BEYOND ACCURACY RANKINGS: RELATIONAL COMPETENCE IN OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Two object detectors can achieve nearly identical scalar performance while succeeding on different instances. We ask whether these differences are incidental or reflect reproducible structure in how models solve the task. We introduce *relational competence*, a population-relative characterization of detector behavior based on residual co-observability after exact control of detector coverage and instance observability with a fixed-margin null. Across eight detectors, residual relational structure remains after this marginal control. Independent retrainings also recover design-associated relational neighborhoods: Faster R-CNN and RetinaNet checkpoints matched to within 0.009 AP have mean relational distances of 0.0113 within design and 0.1537 across designs, yielding $\Delta D = 0.1424$ with a 95% bootstrap interval of $[0.09, 0.16]$. Under controlled resolution and JPEG degradation, population-relative signature changes are likewise reproducible across retrainings, but a predefined same- versus cross-intervention comparison does not establish intervention-specific directions. These results distinguish scalar performance, relational organization, retraining reproducibility, and response to intervention as separate aspects of model behavior. Relational competence therefore complements conventional evaluation by measuring not only how often models succeed, but how their successes are organized relative to a model population.

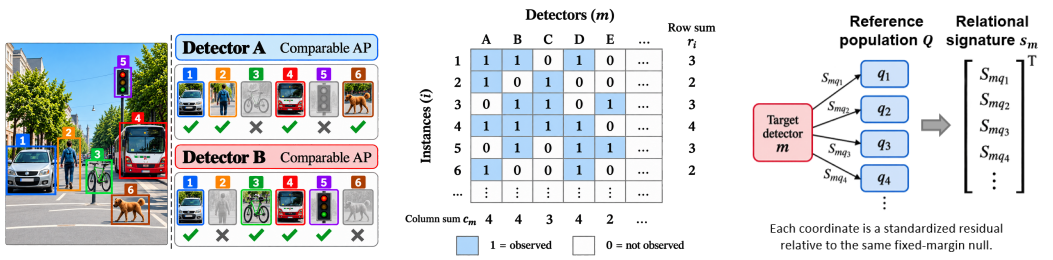
1 INTRODUCTION

Two object detectors can achieve nearly identical average precision while succeeding on different ground-truth instances. Scalar metrics such as average precision (AP) (Lin et al., 2014) summarize how much of a task a model solves, but they collapse the organization of instance-level successes and failures into a single number. Prior work has shown that similarly performing models can still differ in their predictions and failure modes (D'Amour et al., 2022; Geirhos et al., 2020a; Shah et al., 2023; Theodoridis et al., 2026). This raises a more specific question: do these differences merely reflect marginal performance, or do they form reproducible structure across a population of models? **We ask whether detector competence contains population-relative relational structure that scalar performance rankings do not capture.**

Pairwise disagreement alone does not resolve this question. Apparent similarity or difference between two detectors can reflect their overall success rates, the fact that some instances are broadly observable across models, or pair-specific organization beyond these effects. Existing analyses of disagreement, exclusive errors, and structured failure modes reveal substantial behavioral heterogeneity (Bolya et al., 2020; Eyuboglu et al., 2022; Sohoni et al., 2020; Lee et al., 2023; Theodoridis et al., 2026), but heterogeneity alone does not establish relational structure. A relational analysis must therefore separate pair-specific dependence from the detector and instance marginals that shape raw agreement.

We introduce *relational competence* to make this distinction measurable. We represent detector behavior with a binary observability matrix over ground-truth instances and compare pairwise co-observability against a fixed-margin null that exactly preserves detector coverage and instance observability (Strona et al., 2014; Carstens, 2015). The resulting standardized residuals capture pair-specific organization beyond these observed marginals and define each detector's relational signature relative to a fixed reference population. Relational competence is therefore population-relative by

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.



(a) (b) (c)

Figure 1: **Relational competence measures structure in how detector successes are organized.** **(a)** Schematic illustration: detectors with comparable scalar performance can succeed on different instances. **(b)** The observability matrix records instance-level successes across a detector population; detector coverage and instance observability define its observed column and row marginals. **(c)** Pairwise co-observability is compared with a fixed-margin null that preserves these marginals exactly. The resulting standardized residuals define each detector's population-relative relational signature with respect to a fixed reference population.

construction: it describes a detector's position within a specified comparison system rather than an intrinsic model property. Figure 1 summarizes this construction.

We study relational competence through three questions. First, does relational organization remain after detector coverage and instance observability are controlled? Across eight heterogeneous detectors, the observed co-observability structure departs systematically from the fixed-margin null. Second, does this organization persist across independent retraining? Faster R-CNN (Ren et al., 2015) and RetinaNet (Lin et al., 2017), each trained with three independent seeds and matched to within 0.009 AP, have mean relational distances of 0.0113 within design and 0.1537 across designs. Third, how does relational organization change under controlled input degradation? Resolution and JPEG degradation produce reproducible population-relative changes across the tested retrainings, although a predefined same- versus cross-intervention comparison does not establish intervention-specific directions.

Our contributions are:

- **A population-relative measure of detector competence.** We define relational competence from instance-level observability and use exact marginal control to isolate pair-specific co-observability beyond detector coverage and instance observability.
- **Evidence of residual relational organization and retraining stability.** We show that similar scalar performance can conceal different relational organization and that, in the tested controlled setting, same-design retrainings remain closer in relational space than cross-design models.
- **A controlled analysis of relational change under degradation.** We characterize reproducible population-relative changes under resolution and JPEG degradation while finding no predefined evidence for intervention-specific directions.

Similar accuracy does not imply similar relational competence. Scalar performance measures how much a detector gets right; relational competence measures how those successes are organized relative to other models.

2 Related Work

2.1 Object Detection Evaluation Beyond Scalar Performance

Object detection is commonly summarized by aggregate metrics such as AP and AR (Everingham et al., 2010; Lin et al., 2014). Later work introduced diagnostics for error types, calibration, corruption, and distribution shift (Bolya et al., 2020; Küppers et al., 2020; Pathiraja et al., 2023; Michaelis et al., 2019; Mao et al., 2023). These methods reveal behavior hidden by a single performance num-

ber, but they primarily characterize each detector through its own errors, confidence, or degradation. We instead study how instance-level successes are organized across a population of detectors.

2.2 From Error Agreement to Relational Structure

Models with similar aggregate performance can differ substantially at the example level (Geirhos et al., 2020b; Bhojanapalli et al., 2021). Prior work measures such differences through error consistency, prediction disagreement, or broader functional similarity (Geirhos et al., 2020b; Bahri & Jiang, 2021; Lee et al., 2023; Klabunde et al., 2025). These approaches establish whether models behave similarly or differently, but pairwise disagreement alone does not determine whether those differences form a reproducible population-relative structure. We focus on this distinction by studying residual co-observability across a detector population after controlling for detector and instance marginals.

2.3 Beyond Marginal Explanations

Ability–difficulty models explain structured evaluation outcomes through model- and instance-level propensities (Martínez-Plumed et al., 2019; Chen et al., 2019; Martínez-Plumed et al., 2016). We begin from this scalar view but ask whether detector co-observability retains structure after the observed row and column marginals are preserved exactly. Under distribution shift, conventional robustness evaluation asks how much aggregate performance changes (Michaelis et al., 2019; Mao et al., 2023); we instead examine how the residual relational organization changes. Our framework therefore treats marginal performance as structure to control and the remaining population-relative dependence as the object of study.

3 Measuring Relational Competence

3.1 Observability and Marginal Structure

We characterize detector behavior through instance-level success patterns. Let an evaluation set contain N ground-truth instances and a comparison population contain M detectors. We define the binary observability matrix $\mathbf{Z} \in \{0,1\}^{N\times M}$ by

$$Z_{im} = \begin{cases} 1, & \text{if detector } m \text{ successfully observes instance } i, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

We select detector-specific confidence thresholds on a disjoint calibration split and fix them on the analysis split. We then match predictions to non-crowd ground-truth instances using a deterministic, class-aware, one-to-one procedure. An instance is observed only when its matched prediction satisfies the prescribed overlap criterion.

The matrix $\mathbf{Z}$ induces two observed marginals,

$$c_m = \sum_{i=1}^{N} Z_{im}, \qquad r_i = \sum_{m=1}^{M} Z_{im}, \tag{2}$$

where c_m denotes detector coverage and r_i denotes instance observability within the comparison population. These marginals determine how often each detector succeeds and how broadly each instance is observed, but not which detectors succeed on the same instances. Two observability matrices can therefore share identical $\{c_m\}$ and $\{r_i\}$ while differing in their pairwise organization.

We formalize a scalar explanation with a Rasch/1PL model (Rasch, 1993),

$$\Pr(Z_{im} = 1 \mid a_m, d_i) = \sigma(a_m - d_i), \qquad \sigma(x) = \frac{1}{1+\exp(-x)}, \tag{3}$$

where a_m denotes detector ability and d_i denotes shared instance difficulty. The model contains no detector–detector interaction. We fit it by L_2-regularized maximum likelihood, with the ridge penalty resolving the joint shift ambiguity. The Rasch model serves only as a parametric scalar baseline; the relational construction below does not depend on its fitted parameters. Because this baseline reproduces the observed marginals only approximately, we next ask what remains when both marginals are preserved exactly.

3.2 Residual Relations Beyond Exact Marginals

We construct a fixed-margin null over the full observability matrix by randomizing the placement of successes while preserving detector coverage and instance observability exactly:

$$\sum_{i=1}^{N} \widetilde{Z}_{im} = c_m \quad \forall m, \qquad \sum_{m=1}^{M} \widetilde{Z}_{im} = r_i \quad \forall i. \tag{4}$$

We generate the null ensemble using Curveball (Strona et al., 2014). Each null matrix preserves how often detectors and instances participate in successful observations while randomizing how those successes are arranged across detector–instance pairs.

For detectors a and b, the observed co-observability is

$$O_{ab}^{\text{obs}} = \sum_{i=1}^{N} Z_{ia} Z_{ib}. \tag{5}$$

Evaluating the same statistic over the shared null ensemble yields a pair-specific mean μ_{ab}^{null} and standard deviation $\sigma_{ab}^{\text{null}}$. We define the residual relation as

$$S_{ab} = \frac{O_{ab}^{\text{obs}} - \mu_{ab}^{\text{null}}}{\sigma_{ab}^{\text{null}}}. \tag{6}$$

Positive values indicate more shared successes than expected under the preserved marginals, whereas negative values indicate fewer.

This construction separates relational organization from raw agreement. Raw co-observability records how often two detectors succeed together, whereas S_{ab} asks whether their joint success remains unusual after detector coverage and instance observability are fixed exactly. Thus, S_{ab} is a standardized residual relative to the fixed-margin null rather than a direct measure of behavioral effect magnitude.

3.3 Population-Relative Relational Signatures

We characterize each detector by its residual relations to a common reference population. Let $\mathcal{Q} = \{q_1, \ldots, q_K\}$ denote an ordered panel of K reference detectors. For $m \notin \mathcal{Q}$, we define

$$\mathbf{s}_m = \left[S_{mq_1}, S_{mq_2}, \ldots, S_{mq_K}\right]^\top. \tag{7}$$

The reference panel fixes which relations define the signature coordinates. Their values remain population-relative because each S_{mq} is standardized with respect to the fixed-margin null of the evaluation matrix in which it is measured. The signature $\mathbf{s}_m$ therefore represents a detector's position within a specified comparison system rather than an intrinsic model property.

We compare two detectors using

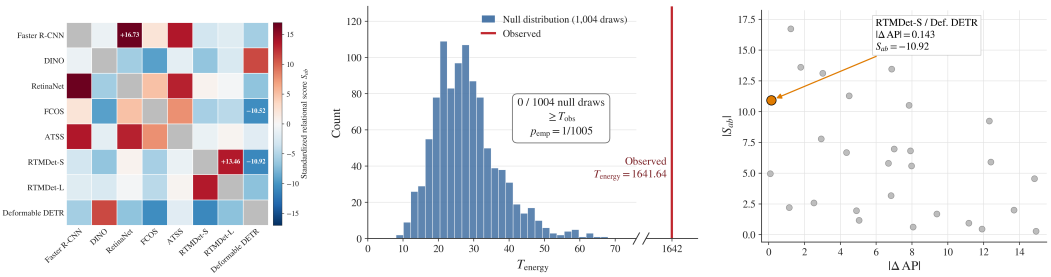
$$D(m, n) = 1 - \rho(\mathbf{s}_m, \mathbf{s}_n), \tag{8}$$

where ρ denotes the Pearson correlation. Small values indicate similar relational profiles, while larger values indicate greater separation. When comparing two detectors within $\mathcal{Q}$, we exclude their two self-associated coordinates and compute the correlation over the remaining $K - 2$ coordinates. We keep the reference panel and its ordering fixed when comparing independently trained checkpoints.

3.4 Evaluation and Statistical Inference

We keep calibration, uncertainty estimation, and hypothesis testing separate. Detector-specific thresholds are selected only on the calibration split and fixed thereafter. The primary setting is P80 at IoU 0.50, where P80 maximizes recall subject to precision of at least 0.80. The remaining combinations of P80/P90 and IoU $0.50/0.75$ serve as predefined sensitivity conditions.

For the primary retraining comparison, we quantify sampling uncertainty with an image-cluster bootstrap that resamples images together with all associated ground-truth instances. Because resampling changes the observed marginals, we recompute the fixed-margin null for every bootstrap



(a) (b) (c)

Figure 2: **Relational structure remains after exact marginal control and is not specified by AP proximity.** **(a)** Standardized residual relations S_{ab} across the eight-detector population under the primary P80, IoU 0.50 condition. **(b)** The observed population-level energy, $T_{\text{energy}} = 1641.64$, exceeds all 1,004 fixed-margin null realizations, giving $p_{\text{emp}} = 1/1005$. **(c)** Each point denotes one detector pair. RTMDet-S and Deformable DETR differ by only $|\Delta\text{AP}| = 0.143$ but have $S_{ab} = -10.92$.

matrix before recalculating relational scores, signatures, and distances. We use 100 bootstrap replicates and 84 null samples per replicate and report percentile confidence intervals. These intervals quantify image-cluster sampling uncertainty for the fixed evaluated checkpoints, not variability over possible training seeds. We report the remaining sensitivity conditions descriptively unless stated otherwise.

For pairwise inference, we compute two-sided empirical p-values with a finite-sample $+1$ correction and control the false discovery rate across pairs using the Benjamini–Hochberg procedure. We assess population-level departure separately using

$$T_{\max} = \max_{a<b} |S_{ab}|, \qquad T_{\text{energy}} = \sum_{a<b} S_{ab}^2,$$

with one-sided upper-tail empirical tests. Additional details on the detector population, calibration protocol, fixed-margin sampling, and primary bootstrap procedure are provided in Appendix A.

4 Relational Structure Beyond Exact Margins

4.1 Residual Structure Beyond Exact Marginal Control

We first test whether pairwise organization remains after detector coverage and instance observability are preserved exactly. Across the eight-detector population, the residual relation matrix contains substantial positive and negative structure (Figure 2A). This departure is evident at the population level: $T_{\max} = 16.73$ and $T_{\text{energy}} = 1641.64$, and both exceed all 1,004 fixed-margin null realizations, yielding $p_{\text{emp}} = 1/1005$ (Figure 2B). The evidence also extends across detector pairs. After Benjamini–Hochberg correction, $20/28$ pairs satisfy $q < 0.05$, including $18/28$ at $q < 0.01$. Detector coverage and instance observability therefore do not determine the observed pairwise organization.

4.2 Similar Accuracy Does Not Determine Relational Structure

The population-level result leaves open whether near-equal scalar performance necessarily implies a near-null relational difference. RTMDet-S and Deformable DETR provide a direct counterexample. Their COCO AP values are 44.650 and 44.507, respectively, giving $|\Delta\text{AP}| = 0.143$. Yet they share 8089 observed successes, compared with a fixed-margin expectation of 8296.09. The resulting residual is -207.09 shared successes, or -2.50% relative to the null expectation, with a standardized relation of $S_{ab} = -10.92$ (Figure 2C). Table 1 summarizes the population-level, pairwise, and near-equal-AP evidence.

The count residual and standardized score capture different aspects of the same departure. The former quantifies its magnitude in shared successes, whereas S_{ab} measures how unusual that magnitude

Table 1: **Evidence for relational organization beyond exact observed marginals.** Null-based results use $1{,}004$ fixed-margin realizations. The final block reports the near-equal-AP contrast and separates standardized null departure from its count-scale effect.

Evidence	Statistic	Result
Population-level departure		
Omnibus maximum	$T_{\max}$	16.73
	empirical p	$1/1005$
Omnibus energy	T_{energy}	1641.64
	empirical p	$1/1005$
Pairwise breadth		
Formal inference	BH $q < 0.05$	$20/28$
Formal inference	BH $q < 0.01$	$18/28$
Largest departure	Faster R-CNN vs. RetinaNet	$\lvert S_{ab}\rvert = 16.73$
Near-equal-AP contrast		
Aggregate performance	RTMDet-S vs. Deformable DETR	44.650 vs. 44.507
Performance gap	$\lvert\Delta\text{AP}\rvert$	0.143
Observed co-observability	O_{ab}^{obs}	8089
Fixed-margin expectation	μ_{ab}^{null}	8296.09
Residual co-observability	$O_{ab}^{\text{obs}} - \mu_{ab}^{\text{null}}$	-207.09
Relative residual	$(O^{\text{obs}} - \mu^{\text{null}})/\mu^{\text{null}}$	-2.50%
Standardized departure	S_{ab}	-10.92

is under the preserved marginals. This example does not imply statistical independence between AP and relational structure. It shows that proximity in AP does not determine the residual relation between detectors. We next ask whether this population-relative organization persists across independent retraining.

5 Relational Profiles Persist Across Independent Retraining

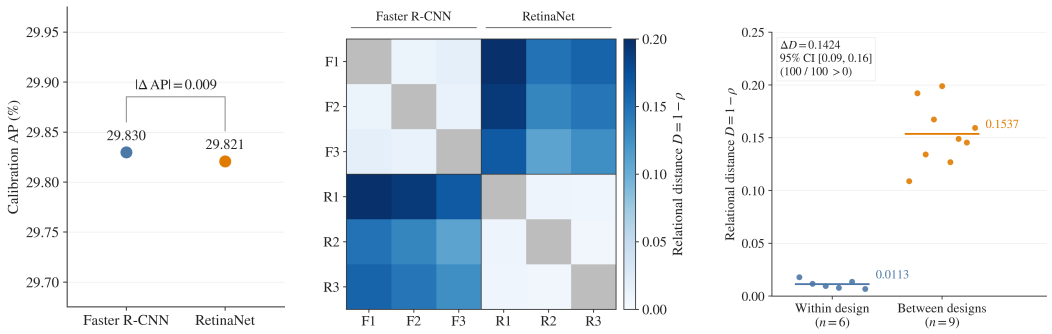
5.1 Independent Retrainings Recover Relational Neighborhoods

We next test whether relational organization survives independent retraining. Faster R-CNN and RetinaNet are each trained with three independent seeds, and checkpoints are selected using only the disjoint calibration split. The selected checkpoints have calibration AP 29.830 and 29.821, respectively, giving $|\Delta\text{AP}| = 0.009$ (Figure 3a). Their calibration AR values differ by 0.291, so the control closely matches AP rather than all scalar performance measures. We evaluate all six checkpoints against the same ordered eight-detector reference panel.

The retrained checkpoints form two clearly separated relational neighborhoods (Figure 3b). The mean distance is $D_{\text{within}} = 0.0113$ within design and $D_{\text{between}} = 0.1537$ across designs, yielding $\Delta D = 0.1424$ (Figure 3c). An image-cluster bootstrap that recomputes the fixed-margin null within each resampled matrix gives a 95% percentile interval of approximately $[0.09, 0.16]$, with all 100 replicates remaining positive. This interval quantifies sampling uncertainty over evaluation images for the six fixed checkpoints, not variability over possible training seeds; implementation details are provided in Appendix A. The result therefore supports reproducible design-associated relational organization in the tested setting without establishing architecture-level causality.

5.2 The Separation Is Population-Relative

The primary separation persists as the reference representation changes, but its magnitude does not remain fixed. Removing both direct reference anchors reduces ΔD from 0.1424 to 0.0899. Across leave-one-reference-out variants, ΔD remains positive but ranges from 0.0376 to 0.1754, with the strongest attenuation occurring when Faster R-CNN is removed. No single reference coordinate is therefore necessary for the primary separation, but the measured separation depends on the reference population.



(a) (b) (c)

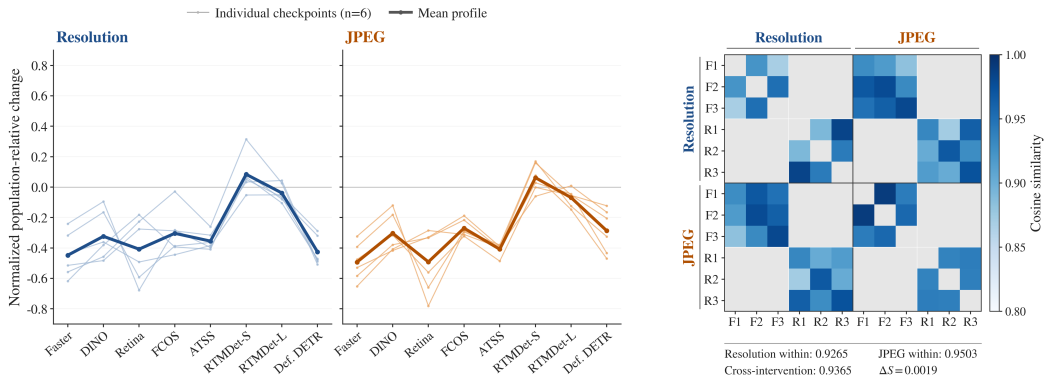
Figure 3: **Relational profiles persist across independent retraining under closely matched AP.** **(a)** Faster R-CNN and RetinaNet are selected at calibration AP 29.830 and 29.821, respectively ($|\Delta\text{AP}| = 0.009$). **(b)** Pairwise relational distances among three independently trained checkpoints of each design reveal tight within-design neighborhoods and larger cross-design separations under the primary P80, IoU 0.50 condition. **(c)** The mean distance is 0.0113 within design and 0.1537 between designs, yielding $\Delta D = 0.1424$. The corrected image-cluster bootstrap gives a 95% percentile interval of approximately $[0.09, 0.16]$, with all 100 replicates remaining positive.

Table 2: **Retraining stability and its population-relative boundary.** The primary P80, IoU 0.50 result reports corrected image-cluster bootstrap inference. Operating-point and reference-composition analyses are descriptive point estimates.

Evidence	Statistic	Result
Primary retraining separation		
Within vs. between	$D_{\text{within}}/D_{\text{between}}$	0.0113 / 0.1537
Separation	ΔD	0.1424
Bootstrap uncertainty	95% percentile CI	[0.09, 0.16]
Bootstrap sign	positive replicates	100/100
Operating-point sensitivity		
P80 / IoU 0.50	full 8D / post-hoc 6D ΔD	0.1424 / 0.0899
P90 / IoU 0.50	full 8D / post-hoc 6D ΔD	0.1058 / 0.0224
P80 / IoU 0.75	full 8D / post-hoc 6D ΔD	0.1089 / 0.0151
P90 / IoU 0.75	full 8D / post-hoc 6D ΔD	0.0843 / −0.0103
Reference dependence		
Direct-anchor exclusion	P80 / IoU 0.50, 6D ΔD	0.0899
Leave-one-reference-out	range of ΔD	0.0376–0.1754
Strongest attenuation	remove Faster R-CNN	0.0376

The operating-point analysis shows the same boundary. In the full eight-dimensional representation, the descriptive ΔD estimates remain positive across all four tested conditions: 0.1424, 0.1058, 0.1089, and 0.0843. After direct-anchor exclusion, the corresponding estimates are 0.0899, 0.0224, 0.0151, and −0.0103. These non-primary results are descriptive point estimates; the complete sensitivity summary is provided in Appendix B. The retraining separation is therefore reproducible under the primary analysis, but its magnitude depends on the reference representation and evaluation condition (Table 2).

Relational profiles are therefore reproducible across the tested independent retrainings, while the magnitude of their separation remains tied to the comparison system in which they are measured. We next ask how this population-relative organization changes under controlled input degradation.



(a) (b)

Figure 4: **Relational change is reproducible across retrainings but does not clearly identify the degradation.** **(a)** Unit-normalized population-relative signature changes across the eight reference relations. Independent retrainings exhibit similar coordinate-wise profiles within resolution degradation and JPEG compression, with mean directional cosines of 0.9265 and 0.9503, respectively. **(b)** Directional-alignment structure for the predefined comparisons. Same-intervention alignment remains high, but cross-intervention alignment is comparably strong, with $S_{\text{same}} = 0.9384$ and $S_{\text{cross}} = 0.9365$.

6 Reproducible Relational Change under Input Degradation

6.1 Degradation Induces Reproducible Relational Change

We next test whether relational organization changes reproducibly under input degradation. For each checkpoint, we define $\Delta\mathbf{s} = \mathbf{s}_{\text{degraded}} - \mathbf{s}_{\text{clean}}$ and compare the unit-normalized change directions across independent retrainings. Because the clean and degraded signatures are standardized within their respective evaluation matrices, $\Delta\mathbf{s}$ represents change in the population-relative signature representation rather than an intrinsic model-space displacement. Appendix C provides the degradation settings, change-vector construction, normalization, and predefined directional-comparison protocol.

The change directions remain highly consistent within both tested degradations. Mean within-intervention cosine similarity is 0.9265 for resolution degradation and 0.9503 for JPEG compression. Figure 4a shows that this agreement extends across the eight reference coordinates: independently trained checkpoints exhibit similar coordinate-wise change profiles under the same degradation. The corresponding performance-retention ratios are 0.8274 for resolution and 0.7709 for JPEG (Table 3). Relational change is therefore reproducible across the tested retrainings, but reproducibility alone does not imply that the change direction identifies the intervention.

6.2 Reproducibility Does Not Establish Intervention Specificity

The predefined specificity contrast shows little separation between the two degradation families. Mean same-intervention alignment is $S_{\text{same}} = 0.9384$, while mean cross-intervention alignment is $S_{\text{cross}} = 0.9365$, giving $\Delta S = 0.0019$. Figure 4b shows the same pattern at the alignment-structure level: cross-intervention similarities remain comparable to those observed within each degradation. The predefined analysis therefore does not establish intervention-specific relational change.

The degradation result therefore has a clear boundary. Population-relative relational change is reproducible across the tested retrainings, but the observed direction does not clearly distinguish resolution degradation from JPEG compression. We do not interpret the present evidence as establishing an intrinsic change direction, an intervention-specific mechanism, or a universal degradation response.

Table 3: **Relational change under controlled input degradation.** The table separates degradation strength, across-retraining reproducibility, and the predefined intervention-specificity contrast. Specificity results are descriptive.

Evidence	Statistic	Result
Degradation conditions		
Resolution	scale factor	0.5
	performance retention	0.8274
JPEG	quality	30
	performance retention	0.7709
Across-retraining reproducibility		
Resolution	mean within-intervention cosine	0.9265
JPEG	mean within-intervention cosine	0.9503
Intervention specificity		
Same intervention	S_{same}	0.9384
Cross intervention	S_{cross}	0.9365
Specificity contrast	ΔS	0.0019
Interpretation	predefined contrast	not established

7 Discussion: What Relational Competence Reveals

Similar performance does not imply similar organization of success. Scalar metrics summarize how much of a task each detector solves, whereas relational competence characterizes how successes are organized relative to other detectors. The fixed-margin null preserves detector coverage and instance observability exactly, isolating residual pairwise structure beyond these marginals. Detectors with similar scalar performance can therefore occupy different relational positions. Relational competence complements conventional performance measures by capturing this additional structure.

Reproducibility does not make a relational profile intrinsic. Independent retrainings recover similar relational neighborhoods, yet the measured separation varies with the reference representation. This dependence follows from defining each signature through standardized relations to an ordered comparison population. The reference population is therefore part of the measurement itself. A relational profile represents a reproducible position within a specified comparison system, not an intrinsic fingerprint that should remain invariant across model populations.

Reproducible change does not imply intervention specificity. Controlled degradation produces reproducible population-relative change, but the predefined analysis does not separate same-intervention from cross-intervention alignment. These results distinguish scalar performance, relational organization, retraining reproducibility, and response to intervention as separate levels of evidence. An open question is which aspects of this organization persist as the reference population, task, or evaluation distribution changes.

8 Conclusion

We introduced relational competence as a population-relative measure of how model successes are organized after controlling exactly for detector coverage and instance observability. Across the evaluated detector population, residual relational structure remains after this control and can differ even between models with similar scalar performance. Independent retrainings recover similar relational neighborhoods, while controlled degradations produce reproducible population-relative changes without establishing intervention-specific directions. These findings separate scalar performance, relational organization, retraining reproducibility, and response to intervention as distinct aspects of model behavior. Relational competence therefore complements conventional evaluation by asking not only how often models succeed, but how their successes are organized relative to one another.

AI Assistance Statement

Generative AI tools were used to assist with language editing and polishing of the manuscript and with literature retrieval and discovery of potentially relevant prior work. The authors independently verified the cited literature and reviewed all AI-assisted text for technical accuracy, consistency with the reported evidence, and appropriate claim scope. Generative AI outputs were not treated as sources of scientific evidence. The authors take full responsibility for the final content of the paper.

Reproducibility statement

We provide the information needed to reproduce the main analyses throughout the paper and appendix. Section 3 defines the observability matrix, fixed-margin null, standardized pairwise relations, relational signatures, and distance measure. Sections 4–6 describe the primary evaluation settings and reported analyses. Appendix A provides the detector population, calibration protocol, fixed-margin sampling procedure, and primary bootstrap implementation. Appendix B reports the supporting sensitivity analyses, and Appendix C specifies the degradation settings and directional-comparison construction used in Figure 4. Together, these sections document the measurement procedure, analysis settings, uncertainty estimation, and supporting evaluations used to obtain the reported results.

References

Dara Bahri and Heinrich Jiang. Locally adaptive label smoothing improves predictive churn. In *International Conference on Machine Learning*, pp. 532–542. PMLR, 2021.

Srinadh Bhojanapalli, Kimberly Wilber, Andreas Veit, Ankit Singh Rawat, Seungyeon Kim, Aditya Menon, and Sanjiv Kumar. On the reproducibility of neural network predictions. *arXiv preprint arXiv:2102.03349*, 2021.

Daniel Bolya, Sean Foley, James Hays, and Judy Hoffman. Tide: A general toolbox for identifying object detection errors. In *European Conference on Computer Vision*, pp. 558–573. Springer, 2020.

Corrie J Carstens. Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast curveball algorithm. *Physical Review E*, 91(4):042812, 2015.

Yu Chen, Telmo Silva Filho, Ricardo B Prudencio, Tom Diethe, and Peter Flach. β^3-IRT: A New Item Response Model and its Applications. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1013–1021. PMLR, 2019.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022.

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

Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in neural information processing systems*, 33:13890–13902, 2020b.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Fabian Küppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. Multivariate confidence calibration for object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1322–1330. IEEE, 2020.

Donghwan Lee, Behrad Moniri, Xinmeng Huang, Edgar Dobriban, and Hamed Hassani. Demystifying disagreement-on-the-line in high dimensions. In *International Conference on Machine Learning*, pp. 19053–19093. PMLR, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. Coco-o: A benchmark for object detectors under natural distribution shifts. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6316–6327. IEEE, 2023.

Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Making sense of item response theory in machine learning. In *ECAI 2016*, pp. 1140–1148. IOS Press, 2016.

Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial intelligence*, 271:18–42, 2019.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

Bimsara Pathiraja, Malitha Gunawardhana, and Muhammad Haris Khan. Multiclass confidence and localization calibration for object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19734–19743. IEEE, 2023.

Georg Rasch. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Harshay Shah, Sung Min Park, Andrew Ilyas, and Aleksander Madry. Modeldiff: A framework for comparing learning algorithms. In *International Conference on Machine Learning*, pp. 30646–30688. PMLR, 2023.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Giovanni Strona, Domenico Nappo, Francesco Boccacci, Simone Fattorini, and Jesus San-Miguel-Ayanz. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature communications*, 5(1):4114, 2014.

Johannes Theodoridis, Johannes Maucher, and Andreas Schilling. Differences in detection: Explainability where it matters. *arXiv preprint arXiv:2606.07503*, 2026.

Table 4: **Sensitivity of retraining separation.** Alternative reference representations and evaluation conditions change the magnitude of ΔD. Non-primary entries are descriptive.

Analysis	Condition	ΔD
Reference sensitivity		
Primary	full 8D	0.1424
Anchor exclusion	6D	0.0899
Leave-one-out	range across removals	0.0376–0.1754
	remove Faster R-CNN	0.0376
Evaluation sensitivity		
P80 / 0.50	8D / 6D	0.1424 / 0.0899
P90 / 0.50	8D / 6D	0.1058 / 0.0224
P80 / 0.75	8D / 6D	0.1089 / 0.0151
P90 / 0.75	8D / 6D	0.0843 / -0.0103

A Experimental and Measurement Details

Detector population and evaluation protocol. We evaluate eight object detectors: Faster R-CNN, DINO, RetinaNet, FCOS, ATSS, RTMDet-S, RTMDet-L, and Deformable DETR. All relational quantities are computed from detector outputs on the same evaluation instances. Confidence operating points are selected on a disjoint calibration split and frozen before analysis. The primary setting uses the calibrated P80 operating point at IoU 0.50; the other predefined operating-point and IoU settings are used only for sensitivity analysis. The ordered detector population also defines the reference coordinates of each relational signature.

Observability and relational measurement. For each evaluation instance i and detector m, we construct the binary observability variable Z_{im} defined in Section 3. The detector margins $c_m = \sum_i Z_{im}$ measure detector coverage, while the instance margins $r_i = \sum_m Z_{im}$ measure how broadly each instance is observable across the detector population. For each detector pair (a, b), we compare the observed co-observability O_{ab}^{obs} with its fixed-margin null distribution, which preserves all detector and instance margins exactly. The standardized residual S_{ab} defines the pairwise relation used to construct the population-relative signatures and relational distances in the main analysis.

Fixed-margin null sampling. We generate null matrices with Curveball updates, which randomize the arrangement of successes while preserving every row and column sum. The primary population-level analysis uses $1{,}004$ null realizations. Pairwise empirical p-values use the standard $+1$ correction, and the Benjamini–Hochberg procedure adjusts for multiple pairwise tests. The omnibus $T_{\max}$ and T_{energy} statistics are evaluated against the same null ensemble. Because S_{ab} scales the count residual by its null variability, we interpret it as a standardized departure from the preserved marginals rather than as a count-scale effect size.

Retraining inference. The primary retraining analysis uses three independently trained Faster R-CNN checkpoints and three independently trained RetinaNet checkpoints selected on the calibration split. We estimate uncertainty in the primary separation statistic with an image-cluster bootstrap that resamples evaluation images together with all associated instances. Each of the 100 bootstrap replicates reconstructs the evaluation matrix and recomputes the fixed-margin null using 84 null samples across four chains before recalculating the relational signatures and the within- versus between-design distance contrast. The resulting percentile interval therefore quantifies sampling uncertainty over evaluation images for these six fixed checkpoints, not variability over the broader population of possible training seeds.

B Supporting Sensitivity Analyses

Table 4 summarizes the supporting sensitivity analyses for the retraining separation $\Delta D = D_{\text{between}} - D_{\text{within}}$. The primary inferential result is the P80/IoU 0.50 eight-dimensional analysis reported in the main text; all other entries are descriptive point estimates.

The separation remains positive across all leave-one-reference-out variants and across all four evaluation conditions in the full 8D representation. Its magnitude depends on the reference representation, and the 6D variant decreases to -0.0103 at P90/IoU 0.75. This sensitivity is consistent with the population-relative interpretation of relational competence in the main text.

C Degradation Evaluation Details

Degradation settings. We evaluate two controlled input degradations using the six checkpoints from the retraining analysis. Resolution degradation uses a scale factor of 0.5, and JPEG compression uses quality 30. Each degraded checkpoint is compared with its corresponding clean checkpoint under the same relational measurement procedure.

Population-relative change. For each checkpoint and degradation, we define $\Delta\mathbf{s} = \mathbf{s}_{\text{degraded}} - \mathbf{s}_{\text{clean}}$. Because the clean and degraded signatures are standardized within their respective evaluation matrices, $\Delta\mathbf{s}$ describes change in the population-relative signature representation rather than an intrinsic model-space displacement. Figure 4a visualizes the unit-normalized direction $\widehat{\Delta\mathbf{s}} = \Delta\mathbf{s}/\|\Delta\mathbf{s}\|_2$ across the eight reference coordinates.

Directional comparisons. We compute cosine similarity between the normalized change directions using the predefined same- and cross-intervention comparison sets from Section 6.2. These comparisons define S_{same}, S_{cross}, and $\Delta S = S_{\text{same}} - S_{\text{cross}}$. Figure 4b includes only the validated comparisons used in this analysis. We interpret these quantities as directional agreement within the population-relative representation, not as evidence of an intrinsic change direction or intervention-specific mechanism.