Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Contextual Reversals in Visual Explanations: Separating Intrinsic Disagreement from Ranking Error

Mazen Hesham Elnahal

m.hesham2400785@badyauni.edu.eg

Badya University

Research Article

Keywords: visual explanations, contextual reversals, removal-based evaluation, feature ranking, reproducibility

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11341750/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Contextual Reversals in Visual Explanations: Separating Intrinsic Disagreement from Ranking Error

Mazen Hesham Elnahal

Department of AI and Data Science - School of Computing and Data Science - Badya University – 6th Of October City - Egypt

m.hesham2400785@badyauni.edu.eg / mazenelnahal51@gmail.com

Corresponding author: Mazen Hesham Elnahal

Abstract

A fixed regional ranking is often used to predict which image region will produce a larger model response when removed. Context can change this order, so disagreement may reflect both the prediction task and the selected ranking. This study audits these components by exhaustively evaluating a finite removal family on 100 fresh Imagenette images shared by MobileNetV3-Large and ResNet34. The primary cohort contains 68 originals correctly classified by both models. For each image, model and replacement operator, 1,820 size-four removal masks determine 43,680 pair/context comparisons. Across mean-fill and blur probes, the best fixed order incurs 17.26–23.95% mean disagreement, while additional ordering-constraint loss is only 0.0039–0.0110 percentage points. On evaluation masks disjoint from fitting masks, a 16-query singleton ranking incurs 2.44–5.21 additional points and outperforms signed Integrated Gradients and positive Grad-CAM in all four settings. Ninety-six-query fitted fixed rankings improve consistently; a context-conditioned predictor exceeds the fixed-order ceiling by 1.20–1.63 points. All primary blur-reference Integrated Gradients runs pass the numerical checks, whereas nine primary MobileNet mean-reference runs fail completeness and remain in the analysis. Disagreement persists under SLIC-derived regions on a matched 13-image subset. Independent integer programming certifies 960 optimization results. The findings support reporting a task-specific fixed-ranking ceiling alongside explainer agreement in finite removal-effect benchmarks.

Keywords: visual explanations; contextual reversals; removal-based evaluation; feature ranking; reproducibility

1 Introduction

A visual explanation often summarizes a classifier with one spatial map. A reader may interpret a region with a higher score as having a larger effect on the prediction when removed. This interpretation adds a prediction task to the explanation: a single regional order is expected to forecast relative model responses across different combinations of visible and removed content. The present study asks how much of the disagreement in that task is unavoidable for every fixed order, and how much is additional error of the chosen ranking.

Integrated Gradients (IG) attributes a prediction difference relative to a reference image by integrating gradients along a path [1]. Gradient-weighted Class Activation Mapping (Grad-CAM) summarizes class-related spatial evidence using convolutional activations and gradients [2]. These constructions have specific targets. Neither construction, by itself, supplies a guarantee

that one regional order will reproduce every comparison of finite removals. Removal-based evaluation must therefore specify the intervention and the response being predicted [3].

The distinction matters when context changes a pair's order. If region i has the larger removal effect in some contexts and region j has the larger effect in others, a fixed order must disagree with at least one direction. Observing a reversal establishes that this incompatibility exists, but its frequency across pairs does not measure the number of incorrect comparison predictions. A pair can reverse once among hundreds of contexts and remain easy to predict by its dominant direction. A second issue arises when dominant directions across several pairs cannot all belong to the same total order.

This paper measures both effects through an exhaustive finite experiment. The response table for 16 regions contains every removal of exactly four regions. It determines 43,680 pair/context comparisons per image, model and replacement operator. The best pairwise majority choices and the best realizable total order separate contextual disagreement from ordering constraints. Comparing a real ranking with the latter then measures its additional disagreement on the same outcomes.

The confirmation uses 100 fresh Imagenette origins shared by MobileNetV3-Large and ResNet34, with 68 originals correctly classified by both models. Across four primary settings, the optimal fixed order incurs 17.26–23.95% mean disagreement. Additional ordering constraints cost only 0.0039–0.0110 percentage points. Thus the main empirical obstacle is variation across contexts, while dominant pair directions remain almost compatible with a global order. This revises the developmental expectation that majority-cycle cost would provide the principal explanation.

The study also separates evaluation diagnostics from prediction. The exact optimum is fitted to evaluated outcomes and serves as a ceiling for fixed orders. Singleton and ridge predictors receive only a frozen fitting set; their evaluation comparisons use disjoint response masks. A 16-query singleton ranking outperforms the evaluated IG and Grad-CAM aggregations. Both 96-query fitted fixed rankings improve over singleton in every primary setting, while a predictor that receives the shared context exceeds the fixed-order ceiling.

The contribution is a verified empirical audit and a practical interpretation of removal-order benchmarks. The accounting identity, optimization formulations and response predictors are established tools. The evidence consists of the relative sizes of the error components on a fresh shared cohort, held-out response comparisons, alternative-region and threshold checks, numerical attribution checks, and independent certification of the finite optima. Reporting these quantities together lets an evaluator distinguish a restrictive output representation from avoidable ranking error.

2 Background and related work

2.1 Explanation objectives and evaluation targets

Explanation quality depends on what the explanation is intended to convey. IG's reference-relative completeness concerns the sum of attributions and an endpoint prediction difference [1]. Grad-CAM constructs a spatial localization map [2]. The present evaluation converts both to regional scores and asks whether their order predicts a declared finite-removal contrast. This is an operational agreement criterion; its interpretation follows from that task rather than from visual plausibility alone.

Adebayo et al. show why visually persuasive saliency maps require checks of their dependence on the model and data [4]. Those checks motivate explicit quantitative evaluation, but answer a different question from the contextual ordering audit. Yeh et al. formalize infidelity and sensitivity and demonstrate how the perturbation distribution influences the explanation objective [5]. Here the target is the sign of a pairwise removal-effect difference, rather than squared reconstruction of a response change or local sensitivity of a map.

2.2 Context interactions and uncertainty in rankings

Context dependence is established in work on feature interactions. Archipelago detects and attributes interactions [6], and Faith-Shap develops a faithful Shapley interaction index with polynomial representations [7]. The quadratic response baseline in this study uses standard pairwise features; it is not introduced as a new interaction attribution method. The audit instead counts the contextual directions produced by a fixed response table and quantifies the loss of representing them by one order.

Partial Order in Chaos examines feature-attribution relations across a Rashomon set of plausible models [8]. This study holds each evaluated model fixed while varying removed context. Confident Feature Ranking addresses uncertainty in estimated global feature ranks [9]. That uncertainty is distinct from systematic disagreement between a global rank and individual contexts: even a precisely estimated majority order can make errors on the minority direction.

2.3 Removal mechanisms and attainable benchmark limits

Explaining by Removing organizes explanation methods around choices of removal mechanism, model behavior and summary [3]. This framework makes clear that replacing pixels with a mean or a blur defines a particular probe. These operators are inexpensive and reproducible, but their responses need not equal effects of removing a semantic object from a natural scene.

Remove And Retrain (ROAR) evaluates importance using retraining after removal [10]. F-Fidelity addresses faithfulness evaluation with fine-tuning to mitigate distributional difficulties [11]. The present study retains pretrained weights so that contextual directions belong to a single response function. It does not estimate whether a retrained model would recover performance or whether generated interventions would preserve a natural image distribution.

Normalized Area Over the Perturbation Curve (Normalized AOPC) compares perturbation scores with model-specific attainable limits [12]. This is the closest precedent for interpreting a score relative to an oracle. The distinction here is the loss: pair/context ordering agreement permits separate calculation of contextual minority loss, additional ordering-constraint loss and ranking-specific excess. The study supplies exact finite ceilings and tests low-budget predictors on responses excluded from fitting.

SAFARI evaluates explanation discrepancies through probabilistic and worst-case analyses around perturbed inputs [13]. Its robustness question concerns changes in explanations. The present target concerns changes in relative removal effects across shared contexts. An explanation can be spatially stable while failing to forecast these contextual directions, or vary without creating the specific ordering loss evaluated here.

These relationships support a narrow empirical contribution. The paper does not claim to discover context dependence, introduce the arithmetic decomposition, or establish the first use of an attainable explanation limit. It evaluates how the components behave together under a

controlled visual removal protocol and provides evidence for reporting them alongside actual predictor performance.

3 Finite contextual ordering task

3.1 Target score and removal contrast

Let x be an image, t its ground-truth ImageNet class, and z the logits of a fixed classifier. Both architectures use the same target t for an origin. The primary response is the target margin against the aggregate of all other ImageNet logits, not merely the other nine Imagenette classes:

$$m_t(x) = z_t(x) - log\left(\Sigma_{\ell \neq t}\, exp\left(z_\ell(x)\right)\right) \quad (1)$$

For a removed region set S and a replacement operator R, define $D(S)=m_t(x)-m_t(R(x,S))$. Larger D means a larger decrease in the target margin. Let i and j be distinct regions outside a shared three-region context C. The relative contrast is:

$$d(i,j;C) = D(C \cup \{i\}) - D(C \cup \{j\}) \quad (2)$$

A positive contrast favors i; a negative contrast favors j. Both images have four removed regions, so the comparison holds removed-region cardinality constant. The unmodified score cancels. If one expresses each effect relative to the context-only removal D(C), that common term also cancels. Consequently, all pair contrasts follow from the size-four response table without evaluating every context-only image.

This cancellation determines relative effects. It does not recover the two individual marginal magnitudes from that table alone. The task also allows removals to change the predicted class; eligibility concerns the original image, and is not reimposed after each intervention. The target remains fixed throughout.

Figure 1 shows one recorded reversal. The shared gray regions change between the two contexts, while the alternatives remain regions 2 and 15. The displayed margins imply contrast +7.387 in the first context and −5.384 in the second. This example illustrates the mechanism; aggregate estimates use the entire finite family.

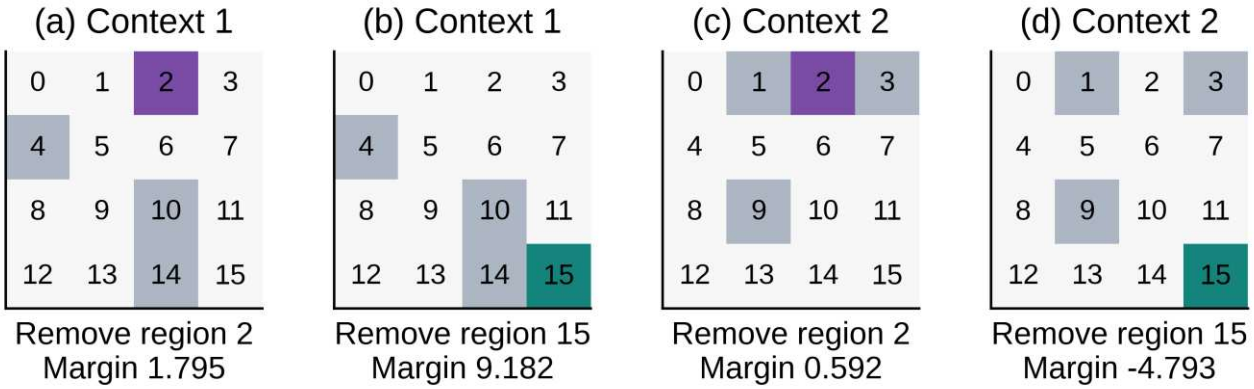


Fig. 1 Measured contextual reversal under ResNet34 mean fill for image 15668391b9f7f80f in the parachute class. Gray cells form the removed shared context; purple and teal identify the alternatives. Panels a and b use context 1 and panels c and d context 2. The relative removal contrast switches from +7.387 to −5.384. The image nearest the setting median gap and its

strongest balanced reversal pair were selected post hoc for illustration. These diagrams show masks rather than semantic object boundaries

3.2 Comparison population ties and coverage

Sixteen regions yield 120 unordered pairs. Each pair has every three-element context among the other 14 regions, giving 364 contexts and 43,680 comparison rows. The 1,820 size-four responses are reused across these rows. Comparisons within an image are therefore dependent; exhaustive enumeration does not turn them into independent population samples.

At threshold ε, retain only comparisons with |d|>ε. Exact truth ties are excluded at ε=0. A predicted strict direction receives one for agreement and zero for disagreement; a predicted tie receives one half. Coverage is the fraction of allowed comparisons retained by thresholding. Its denominator is the complete family or the held-out family, as applicable, before threshold exclusions.

The primary analysis uses ε=0 and target margin. Frozen sensitivities use margin thresholds 0.01, 0.1 and 0.5, and target-logit alternatives. Thresholding changes the evaluated comparison population. It does not simply improve the prediction method on an unchanged test set, so threshold-conditioned agreement must be read with coverage.

3.3 Attainable agreements and error components

Let w(i,j) count retained contexts favoring i over j, and let N be the total retained comparison count. Choosing the majority direction separately for every pair achieves agreement B. Let K be the maximum agreement of any strict total order π over the regions. A chosen fixed score ranking achieves agreement A on those same comparisons:

$$B = \frac{\Sigma_{i<j} \max\{w(i,j), w(j,i)\}}{N}, \quad K = \max_{\pi} A(\pi) \quad (3)$$

$$1 - A = (1 - B) + (B - K) + (K - A) \quad (4)$$

The first component, 1−B, is contextual minority-direction loss: the error remaining even when each pair can select its own dominant direction. The second, B−K, is additional loss required to make the pair choices compatible with one order. The third, K−A, is ranking-specific excess. Each term is nonnegative for a fixed ranking because A≤K≤B.

Tie credit preserves the bound. A tied score group can be ordered uniformly at random internally; each tied pair then agrees on average half the time. The score ranking's agreement is an average over compatible strict orders and cannot exceed their maximum. In contrast, a predictor that changes direction with C is outside the fixed-order class and can exceed K.

For intuition, if a pair favors i in eight contexts and j in two, its majority predictor still misses two of ten comparisons. The pair counts as reversed, but its disagreement is 20%, not 100%. For several regions, majority directions can also form a cycle, forcing an order to sacrifice a weighted edge. This illustrative arithmetic explains why reversal prevalence and ordering cost must be measured separately.

Reversal prevalence is the proportion of active unordered pairs for which both directions occur at least once. It counts pairs, while agreement counts comparisons. A small B−K also does not imply that there are no majority cycles; it states that the weighted penalty for enforcing an order is small on the evaluated family.

3.4 Exact optimization and independent verification

A subset dynamic program maximizes the weighted order objective over 2^{16} states. If F(S) is the best reward for ordering S and F(∅)=0, placing j last gives:

$$F(S) = max_{j \in S} \{F(S \setminus \{j\}) + \Sigma_{i \in S \setminus \{j\}} w(i,j)\} \quad (5)$$

The reconstruction is checked against the optimal reward. The recurrence takes $O(n2^n)$ state transitions and $O(2^n)$ state storage, making the small finite region count central to feasibility. The exact census is an audit procedure; its cost grows rapidly if the region representation is expanded.

An independent mixed-integer linear program (MILP) uses binary order variables and transitivity constraints. It certifies 960 planned complete-family and held-out zero-threshold margin optima at zero solver gap. These cover all 100 grid origins and all 20 geometry-subset origins, two models and two operators. The higher thresholds use the dynamic program and are not separately claimed as MILP-certified.

K uses evaluated outcomes. It is an evaluation-fitted diagnostic, not a predictor trained on the fitting masks. Complete-family K and held-out K are calculated separately. In particular, a held-out ranking's excess is measured against the optimum for its exact held-out comparison population.

4 Experimental design

4.1 Fresh shared cohort and model eligibility

The frozen protocol selects 100 unique Imagenette-160 validation origins [14], ten from each of ten classes, without inspecting model responses. All 580 known project-development origins are excluded by archive path and image-byte hash. The data seed is 2026101131. Freshness refers to prior project use; it does not establish independence from the ImageNet pretraining corpus.

MobileNetV3-Large [15] and ResNet34 [16] use torchvision IMAGENET1K_V1 weights. They receive identical origins and ground-truth targets. The primary cohort is the intersection of correctly classified originals, containing 68 images. All 100 selected images remain in a secondary analysis, and misclassified images are not replaced. Table 1 shows how original-image eligibility varies by class.

Class	Selected	MobileNet	ResNet34	Joint
English springer	10	7	7	6
French horn	10	8	9	8
cassette player	10	7	7	6
chain saw	10	6	7	6
church	10	2	5	2
garbage truck	10	8	7	6
gas pump	10	9	9	9
golf ball	10	10	9	9
parachute	10	9	10	9
tench	10	7	7	7

Table 1 Original-image correctness by class. Joint correctness defines the 68-image primary cohort. Selection is balanced before eligibility; the primary cohort contains two church images and nine images in each of three classes

Sharing origins and targets removes the developmental confounding in which each architecture had a different image cohort. Differences between these two trained systems can now be computed within image. They still combine architecture and pretrained-weight differences; this is not an experiment that isolates architecture while controlling training history.

RGB inputs are resized bilinearly with antialiasing to a short side of 256 and center-cropped to 224×224. Standard ImageNet normalization is retained. The original target margin and predicted label are saved before interventions, so correctness conditioning is explicit and traceable.

4.2 Regions and replacement operators

The primary geometry is a 4×4 grid with 16 equal 56×56 cells. Mean fill replaces removed pixels with the ImageNet RGB mean. Blur replaces removed pixels with corresponding pixels of the original image after Gaussian filtering with kernel size 31 and standard deviation 7. Blur preserves low-frequency spatial information and should be interpreted as information degradation rather than complete content erasure.

A response-independent subset of 20 selected origins also receives regions derived from Simple Linear Iterative Clustering (SLIC) [17]. Adjacent partitions merge deterministically by Lab-color Ward cost until exactly 16 nonempty regions remain. Thirteen of these origins meet joint correctness. Matched geometry comparisons use those same 13 origins for both grid and SLIC16.

Four grid regions always cover one quarter of the input. Four SLIC16 regions can cover different pixel areas. A cardinality control therefore addresses region count, while an area-ranking control addresses unequal covered area. SLIC-derived partitions provide an alternative operational geometry; they are not ground-truth object segments or semantic inpainting.

4.3 Fitting masks and held-out response evaluation

Each image has a deterministic fitting pool of 16 singleton removals and 80 randomly selected size-four removals. Linear-32 receives the singletons and the first 16 four-region masks. Linear-

96 and both Quadratic-96 outputs receive all 96 fitting responses. The remaining 1,740 four-region masks are evaluation-only. The split seed uses 2026101132 and the image identifier; the assignment is shared across models, operators and geometries.

A pair/context comparison enters held-out evaluation only if both response masks underlying its contrast are evaluation masks. Holding out comparison rows alone would be insufficient because a fitted response might otherwise reappear through a different comparison. Excluding both constituent responses prevents that route of outcome leakage into fitting.

Predictors are fitted separately for each image, model, operator and geometry. The protocol holds out intervention responses within images; it does not train a single predictor and test it on new image identities. The fresh-image confirmation and the within-image response split address different questions, and are reported separately.

4.4 Baselines and output capabilities

Singleton ranks the 16 directly observed drops $D(\{i\})$. Linear ridge fits 16 removed-region indicators. Quadratic ridge adds all 120 pairwise indicator products. There is no intercept. The frozen penalties are 0.32 for main coefficients and 3.2 for pair interactions; no evaluated response is used for fitting or hyperparameter selection.

For a quadratic response $\hat{D}(S)=\Sigma_{i\in S}\beta_i+\Sigma_{i<j,i,j\in S}\theta_{ij}$, the fixed regional score is $\beta_i+(3/14)\Sigma_{j\neq i}\theta_{ij}$. Its score difference equals the mean predicted contrast over uniform size-three contexts excluding the pair. The common θ_{ij} term cancels. The contextual output instead predicts $\hat{D}(C\cup\{i\})-\hat{D}(C\cup\{j\})$ for each observed context C.

Thus the two quadratic outputs use the same fitting responses and coefficients but different prediction representations. The fixed output supplies one order for the image; the contextual output supplies a direction for each pair and context. Their comparison tests whether conditioning on context helps with this task, not whether one saliency map is better than another.

Cardinality predicts ties for every fixed-size contrast. Equal grid area also predicts ties, giving 50% agreement through half credit. The study does not infer a random-ranking empirical result from those controls or claim an adaptive mask-selection advantage from the randomly selected fitting pool.

4.5 Attribution construction and numerical checks

IG integrates target-margin gradients on the straight path from a mean-fill or original-blur reference to the original input. Channel attributions are summed and then aggregated by region, preserving signs. Midpoint sums at 512 steps supply the evaluated scores; 256 steps provide a refinement check. Grad-CAM uses margin gradients at the final convolutional block, applies ReLU and bilinear upsampling, and scores regions by the mean of the resulting positive map.

These aggregation rules are frozen. Signed IG sums and positive CAM means are different summaries, especially for irregular regions. The findings compare their use as predictors of the declared removal order rather than establishing that either summary is appropriate for every explanation purpose.

At each IG resolution, the absolute completeness residual must not exceed $0.05+0.05|m_t(x)-m_t(\text{reference})|$. Regional refinement requires

$\|IG_{256} - IG_{512}\|_2 / (\|IG_{512}\|_2 + 10^{-12}) \leq 0.1$. The joint flag requires both completeness checks and refinement. Failed runs remain in the primary results, with matching passing-only results supplied as a secondary sensitivity.

The saved pipeline enables deterministic Torch options and disables TF32. Regional sums use float64 after float32 model gradients. All-zero Grad-CAM maps remain tied scores. Neither numerical failures nor degenerate maps cause post hoc image replacement.

4.6 Estimation uncertainty and frozen outcomes

Image means receive equal weight. The primary outcomes are complete-family 1−K and held-out singleton K−A at zero-threshold target margin. Intervals are 95% percentile intervals from 2,000 class-stratified bootstrap draws with seed 2026101133. Resampling preserves each observed class count, and differences use paired image identities.

The bootstrap unit is the image, not the comparison. The intervals condition on the dataset, selected and eligible cohort, pretrained systems, intervention family, region construction and fitting-mask assignment. They are descriptive intervals; there are no comparison-level p-values, multiplicity-adjusted rejection decisions or effect-dependent stopping rules in this confirmation. The full protocol and seeds are supplied in Online Resource 1 (Supplementary_Information.docx).

5 Results

5.1 Contextual disagreement dominates ordering cost

Table 2 reports the two frozen endpoints and the additional ordering-constraint loss. All four settings use the same 68 jointly correct originals. Even the exact best fixed order disagrees on 17.26–23.95% of comparisons on average. This gives a direct limit for interpreting the agreement of an explanation converted to a fixed regional ranking.

The mean ordering-constraint cost is only 0.0039–0.0110 percentage points. Its largest primary image value is 0.1007 points. Figure 2 shows the image-level distributions with separate vertical scales. These costs are nonzero, but are small relative to the total fixed-order disagreement. The evidence therefore supports contextual minority directions as the dominant component in this finite family.

Setting	Complete gap %	Ordering cost pp	Held-out excess pp
MobileNet / Mean	17.26 [16.45, 18.02]	0.0039 [0.0017, 0.0063]	2.44 [2.08, 2.80]
MobileNet / Blur	18.31 [17.46, 19.24]	0.0067 [0.0036, 0.0102]	5.21 [4.40, 6.05]
ResNet34 / Mean	23.95 [23.13, 24.78]	0.0079 [0.0047, 0.0114]	4.62 [3.91, 5.39]
ResNet34 / Blur	21.13 [20.23, 21.99]	0.0110 [0.0069, 0.0156]	4.10 [3.47, 4.77]

Table 2 Frozen primary endpoints and additional ordering cost. Brackets give 95% class-stratified image-bootstrap intervals; pp denotes percentage points. Complete gap uses the full family, while singleton excess uses the separately optimized held-out family

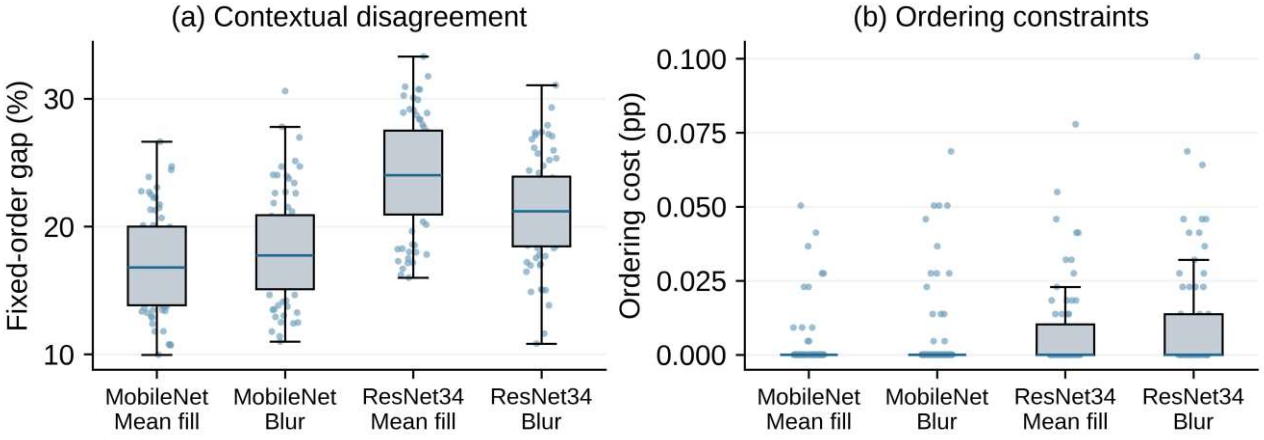


Fig. 2 Complete-family disagreement and ordering-constraint loss on the same 68 originals. Panel a shows the fixed-order gap and panel b the additional ordering cost. Points are images; boxes show medians, quartiles and 1.5-interquartile-range whiskers. Different vertical scales keep the much smaller ordering costs visible

Both directions occur somewhere in 85.20–96.86% of active region pairs. Such widespread reversals coexist with almost coherent aggregate preferences. A fixed majority order can be accurate on a dominant direction while missing many contexts that favor the alternative. Reversal prevalence is consequently not an 85–97% comparison failure rate.

The confirmation also supplies a negative result for the earlier expectation of substantial majority-cycle cost. It would be misleading to present the existence of reversals or occasional cycles as evidence that most disagreement comes from global ordering constraints. The measured weighted decomposition distinguishes these mechanisms.

5.2 Fixed rankings on held-out responses

Table 3 reports agreement on identical held-out comparisons. The diagnostic ceiling is 76.03–82.73% across settings. Singleton obtains 71.42–80.29% and incurs 2.44–5.21 additional percentage points of disagreement. Both IG references and positive Grad-CAM score lower than singleton in every primary setting.

All primary blur-reference IG runs pass the frozen numerical gates, so that comparison does not rely on failed integration runs. Mean-reference MobileNet includes nine completeness failures. These failures remain visible and are addressed by the passing-only sensitivity rather than being removed to improve the displayed result.

Method	M mean	M blur	R mean	R blur
Fixed-order ceiling	82.73	81.72	76.03	78.89
Singleton 16	80.29	76.51	71.42	74.79
Linear 32	78.73	76.87	70.90	73.54
Linear 96	80.83	79.90	73.85	76.48
Quadratic fixed 96	81.28	80.27	74.35	77.05
IG mean reference	59.36	60.01	59.22	59.29
IG blur reference	60.98	66.33	61.96	64.90
Grad-CAM	64.32	67.96	67.02	66.02
Quadratic context 96	84.03	82.95	77.66	80.09

Table 3 Mean held-out agreement in percent. M denotes MobileNetV3-Large and R ResNet34. The ceiling is an evaluation-fitted diagnostic. All methods use the same comparisons. Context prediction has a different output capability; cardinality and equal-grid-area controls score 50%

Figure 3 partitions each fixed method’s disagreement into the common held-out fixed-order gap and method-specific excess. The display keeps the unavoidable component visible rather than treating all disagreement as error attributable to the explanation. The larger excesses for IG and Grad-CAM remain specific to the frozen target, operators and regional aggregation rules.

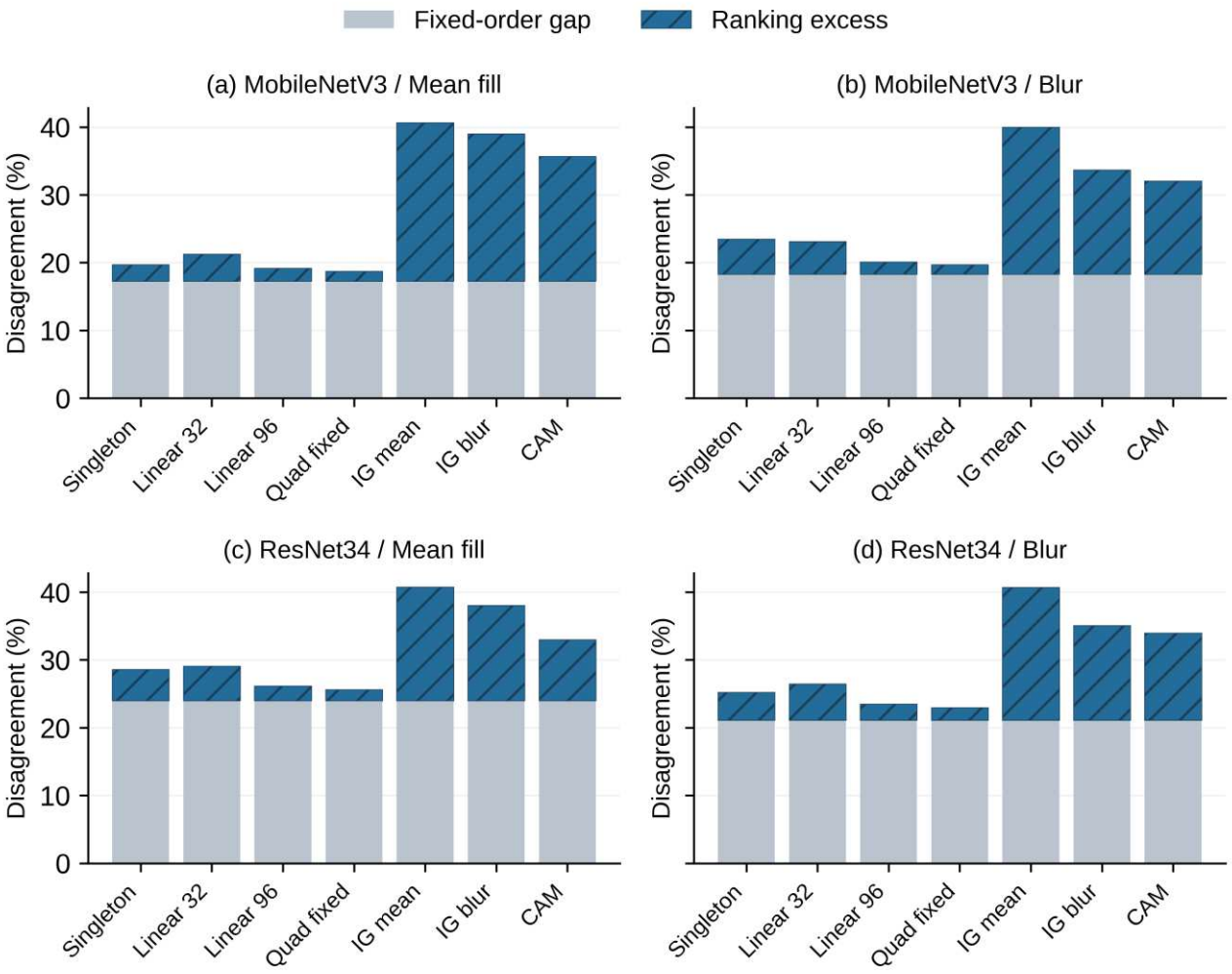


Fig. 3 Held-out fixed-ranking disagreement on 68 images. Panels a and b show MobileNet mean fill and blur; panels c and d show the corresponding ResNet34 settings. Solid gray denotes the common fixed-order gap and hatched blue ranking-specific excess. The bars sum to total disagreement. Mean-reference MobileNet IG includes its nine completeness failures

5.3 Fitting budget and response prediction

Linear-32 does not consistently improve over singleton. Its paired differences range from −1.56 to +0.36 percentage points; only MobileNet blur has a positive point estimate, and its interval crosses zero. The developmental preference for a universal 32-query winner therefore does not carry into this fresh confirmation.

Linear-96 improves over singleton by 0.54–3.39 points, and Quadratic fixed-96 by 0.99–3.76 points. Each method has a positive paired interval in all four primary settings (Fig. 4). The two 96-query outputs access 80 more fitting responses than singleton. This additional access is part of the interpretation: the results demonstrate a useful low-budget response-fitting comparison, not a cost-free improvement.

Within the 96-query budget, the quadratic fixed output scores higher than the linear output in all four displayed settings. The primary paired intervals were constructed against singleton, so that observation alone is not presented as a separately tested quadratic-versus-linear advantage. No budget or ridge penalty is retuned after seeing evaluated responses.

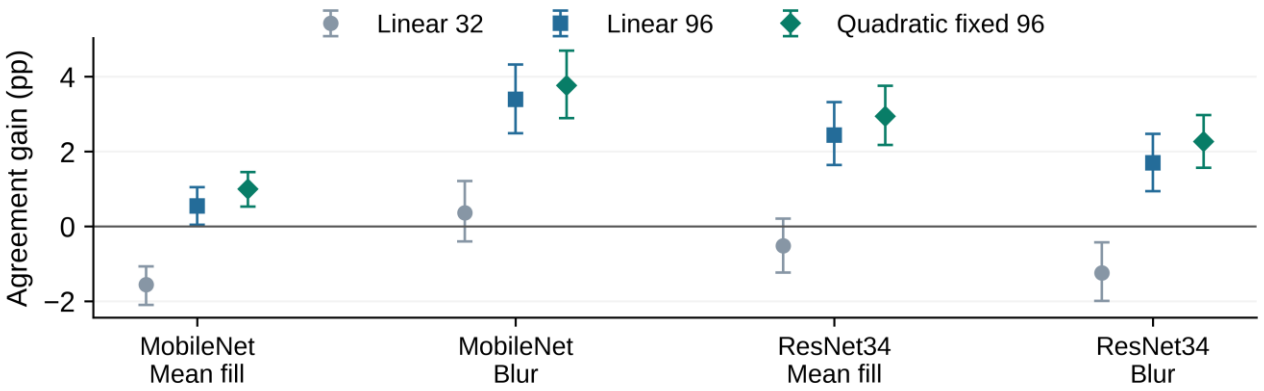


Fig. 4 Paired held-out agreement gains over singleton on 68 originals. Error bars are 95% class-stratified image-bootstrap intervals. Gray circles show Linear-32, blue squares Linear-96 and green diamonds Quadratic fixed-96. Both 96-query fixed rankings have positive intervals in all four settings; the 32-query ranking does not improve consistently

5.4 Context prediction exceeds the fixed-order limit

The contextual quadratic output obtains 77.66–84.03% agreement. It exceeds the held-out fixed-order ceiling by 1.20–1.63 percentage points, with positive paired intervals in every setting (Table 4). Between 54 and 56 of the 68 images exceed their own fixed-order ceilings. Figure 5 shows agreement for the two output capabilities.

This result is compatible with A≤K for fixed rankings. The contextual predictor uses C and may select different directions for the same pair in different contexts. It is therefore evaluated as a predictor of contextual comparisons rather than as one regional order. The benefit is modest relative to the full fixed-order gap, so the evidence supports some recovery of contextual variation without claiming that the fitted quadratic representation eliminates it.

Setting	Context minus ceiling pp	Images above K
MobileNet / Mean	1.30 [0.93, 1.66]	54/68
MobileNet / Blur	1.23 [0.87, 1.59]	56/68
ResNet34 / Mean	1.63 [1.30, 1.97]	54/68
ResNet34 / Blur	1.20 [0.84, 1.55]	55/68

Table 4 Paired context-predictor agreement minus the held-out fixed-order ceiling. Both outputs use the same frozen 96-response quadratic fit. Brackets give image-bootstrap intervals; the final column counts individual images with agreement above their own diagnostic ceiling

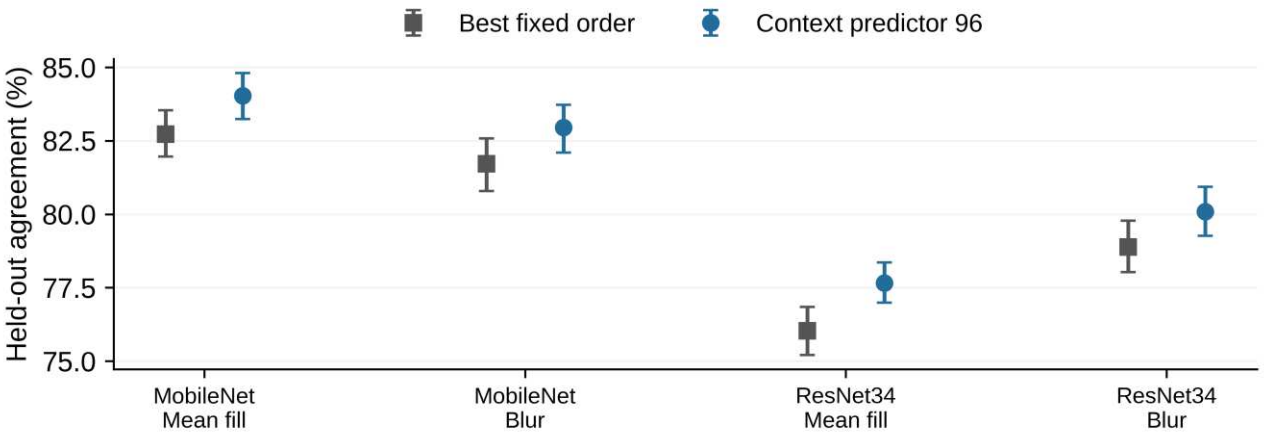


Fig. 5 Mean held-out agreement of the best fixed order and the 96-query contextual predictor. Squares mark the evaluation-fitted fixed-order diagnostic and circles the predictor fitted only to training responses. Error bars show each estimate's image-bootstrap interval. Table 4 gives the paired difference intervals

5.5 Thresholds and score sensitivity

Figure 6 evaluates the frozen target-margin thresholds with retained coverage. At ε=0.5, the complete-family gap remains 7.09–18.05%, with 59.14–75.89% coverage. Higher thresholds remove small contrasts and usually lower the displayed gap, but the amount retained differs by model and operator.

The target-logit sensitivity also retains substantial fixed-order disagreement. At zero logit threshold, primary gaps are 16.29%, 16.54%, 23.36% and 21.02% for MobileNet mean fill, MobileNet blur, ResNet mean fill and ResNet blur respectively. The exact magnitude depends on the score function; it is not appropriate to transfer a margin-based percentage to a different target score without recomputation.

The nonzero thresholds additionally expose near-tie sensitivity. They do not establish that large contrasts are semantic interventions or clinically important. Online Resource 1 gives the frozen score/threshold tables, and Online Resource 2 contains the full interval and coverage records.

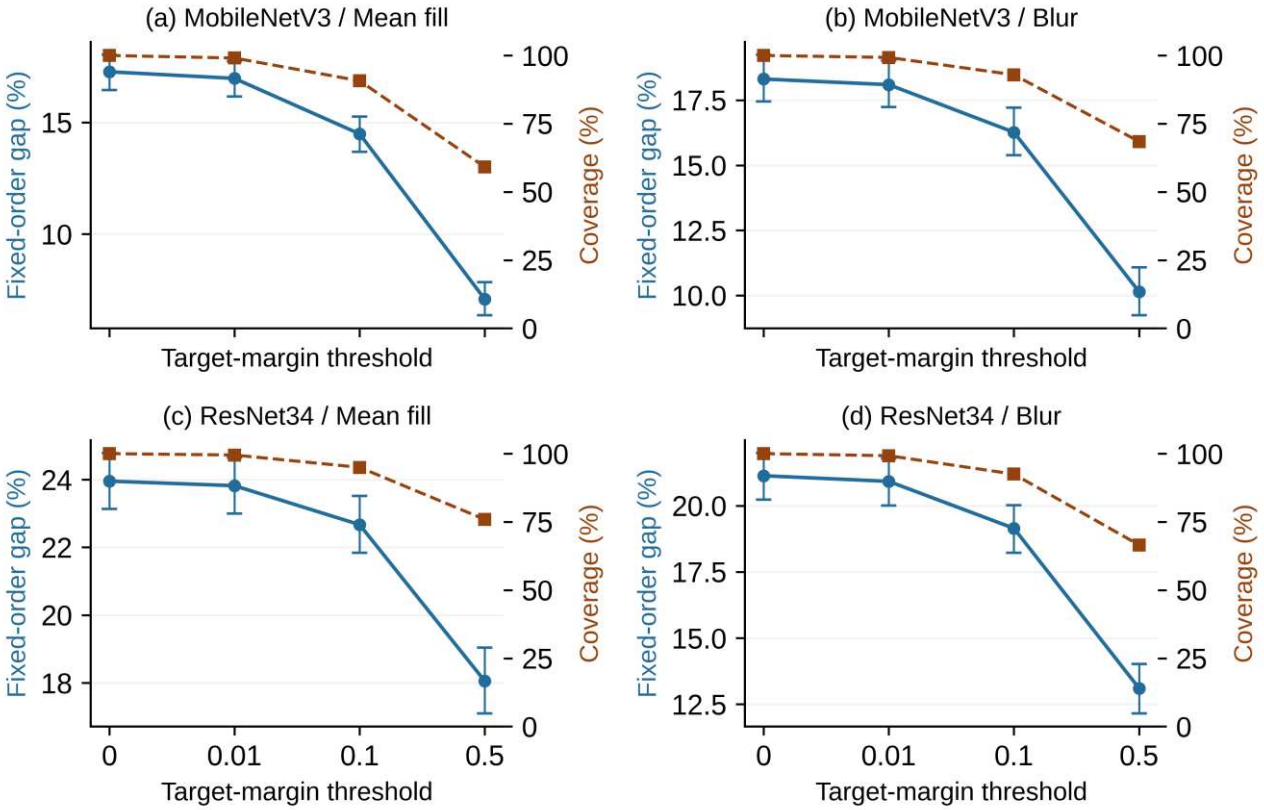


Fig. 6 Complete-family fixed-order gaps and coverage at target-margin thresholds 0, 0.01, 0.1 and 0.5. Panels a and b show MobileNet mean fill and blur; panels c and d show ResNet34. Blue circles and error bars give gap means and image-bootstrap intervals; orange dashed curves with squares give mean retained coverage. Thresholds have categorical spacing and change the evaluated comparison population

5.6 Alternative regions on a matched subset

SLIC16 gaps range from 17.97% to 22.19% on the matched 13-image primary geometry subset. Table 5 compares grid and SLIC16 on exactly those origins. The paired change is positive for MobileNet mean fill and negative for the other displayed settings, with intervals near or crossing zero for some of the decreases.

Figure 7 shows both the complete-family gap and held-out singleton excess. This check supports persistence of substantial disagreement beyond the grid under one deterministic image-derived partition. It does not support geometry invariance: changing region boundaries and areas changes the intervention family, and the matched subset is small.

Unequal SLIC areas remain an operational consideration. The supplied paired geometry records include area-ranking controls and other method outcomes. Matching four removed regions controls cardinality, while covered pixel area can still vary. These distinct controls prevent interpreting every geometry difference as a pure semantic-context effect.

Setting	Grid gap %	SLIC16 gap %	SLIC minus grid pp
MobileNet / Blur	19.17	17.97	-1.20 [-2.39, -0.02]
MobileNet / Mean	16.62	20.48	3.86 [3.02, 4.70]
ResNet34 / Blur	21.11	19.66	-1.45 [-3.67, 0.77]
ResNet34 / Mean	23.86	22.19	-1.68 [-3.37, 0.01]

Table 5 Matched complete-family geometry comparison on the same 13 jointly correct originals. Brackets give paired image-bootstrap intervals. The grid estimates condition on this subset and therefore differ from the 68-image primary estimates

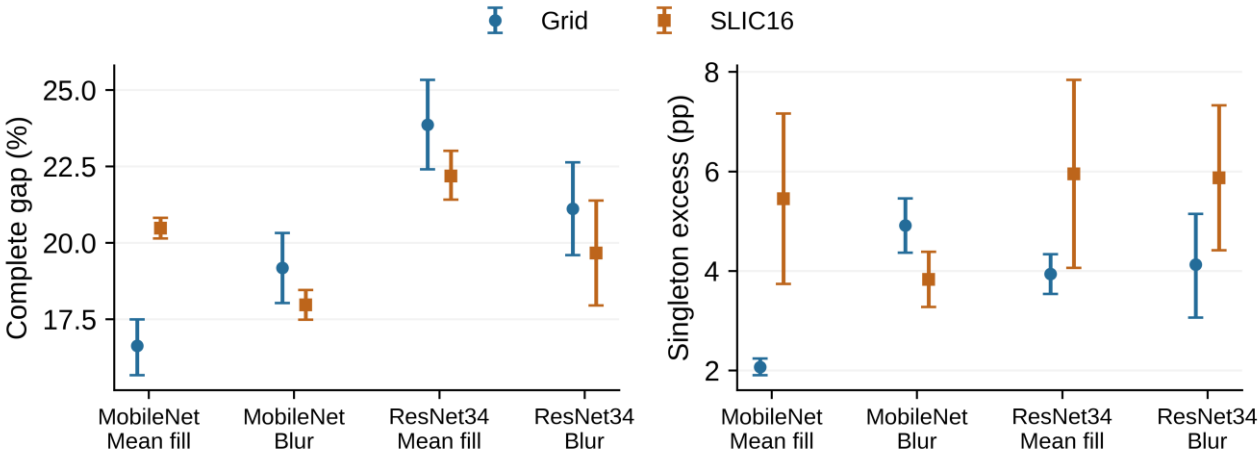


Fig. 7 Matched geometry sensitivities on 13 originals. The left panel gives complete-family gaps and the right held-out singleton excess. Blue circles identify grid and orange squares SLIC16. Error bars are image-bootstrap intervals. Covered area varies among four-region SLIC16 removals

5.7 Numerical quality of the attribution runs

Across all 400 model/image/reference IG runs, 27 fail completeness and none fails regional refinement. Failures comprise 25 MobileNet mean-reference and two ResNet mean-reference runs. All 200 blur-reference runs pass. Among the 68 primary origins, only nine MobileNet mean-reference runs fail; the other three model/reference combinations pass 68/68 (Fig. 8).

The primary analysis retains these failures because numerical outcomes are themselves observed results. A matching passing-only sensitivity compares mean-reference MobileNet IG and singleton on the same 59 passing originals. IG agreement is 58.96% for mean fill and 59.78% for blur, versus singleton at 80.27% and 76.67%. Paired IG-minus-singleton intervals are −21.31 [−23.39, −19.25] and −16.89 [−18.89, −14.97] percentage points respectively.

That subset is selected by an observed numerical outcome and remains secondary. It shows that the singleton advantage persists among the passing runs, but cannot replace the primary cohort or establish the behavior of an ideal exact integral. Online Resource 1 shows normalized integration residuals at both resolutions. One ResNet Grad-CAM map is all zero and receives tie credit; it is retained without replacement.

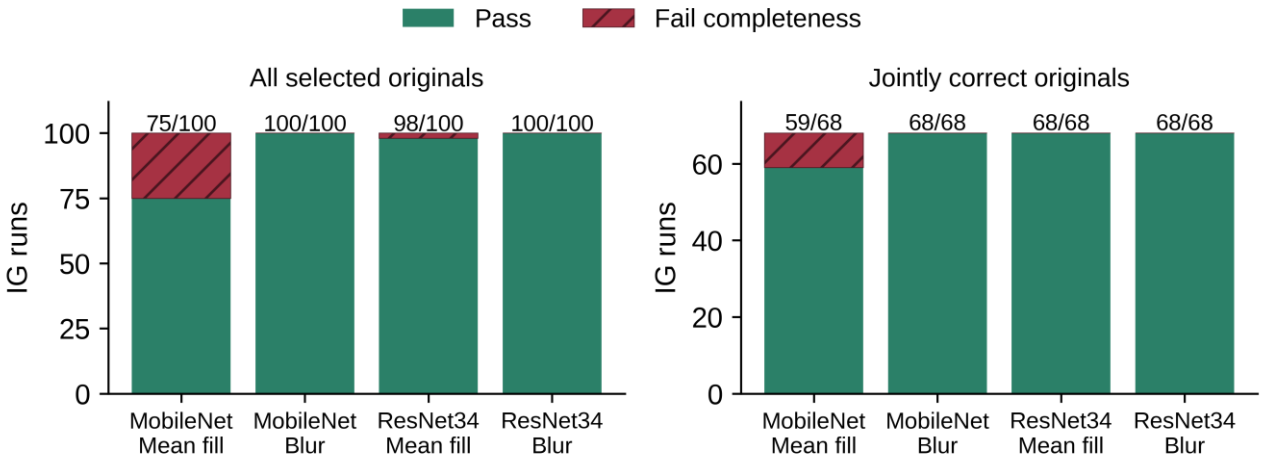


Fig. 8 Numerical IG outcomes by model and reference. The left panel includes all 100 selected origins per combination and the right the 68 jointly correct originals. Solid green marks passing runs and hatched red completeness failures; labels give passing counts over totals. No run fails the regional refinement gate, and every blur-reference run passes

5.8 Shared-image comparisons and selection sensitivity

Table 6 and Figure 9 report paired differences between the two evaluated systems on the same 68 originals. ResNet34 has a larger complete-family gap by 6.69 percentage points under mean fill and 2.82 points under blur. The singleton excess changes in different directions across operators: ResNet minus MobileNet is +2.18 points under mean fill and −1.10 points under blur.

These differences describe the two particular pretrained systems. They should not be generalized as an architecture-wide causal effect because weights and training histories differ. Their main design value is that cohort differences no longer explain the comparison: the origins, targets, fitting masks and eligibility population are shared.

Retaining all 100 selected originals produces complete gaps of 17.41%, 19.02%, 24.45% and 21.50% in the four primary setting order. The corresponding held-out singleton excesses are 2.30, 5.81, 4.67 and 4.43 percentage points. The all-selected results preserve the main qualitative pattern. They address conditioning on original-image correctness, rather than supplying an independent confirmation dataset. Online Resource 1 includes their intervals and a visual comparison of the two populations.

Operator	Gap difference pp	Singleton excess difference pp
Mean fill	6.69 [5.76, 7.60]	2.18 [1.37, 3.06]
Blur	2.82 [1.69, 3.89]	-1.10 [-2.13, -0.07]

Table 6 Paired ResNet34 minus MobileNet differences on 68 shared originals. Gap uses the complete family; singleton excess uses the held-out family. Brackets give class-stratified paired image-bootstrap intervals. Differences combine architecture and pretrained-weight characteristics

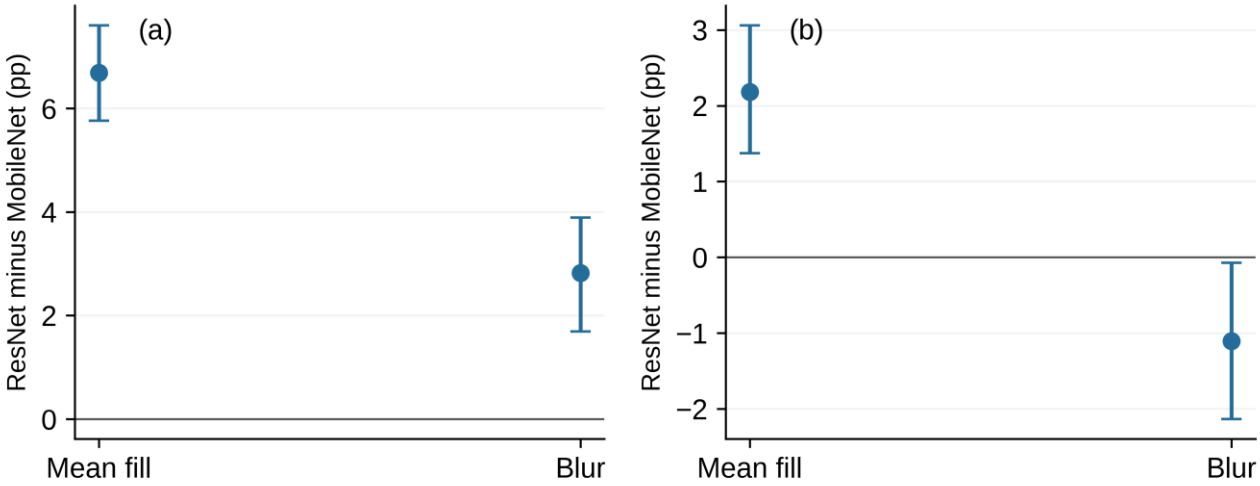


Fig. 9 Paired ResNet34 minus MobileNet differences. Panel a shows the complete fixed-order gap and panel b held-out singleton excess. Points and error bars are means and paired image-bootstrap intervals on 68 originals. The horizontal zero line indicates equal outcomes for the two systems

6 Reproducibility and computational accounting

6.1 Verified response sources and optimization certificates

The completed confirmation snapshot supplies all scientific outcomes. All 30,392 canonical JSON digests pass verification. The 480 response aggregates exactly match 3,840 census chunks; 400 IG results match 9,600 integration chunks. The original-score references, all 120 region maps and 100 stored fitting/evaluation assignments also pass the structural audit.

Independent MILP certifies 960 zero-threshold target-margin optima: 800 grid and 160 SLIC16, including complete and held-out populations for all selected origins. Every certified objective equals the dynamic-programming result, with solver gap zero. The analysis produces 24,960 per-image metric rows and 5,568 interval rows. These counts identify verified computations, not independent sample sizes.

One auxiliary checkpoint, singletons/resnet34_742168c2b303fe1c_grid_blur.json, was omitted from the complete export. It was restored byte-for-byte from the earlier export after verifying identical completed jobs, run identity, model-weight records and the original singleton ledger event. All 16 drops match the completed job exactly. The repaired archive contains all 480 auxiliary checkpoints and preserves every original member unchanged. This archival recovery introduces no new model queries or changes to scientific results.

6.2 Snapshot and arithmetic sensitivity

An earlier grid-only snapshot has 8,212 verified JSON digests and the same frozen run identity, protocol and source. Scientific score tables are not pooled across snapshots. Both original ZIP CRC checks pass, and archive verification checks for duplicate or unsafe member paths.

After the T4/A100 session transition, 158 shared ResNet grid jobs differ numerically across snapshots. The largest shared discrepancy is 0.00007629, below the frozen 0.002 replay tolerance. Complete K changes by at most 0.00458 percentage points and held-out singleton agreement by at most 0.00501 points per affected job. GPU arithmetic is a plausible cause, rather than a reconstructed proof of overwrite provenance.

Zero-threshold direction counts can be sensitive near ties despite small absolute response discrepancies. Keeping the frozen nonzero-threshold sensitivities visible provides context for this issue. The archived-source and input hashes, recovered checkpoint path, recovery verification, numerical comparisons and certificate records are provided in the online resources.

6.3 Method access and observed compute

Method-access costs differ from the cost of collecting an exhaustive diagnostic family. Singleton, Linear-32 and the 96-query predictors receive 16, 32 and 96 intervention responses respectively, plus the shared original score. Evaluating the full size-four response table requires 1,820 interventions per job, and the saved singleton responses are additional. The oracle's evaluated outcomes must not be counted as predictor training access.

IG uses 768 gradient-bearing examples per reference across the 256- and 512-step resolutions, of which the 512-step result supplies the evaluated regional ranking. Grad-CAM uses one gradient-bearing original per model/image. Forward-only removals and gradient-bearing examples are reported separately because a gradient evaluation is not equivalent in cost to a simple forward pass.

Table 7 distinguishes intended logical counts from observed recorded evaluations. The physical ledger combines identical verified event keys across the two snapshots, counts shared events once and preserves 1,501 earlier-only events. It records repeated and restart/replay work without merging scientific response outcomes.

Quantity	Intended logical count	Observed recorded count
Forward-only image evaluations	881,880	1,172,111
Forward and gradient evaluations	307,400	307,400
Verified ledger events	Not a fitting budget	16,462

Table 7 Logical research counts and observed evaluation counts. Observed counts include repeated and restart/replay work. The event count is a ledger property, not an image sample size or method fitting budget

Logged forward-only durations total 1,188.612 seconds, and gradient-bearing durations total 860.301 seconds. IG chunk durations total 842.215 seconds, about 14.04 minutes. These are summed durations of enclosed work, not full cell wall time or GPU billing. Unlogged interruptions and export overhead cannot be reconstructed. The frozen CPU analysis can recompute scientific metrics from collected responses without a GPU.

7 Discussion

7.1 Interpreting fixed visual explanation benchmarks

The main result changes the interpretation of disagreement. When a benchmark asks one ranking to represent many context-dependent directions, low agreement can combine a limit of the representation with deficiencies of a particular ranking. On this family, an optimal order still misses roughly one in six to one in four comparisons on average. Reporting that gap makes the benchmark's demand visible.

The small additional ordering cost gives a more specific account than a general statement that contextual importance is inconsistent. Most dominant pair directions can be arranged in almost

one coherent order, but that order still disagrees with many individual contexts. A global summary can therefore be well defined and accurately optimized while remaining incomplete as a predictor of local removal effects.

The ceiling is useful as an audit reference. It is not available without evaluated outcomes, and subtracting a shared K from competing fixed rankings does not alter their relative ranking. Its value is in allocating their total disagreement between the common task-dependent gap and method-specific excess. A practical benchmark can report raw agreement, K and K−A together without presenting K as a deployable explainer.

7.2 Baseline choice and context as a prediction input

Direct removal responses provide a relevant baseline when the target is a removal-order prediction. The singleton baseline performs strongly here despite its low intervention budget. Ridge fitting at 96 queries improves its agreement consistently, while the 32-query ranking does not. This suggests that the usefulness of a fitted response model depends on its declared access budget, rather than on its name or complexity alone.

Allowing the output to depend on context improves agreement beyond the fixed-order ceiling. A context predictor and a saliency ranking therefore answer different operational requests. If an application needs a direction for a particular removal context, that context can be included in prediction. If it needs one stable visual summary, the fixed representation should be evaluated against its own attainable limit.

These findings concern prediction of the declared model-response contrasts. They do not establish that a clinician or another human would make better decisions using a contextual predictor, nor that the predictor estimates semantic causal importance. Those applications require different data, interventions and outcomes. The present experiment supplies a reproducible audit component for such future work.

7.3 A reporting protocol for removal-order audits

An audit should state the target score, region construction, replacement operator, context population and threshold before interpreting agreement. It should report the number of independent images separately from the number of dependent comparisons. For nonzero thresholds, coverage should accompany performance so that easy retained subsets are not confused with improvements on the original population.

For small finite families, calculate the pair-majority and best-total-order agreements and report the three error terms. Include direct singleton responses and a fitted response baseline with explicit access costs. If fitting responses are held out, exclude every evaluated comparison containing either fitted constituent response. Keep a context-conditioned output separate from one fixed order.

Numerical validity should also accompany the comparison. Completeness and resolution checks can reveal integration failures that would otherwise be hidden by a final map. Retain observed failures in the primary results and identify any passing-only population as selected by those outcomes. Structural checks, a second optimization formulation and snapshot provenance then make the reported diagnostic limit reviewable.

8 Limitations and future evaluation

The study covers one ten-class dataset and two pretrained systems. The original selection is class balanced, but joint correctness produces an uneven 68-image primary cohort. The all-selected sensitivity helps describe correctness conditioning without removing it from the primary estimand. There is one fitting-mask assignment per image and no training-seed or cross-dataset replication.

Mean fill and blur may produce inputs outside the natural-image distribution. Blur can retain substantial low-frequency evidence. SLIC-derived regions are image-based partitions rather than semantic removals, and the eligible geometry subset contains only 13 origins. The findings therefore concern these finite operational interventions and do not estimate object-level or clinical causal effects.

IG completeness failures constrain the mean-reference results, especially MobileNet. Passing the stated finite-resolution gate is also not a proof that the exact continuous integral has been recovered. Grad-CAM uses positive regional means and IG signed regional sums, so the experiment evaluates particular score conversions. Other aggregation conventions or target functions require explicit new analyses.

The finite optimum is fitted to the evaluated outcomes and requires census compute. The held-out predictors have a genuine response split within each image, but are not learned transfers to unseen images. Intervals quantify conditional image variation and do not include uncertainty from dataset choice, intervention choice, pretrained-model training or alternative split seeds.

Small cross-snapshot arithmetic changes motivate keeping near-tie sensitivities visible. The omitted auxiliary singleton checkpoint has been recovered from the original export and verified against the unchanged completed job. Both original checkpoint archives, the repaired copy and the recovery record are retained for full archival reproduction.

Further evaluation could test additional datasets and trained systems, repeated fitting assignments, richer region families and naturalistic interventions. Any semantic or human-decision claim would require suitable outcome measures and validation of the interventions. These extensions are future work; they are not prerequisites inferred from the current finite-response estimates or represented as completed experiments.

9 Conclusion

An exhaustive fresh-image confirmation shows substantial contextual disagreement for fixed regional orders, with very small additional loss from enforcing global ordering. On held-out removal responses, singleton and fitted response baselines expose ranking-specific excess, and context-conditioned prediction recovers some directions that no fixed order can represent. The result supports a concrete evaluation practice: report the attainable fixed-order agreement, contextual minority loss and method-specific excess alongside raw explainer agreement, numerical checks and intervention scope.

Declarations

Funding

This research received no external funding.

.

Data and code availability

Imagenette is available from the public dataset repository [14]. The frozen protocol, collection and analysis source, derived per-image results, interval tables and optimization certificates are supplied in Online Resource 2 (Data_and_Code.zip). The original checkpoint exports and the verified repaired copy are preserved for archival reproduction. The Zenodo repository record for the study is https://zenodo.org/records/23304821.

Ethics and consent

This study analyzes existing publicly available benchmark images. It did not recruit human participants, collect clinical records or conduct experiments on living animals. It reports no clinical intervention. Any institutional declaration applicable to the final submission remains to be confirmed by the author.

Competing interests

The author declares that they have no competing interests

References

[1] Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In: ICML. PMLR. 2017;70:3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html

[2] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: ICCV. 2017:618–626. https://doi.org/10.1109/ICCV.2017.74

[3] Covert I, Lundberg S, Lee SI. Explaining by Removing: A Unified Framework for Model Explanation. J Mach Learn Res. 2021;22(209):1–90. https://jmlr.org/papers/v22/20-1316.html

[4] Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity Checks for Saliency Maps. Adv Neural Inf Process Syst. 2018;31. https://papers.nips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

[5] Yeh CK, Hsieh CY, Suggala AS, Inouye DI, Ravikumar PK. On the (In)fidelity and Sensitivity of Explanations. Adv Neural Inf Process Syst. 2019;32. https://papers.nips.cc/paper_files/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html

[6] Tsang M, Rambhatla S, Liu Y. How Does This Interaction Affect Me? Interpretable Attribution for Feature Interactions. Adv Neural Inf Process Syst. 2020;33. https://arxiv.org/abs/2006.10965

[7] Tsai CP, Yeh CK, Ravikumar P. Faith-Shap: The Faithful Shapley Interaction Index. J Mach Learn Res. 2023;24(94):1–42. https://jmlr.org/papers/v24/22-0202.html

[8] Laberge G, Pequignot Y, Mathieu A, Khomh F, Marchand M. Partial Order in Chaos: Consensus on Feature Attributions in the Rashomon Set. J Mach Learn Res. 2023;24(364):1–50. https://jmlr.org/papers/v24/23-0149.html

[9] Neuhof B, Benjamini Y. Confident Feature Ranking. In: AISTATS. PMLR. 2024;238:1468–1476. https://proceedings.mlr.press/v238/neuhof24a.html

[10] Hooker S, Erhan D, Kindermans PJ, Kim B. A Benchmark for Interpretability Methods in Deep Neural Networks. Adv Neural Inf Process Syst. 2019;32. https://papers.nips.cc/paper/9167-a-benchmark-for-interpretability-methods-in-deep-neural-networks.pdf

[11] Zheng X, Shirani F, Chen Z, Lin C, Cheng W, Guo W, Luo D. F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI. In: ICLR. 2025. https://arxiv.org/abs/2410.02970

[12] Edin J, Motzfeldt AG, Christensen CL, Ruotsalo T, Maaløe L, Maistro M. Normalized AOPC: Fixing Misleading Faithfulness Metrics for Feature Attribution Explainability. In: ACL. 2025:1715–1730. https://doi.org/10.18653/v1/2025.acl-long.86

[13] Huang W, Zhao X, Jin G, Huang X. SAFARI: Versatile and Efficient Evaluations for Robustness of Interpretability. In: ICCV. 2023:1988–1998. https://doi.org/10.1109/ICCV51070.2023.00190

[14] Howard J, fastai contributors. Imagenette. Dataset repository. Accessed 11 October 2026. https://github.com/fastai/imagenette

[15] Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, et al. Searching for MobileNetV3. In: ICCV. 2019:1314–1324. https://doi.org/10.1109/ICCV.2019.00140

[16] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: CVPR. 2016:770–778. https://doi.org/10.1109/CVPR.2016.90

[17] Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. IEEE Trans Pattern Anal Mach Intell. 2012;34(11):2274–2282. https://doi.org/10.1109/TPAMI.2012.120

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx
- DataandCode.zip