Visual_Explanations_for_Convolutional_Neural_Networks_CVPRW_2020_paper.html.

Yipei Wang and Xiaoqian Wang. "why not other classes?": Towards class-contrastive back-propagation explanations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9085–9097, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3b7a66b2d1258e892c89f485b8f896e0-Abstract-Conference.html.

Weiyan Xie, Xiao-Hui Li, Zhi Lin, Leonard K. M. Poon, Caleb Chen Cao, and Nevin L. Zhang. Two-stage holistic and contrastive explanation of image classification. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2335–2345. PMLR, 2023. URL https://proceedings.mlr.press/v216/xie23a.html.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep K. Ravikumar. On the (In)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html.

Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *Computer Vision – ECCV 2016*, volume 9908 of *Lecture Notes in Computer Science*, pp. 543–559. Springer, 2016. doi: 10.1007/978-3-319-46493-0_33.