Directional Confusions Reveal Divergent Inductive Biases Through Rate-Distortion Geometry in Human and Machine Vision

Leyla Roksan Caglar
Windreich Department of AI and Human Health
Icahn School of Medicine at Mount Sinai
New York, NY, USA
leylaroksan.caglar@mssm.edu

Pedro A.M. Mediano
Department of Computing
Imperial College London
London, UK
p.mediano@imperial.ac.uk

Baihan Lin
Windreich Department of AI and Human Health
Department of Psychiatry
Icahn School of Medicine at Mount Sinai
New York, NY, USA
baihan.lin@mssm.edu

Abstract

To humans, a robin seems more like a bird than a bird seems like a robin, but does this asymmetry also hold for machine vision? Humans and modern vision models can match each other in accuracy while making systematically different kinds of errors, differing not in how often they fail, but in who gets mistaken for whom. We show that these directional confusions reveal distinct inductive biases invisible to accuracy alone. Using matched human and deep neural network responses on a natural-image categorization task under 12 perturbation types, we quantify asymmetry in confusion matrices and link its organization to the geometry of the information–error trade-off — how efficiently, and how gracefully, a system generalizes under distortion. We find that humans exhibit broad but weak asymmetries across many class pairs, whereas deep vision models show sparser, stronger directional collapses into a few dominant categories. Training on single distortions or noise mixtures brings overall asymmetry into the human range without recovering the human-like distributed structure. Generative simulations further show that these two asymmetry organizations shift the trade-off geometry in opposite directions even at matched accuracy, explaining why the same scalar asymmetry score can reflect fundamentally different generalization strategies. Together, these results establish directional confusion structure as a sensitive, interpretable signature of inductive bias that accuracy-based evaluation cannot recover.

1 Introduction

People judge a robin to be more similar to a bird than a bird is to a robin. An ellipse seems more like a circle than a circle seems like an ellipse [Tversky and Gati, 1982]. These directional asymmetries are not noise, they reflect the graded structure of cognitive representations — prototype effects [Rosch, 1975], feature salience [Tversky, 1977], category typicality [Rosch and Mervis, 1975] — and they violate the assumptions of symmetric similarity spaces [Shepard, 1964, 1987]. Directional confusions in categorization tasks [Nosofsky, 1986, 1991, Getty et al., 1979, Kahana and Sekuler, 2002] can therefore serve as diagnostic signatures of representational bias and efficiency [Sims, 2018, Jakob and Gershman, 2023].

Do artificial vision systems exhibit the same asymmetric structure or a fundamentally different one? Humans and modern ANNs increasingly reach similar categorical accuracy, but most evaluation pipelines treat confusion as unstructured, asking how often a system errs rather than who gets mistaken for whom and in which direction [Gupta et al., 2021, Attarian et al., 2020]. Yet, this kind of

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

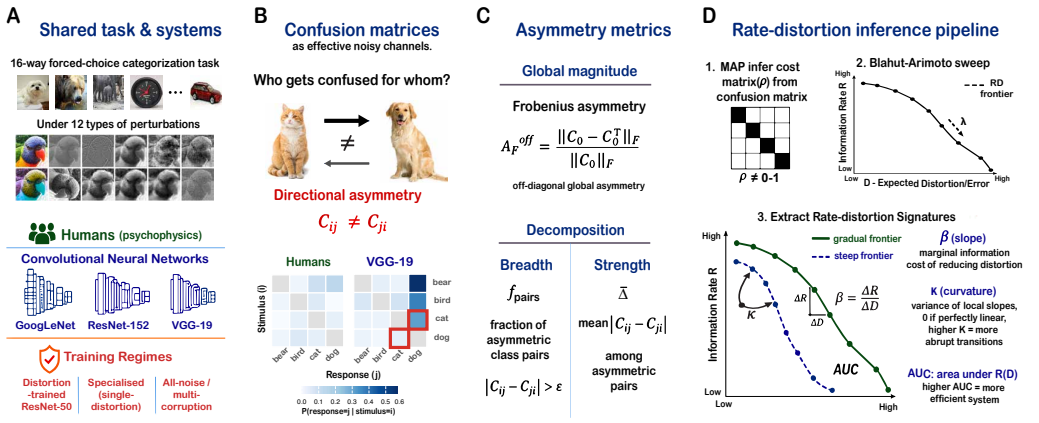

Figure 1: **Analysis pipeline. (A)** Matched human and ANN stimulus–response behavior under perturbation. **(B)** Confusion-matrix construction, illustrating the contrast between human broad–weak errors and ANN sink-like collapses (noise perturbation condition, $\sigma = 0.10$). **(C)** Schematic asymmetry decomposition into breadth and strength, alongside the global Frobenius magnitude. **(D)** Schematic RD frontier inference yielding three geometric signatures (β, κ, AUC).

directional structure is precisely where inductive biases leave their fingerprint, reflecting the priors a system implicitly imposes when mapping ambiguous or degraded inputs to categories.

However, evaluations of ANNs under distribution shift typically abstract away error structure. Metrics such as accuracy or top-k error implicitly treat confusion patterns as symmetric or irrelevant [Geirhos et al., 2020, Recht et al., 2019], even though deep vision models can produce highly structured one-way failures, such as collapsing diverse subcategories into a dominant prototype (e.g., many dog breeds into "Labrador") without the reverse confusion occurring [Shankar et al., 2020, Miller et al., 2021, D'Amato et al., 2025]. Enhanced under adversarial noise or texture-based corruptions, models may confuse texture-diagnostic categories (e.g., "zebra" with "barcode") in one direction only [Geirhos et al., 2018a, Ilyas et al., 2019]. Large-scale evaluations show that this gap between human and machine error consistency persists even in state-of-the-art models [Geirhos et al., 2021], and simplicity bias — the tendency of ANNs to rely on the simplest predictive features available — provides a mechanistic account of why such structured failures arise [Shah et al., 2020]. The most direct precedent, Liu et al. [2025], compares human and machine confusion matrices at the level of aggregate error distributions, but does not characterize the directional organization of asymmetries or link error structure to generalization properties.

This raises a specific question: when vision systems are matched for task, perturbation, and accuracy level, do humans and ANNs exhibit systematically different directional confusion structure, and does that structure predict generalization geometry independently of accuracy? Cognitive accounts explain human asymmetries via graded similarity and attentional biases [Tversky, 1977, Rosch, 1975], predicting broad but weak directional tendencies spread across many class pairs. By contrast, prototype collapse and shortcut-driven feature reliance in ANNs [Shankar et al., 2020, Geirhos et al., 2020, Shah et al., 2020] predict sparser but stronger sink-like failures concentrated on a few classes — representations collapsed onto a small set of dominant decision boundaries rather than the graded similarity gradients that characterize human-like representations [Sorscher et al., 2022, Wei and Woodford, 2025]. These two organizations reflect fundamentally different priors. Distributed asymmetries indicate sensitivity to graded similarity across many feature dimensions, while sink-like collapses indicate rigid reliance on a small set of dominant decision boundaries. Critically, two systems can show similar accuracy (or even similar aggregate asymmetry) while failing for qualitatively different reasons, with different implications for downstream reliability. We therefore not only compare asymmetry structure across systems, but explicitly test, in simulation at matched accuracy, whether the organization of directional errors carries information about generalization that is independent of accuracy.

To formalize this, we adopt a rate–distortion (RD) framework that treats each system as a noisy communication channel and summarizes its generalization behavior as a geometric signature of the information–error trade-off — how efficiently, and how gracefully, a system compresses stimulus–

response information under distortion (see Section 2). Within this formalism, the *organization* of asymmetry, not just its magnitude, is predicted to shape where a system sits on the RD frontier, linking the structure of directional errors directly to generalization geometry. We ask whether humans and ANNs dissociate in asymmetry organization, how that structure relates to generalization geometry independently of accuracy, and whether robustness training closes the gap. We find that (i) humans distribute confusion broadly and weakly across many class pairs while ANNs concentrate failures into a few strong, sink-like collapses, a dissociation invisible to accuracy; (ii) asymmetry covaries with generalization geometry across systems, but scalar-asymmetry couplings to RD signatures are largely accounted for by accuracy, so the same performance can mask different failure regimes; (iii) generative simulations show the two organizations produce opposite effects on generalization geometry even at matched accuracy, explaining why scalar asymmetry metrics are insufficient; and (iv) robustness training reduces overall asymmetry but does not restore the human organization, falling below the human level on breadth, strength and error concentration alike.

2 Rate-Distortion Framework

To formalize how directional error structure relates to generalization, we adopt a rate–distortion (RD) framework grounded in information theory and efficient coding [Shannon, 1948, Sims, 2018]. We treat each system — human or machine — as an effective noisy communication channel whose behavior is summarized by its stimulus–response confusion matrix. The central object is the information–error trade-off: how much mutual information between stimulus and response must be preserved to achieve a given level of categorical distortion. Tracing this trade-off yields a rate–distortion (RD) frontier [Shannon, 1948], and the shape of that frontier characterizes how efficiently and how gracefully a system generalizes under compression [Sims, 2018] (see Fig. 1 for an illustrated inference pipeline).

Concretely, we infer a latent distortion matrix $\rho \in \mathbb{R}_{\geq 0}^{K \times K}$ from each observed confusion matrix via maximum-a-posteriori (MAP) estimation, using Blahut–Arimoto (BA) iterations [Blahut, 1972, Arimoto, 1972] to trace the optimal channel at each compression level. Sweeping an inverse-temperature parameter λ over the resulting frontier yields three compact geometric signatures that are behavioral read-outs of the inferred cost geometry:

- **Slope** (β): the median finite-difference derivative $\Delta R / \Delta D$ along the frontier — how steeply error falls as the system retains more stimulus–response information. Because R decreases with D, β is negative, and a more negative value means a steeper trade-off.
- **Curvature** (κ): the variance of local slopes, capturing how nonuniform the trade-off is across operating points. Low κ corresponds to graceful degradation as information is reduced. High κ to an abrupt, threshold-like collapse.
- **Efficiency** (AUC): the trapezoidal area under $R(D)$ across the swept λ range. A larger AUC means the system retains more stimulus–response information at each level of tolerated error, so class distinctions are preserved rather than lost as error tolerance varies. Because the prior fixes the scale of ρ, the distortion range covered by the fixed λ grid is comparable across systems (Appendix A2.3).

A key property of this framework is that it accommodates asymmetric confusion matrices directly, without forcing behavioral data into a symmetric similarity space. This makes it a natural formalism for the present analysis. The RD frontier is shaped by the full structure of the distortion matrix — including its asymmetric component — so systems with different asymmetry organizations are expected to trace qualitatively different frontiers, even at matched accuracy levels. The organization of directional errors, not just their magnitude, is therefore a first-class input to generalization geometry (Fig. 1, panels B-D). The analysis rests on three assumptions: (i) the behavioral channel is adequately summarized by the confusion matrix; (ii) the cost of confusing two classes is graded rather than binary, and is inferred under a structured, component-wise prior (Appendix A2.3); and (iii) the frontier is generated by efficient coding under that cost, the setting validated on human identification data by Sims [2018]. Recovering ρ from a finite confusion matrix is not guaranteed to be unique, so we do not claim to recover ρ or the full frontier. Rather, the claim is that the three low-dimensional signatures are recoverable, which we test directly in simulations with known ground truth (Section 3.5).

3 Methods and Modeling Framework

3.1 Datasets, Perturbations, and Evaluated Systems

We analyze matched human and ANN model behavior on controlled perturbations of natural images in a $K = 16$ ImageNet-derived categorization setting. The primary stimulus benchmark is the Generalization repository (GEN; [Geirhos et al., 2018b]), which includes twelve perturbation families (e.g., colour/grayscale, contrast, filtering, phase noise, power equalisation, rotation, Eidolon variants, and uniform noise), each parameterized by distortion strength to produce systematic out-of-distribution (OOD) conditions. The data includes ∼83k human psychophysics trials, as well as three baseline pretrained convolutional neural networks (CNNs): GoogLeNet [Szegedy et al., 2015], ResNet-152 [He et al., 2016], and VGG-19 [Simonyan and Zisserman, 2014]. To study training-induced variation, we evaluate models with different robustness regimes. In particular, we include (a) *Distortion-trained* ResNet-50 models trained from scratch with distorted training distributions, (b) *Specialised* single-distortion models trained on one perturbation family and evaluated across all perturbations, and (c) *All-noise / multi-corruption* models trained on mixtures of noise-like corruptions and evaluated across individual perturbations (see the GEN repository [Geirhos et al., 2018b] for further details on benchmark stimuli, model training, and task evaluations).

3.2 Confusion Matrices as Effective Behavioral Channels

We treat each system's confusion matrix as defining an effective noisy channel, then ask what latent distortion structure — and what information–error trade-off — is implied by its pattern of errors, including their directional asymmetries (Fig. 1, panel B). For each system s, experiment e, and distortion level d, we summarize stimulus–response behavior with a $K \times K$ confusion matrix $N^{(s,e,d)}$, where N_{ij} counts responses of class j to stimuli of class i. Row-normalization yields an empirical conditional distribution $C_{ij}^{(s,e,d)} = \Pr_s(y = j \mid x = i; e, d) \approx N_{ij}^{(s,e,d)} / \sum_{j'} N_{ij'}^{(s,e,d)}$, where $C_0^{(s,e,d)}$ denotes $C^{(s,e,d)}$ with its diagonal entries set to zero (i.e., excluding correct responses). Because C is built from hard (argmax) decisions, it is invariant to argmax-preserving recalibration of a model's posterior (e.g., temperature scaling), so calibration is largely orthogonal to the signatures reported here. We infer the latent distortion matrix ρ by maximum-a-posteriori (MAP) estimation, with the likelihood evaluated via the Blahut Arimoto-optimal channel [Blahut, 1972, Arimoto, 1972] under ρ (scale absorbed into ρ), following Sims [2018] and adapted here to directional asymmetry. To trace the rate–distortion (RD) frontier, we scale ρ by an inverse-temperature parameter $\lambda > 0$ and compute the corresponding optimal channel. Specifically, we solve for the RD-optimal channel $q_\lambda(y|x)$ via Blahut–Arimoto fixed-point updates: $q_\lambda(y|x) \propto p(y) \exp(-\lambda \rho(x, y))$, iterated jointly with $p(y) = \sum_x p(x) q_\lambda(y|x)$ until convergence. We then trace the frontier over 50 log-spaced λ values ($\lambda \in [10^{-3}, 10^{3}]$), computing the rate in bits, and extract the three signatures defined in Section 2 (full specification in Appendix A2.3).

3.3 Quantifying Asymmetry

We define directional asymmetry as deviation from matrix symmetry in the row-normalized confusion matrix C, with a threshold $\varepsilon = 10^{-12}$ to suppress numerical noise (Fig. 1, panel C). We use three complementary measures. **Breadth**: $f_{\text{pairs}} = n_{\text{pairs}} / \binom{K}{2}$, where $n_{\text{pairs}} = \sum_{i<j} \mathbb{I}[|C_{ij} - C_{ji}| > \varepsilon]$, captures how many class pairs exhibit any directional asymmetry. **Strength**: $\bar{\Delta} = \mathbb{E}_{i<j}[|C_{ij} - C_{ji}| \mid |C_{ij} - C_{ji}| > \varepsilon]$ captures the mean magnitude of directional deviation among asymmetric pairs, where $\mathbb{E}_{i<j}[\cdot]$ denotes the sample mean over all class pairs $i < j$. **Global magnitude**: quantified by the normalized off-diagonal Frobenius norm $A_F^{\text{off}}(C) = \|C_0 - C_0^\top\|_F / \|C_0\|_F$, which excludes correct responses from both numerator and denominator and is used as the asymmetry predictor in all analyses reported here. Implementation details and the choice of ε are in Appendix A2.5. All analyses use block-wise aggregates (experiment × condition × model instance) to avoid pseudo-replication. We compare groups using Wilcoxon rank-sum tests for robustness to non-normality and Welch's t-tests for effect size estimation, with p-values corrected via Benjamini–Hochberg FDR (BH–FDR) and 95% CIs via the Welch–Satterthwaite approximation.

3.4 Linking Asymmetry to RD Geometry

We test whether directional confusability covaries with RD behavior **(Fig. 1, panel D)**, using $A_F^{\text{off}}(C)$ (Section 3.3) as the asymmetry predictor for each block. Channels with near-deterministic rows are excluded on entropy and response-dominance criteria (Appendix A2.1). Asymmetry–RD relationships

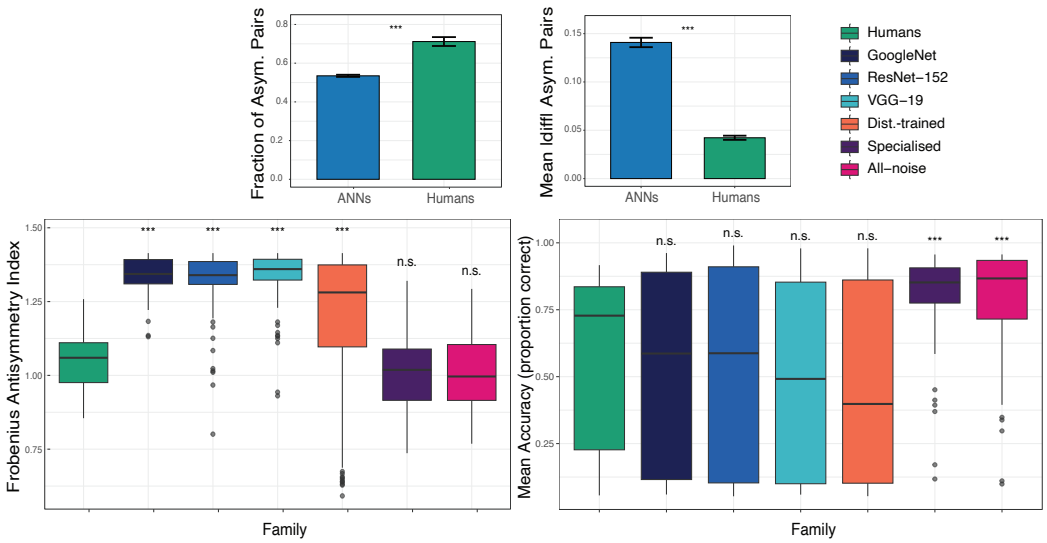

Figure 2: **Asymmetry decomposes into *breadth* vs. *strength*, revealing a dissociation between humans and ANNs invisible to accuracy. Top:** Breadth (f_{pairs}, left) and strength ($\bar{\Delta}$, right) of directional structure, comparing all ANNs pooled against humans. Error bars show s.e.m. across blocks; significance marks two-sided Wilcoxon rank-sum tests. **Bottom:** Frobenius asymmetry index (left) and mean accuracy (right) by model family, each compared individually against humans (planned Wilcoxon tests, BH–FDR corrected). Each point is one block (experiment×condition×model). Groups significantly more asymmetric than humans show no accuracy difference (n.s.), and vice versa, confirming a double dissociation between asymmetry organization and accuracy.

are estimated via rank correlations, block-demeaned interaction models and accuracy-controlled regressions with cluster-robust inference (Appendices A2.2 and A2.4).

3.5 Generative Simulation Linking Asymmetric Inductive Bias to RD Signatures

Our empirical results quantify asymmetry and RD geometry but do not expose the generative mechanisms behind their relationship. We therefore simulate systems with tunable asymmetric distortion structures to test interpretability and recoverability. Specifically, we predict that broad—weak and sink-like asymmetry organizations will produce opposite effects on RD geometry. Distributed asymmetries should expand the RD frontier by preserving information across many class distinctions, while concentrated sink-like asymmetries should collapse it by funneling probability mass into a few dominant responses. This dissociation should persist even after controlling for overall accuracy.

We fix $K = 16$ classes. Each replicate involves a ground-truth distortion matrix ρ_{true}, simulated confusion counts N, inferred distortion $\hat{\rho}$, and derived metrics. We construct $\rho_{\text{true}} = \rho_{\text{sym}} + aA$ where ρ_{sym} is symmetric, A is skew-symmetric ($A = -A^\top$), $\rho_{ii} = 0$, and $a \geq 0$ controls antisymmetry magnitude. We evaluate two antisymmetry types: (i) *broad–weak* (dense skew-symmetric noise), and (ii) *sink-like* (targeted bias toward a small set of classes). Given ρ_{true}, we generate channels via Blahut–Arimoto iterations, draw confusion counts from a multinomial distribution, and recover $\hat{\rho}$ using the same MAP pipeline as for empirical systems. Simulations sweep antisymmetry magnitude a, generation inverse temperature λ_{gen}, and per-class trial count $N_{\text{per row}}$ across both structures with multiple seeds. Full grid details, secondary diagnostics, and sensitivity checks are reported in Appendix A6 (Tables 9–12).

4 Results

4.1 Asymmetry magnitude and sparsity dissociate between humans and ANNs.

We quantified asymmetry using block-wise summaries (one value per unique *experiment*×*condition*×*model*; ANNs: $n = 1569$ blocks; humans: $n = 81$ blocks; representative matrices in Fig. 6). ANNs exhibited larger global asymmetry than humans on the Frobenius index, yet sparser directional structure (all comparisons Wilcoxon rank-sum, $p \leq 6.6 \times 10^{-5}$; block counts and group means in Table 6, per-group test statistics in Table 7, and **Fig. 2**). Humans spread directionality across more class pairs than ANNs (fraction of asymmetric pairs 0.712 vs. 0.535, a gap that widens

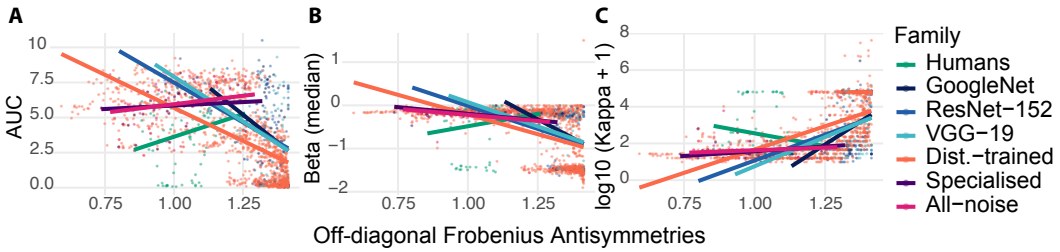

Figure 3: **Directional confusion asymmetry covaries with RD signatures across humans and model families.** These pooled couplings are largely accounted for by accuracy. Each point is one block (experiment×condition×model) in the GEN corpus ($n = 1610$); x-axis is off-diagonal Frobenius asymmetry A_F^{off}. Thick lines show family-wise linear trends. Raw couplings are shown here; the accuracy-controlled versions, and the same analysis in the ModelZoo corpus, are in Fig. 5.**(A)** Efficiency (AUC). Greater asymmetry coincides with reduced RD efficiency in most ANN families, with a distinct human profile. **(B)** Slope (β). Directional confusions track steepness of the information–error trade-off. **(C)** Curvature ($\log_{10}(\kappa + 1)$). Regime-dependent coupling between directional structure and RD nonlinearity. All κ values are strictly positive (range: 8.6–3.5×10^9). Apparent sub-zero region of the y-axis reflects extrapolation of group-wise linear trends beyond the data range.

against baseline CNNs alone), whereas conditional on a pair being asymmetric ANNs showed over three times the per-pair deviation ($\bar{\Delta} = 0.141$ vs. 0.0422). Global asymmetry magnitude therefore conflates two separable properties on which humans and ANNs dissociate in opposite directions: breadth (more pairs in humans) and strength (larger deviations in ANNs; **Fig. 1, panel C**).
Planned humans vs. model comparisons of Frobenius asymmetry (BH–FDR across groups) indicated that baseline CNNs and the Distortion-trained regimes remain significantly more asymmetric than humans, whereas the specialised and all-noise regimes were not significantly different from humans **(Fig. 2)**, suggesting that robustness-oriented training can reduce global asymmetry toward the human range (Table 7). That reduction does not, however, recover the human breadth–strength organization (Section 4.2). Importantly, the dissociation runs in both directions (**Fig. 2**, bottom-right). Groups that are significantly more asymmetric than humans (GoogLeNet, ResNet-152, VGG-19, Distortion trained) show no significant difference from humans in accuracy, whereas groups that match humans on asymmetry (Specialised, All-noise) are significantly more accurate than humans. This double dissociation shows that group differences in asymmetry are not simply group differences in accuracy, and that directional confusion structure carries information not captured by accuracy-based evaluation.
To assess generalizability beyond classic CNNs, we replicated the Frobenius asymmetry analysis on 15 modern architectures including Vision Transformers, large CNNs, and self/semi-supervised models. All families remained significantly more asymmetric than humans (all $p < 0.001$, BH-FDR; **Fig. 8**, see Appendix A4) regardless of accuracy level, confirming the dissociation is not architecture-specific (per-condition matrices; the full breadth–strength analysis on pooled matrices is in Section 4.2).

4.2 Robustness of the dissociation

Robustness training lowers strength without restoring breadth. The two regimes that match humans on global asymmetry reach that match by a different route. In the GEN corpus the specialised and all-noise regimes reach conditional strengths of 0.030 and 0.025, at or below the human value of 0.042, but their breadth reaches only 0.503 and 0.583 against 0.712 for humans — closer to the baseline CNNs (0.446–0.457). Robustness training therefore produces a third organization, narrow and weak, rather than the broad and weak human one, and it is this combination that brings the scalar Frobenius index into the human range (Table 7). Matching on a one-number asymmetry score does not imply matching on its organization.

ANN errors funnel into several shifting receivers, not a single sink. A natural alternative is that ANN sparsity reflects a single "sink" category. It does not: error mass above chance falls on 4.95 classes on average (humans: 6.07), and the dominant receiver shifts across perturbations (modal-receiver consistency 0.21–0.46 across models, versus 0.73 for humans), while ANNs nonetheless concentrate more mass into their most-confused class (top-1 excess over a density-matched null: 0.281 vs. 0.086; Appendix A2.7). The organization is therefore a set of strong, shifting directional collapses rather than a fixed bin, and it persists under accuracy control and in robustness-training.

Concentrated error structure persists under density-matched comparison. To test whether ANN error concentration is a property of their errors rather than of the $K = 16$ benchmark, we evaluated three architectures (ResNet-50, ViT-B/16, and a self-supervised ConvNeXt-V2) on the full ImageNet-1k validation set under three ImageNet-C corruption families at two severities — a $60\times$ larger label space, on different images and categories (also see Appendix A2.8 for robustness to how the original 16 labels are grouped). Against references generated from our own broad–weak and sink cost geometries at matched per-class error and trial counts, real ANN errors concentrate onto $\sim$4–24 effective target classes versus $\sim$46 for the density-matched diffuse reference, in all 11 conditions with adequate error density (Table 5, Fig. 10). The structure is concentrated but not degenerate, falling well short of the 1–2 effective targets of the extreme sink reference (Appendix A5).

Modern architectures do not close the gap. Per-condition ModelZoo matrices are too sparse for the breadth–strength decomposition, so we pooled each perturbation family across its severity levels and repeated the analysis on 15 modern systems. Every architecture family was significantly less broad, higher in strength, more Frobenius-asymmetric and sparser than humans, with all 24 family$\times$metric comparisons surviving BH–FDR correction. Vision Transformers and semi-supervised models are the least broad families (0.469 and 0.451). Critically, the systems this benchmark ranks as most human-like on error consistency (CLIP, BiT-M, ViT-L, SWAG, ResNeXt-SWSL, EffNet-L2) are not human-like in asymmetry organization: breadth 0.495 vs. 0.923 and Frobenius 1.302 vs. 0.882 against humans (both $p_{\mathrm{BH}} \leq 10^{-7}$). Directional structure therefore carries inductive-bias information that accuracy and error-consistency benchmarks do not (Table 4, Fig. 9).

The human profile holds at the individual-observer level. Every individual observer's mean strength (0.024–0.043) falls entirely below the baseline CNN and distortion-trained range in the same corpus (0.121–0.154), so the low-strength half of the human profile is not an artifact of pooling observers. Individual breadth is lower than the pooled value, but a trial-count control indicates this is a density effect rather than a strategy difference (Appendix A2.6). Cross-observer correlations of per-experiment breadth are uniformly high (0.90–0.99), with no evidence of distinct strategy subgroups.

4.3 Asymmetry–RD coupling and its dependence on accuracy.

Asymmetry organization related systematically to RD trade-off geometry, but the strength and direction of coupling depended on model family and training regime. Because naive correlations partly reflect shared dependence on accuracy, we report both pooled associations and accuracy-controlled within-block regressions (specification in Appendix A2.2; cluster-robust inference and collapsed-channel exclusion in Appendix A2.4). In the empirical data, pooled asymmetry–RD associations are largely accounted for by accuracy, signature by signature below. Because power differs across groups ($n = 1182$ Distortion-trained blocks versus $n \leq 81$ elsewhere), a non-significant residual is not evidence of absence, and the accuracy-independent evidence therefore rests mainly on the generative simulations (Section 3.5), which isolate asymmetry organization at matched accuracy.

Efficiency (AUC). Greater asymmetry was associated with lower RD efficiency across most ANN families, but this coupling was largely accounted for by shared dependence on accuracy — with one exception **(Fig. 3A)**. Pooled rank correlations were strongly negative for Distortion-trained models ($\rho = -0.73$, $n = 1182$) and negative for all three baseline CNNs ($\rho = -0.39$ to -0.43), and block-demeaned interaction models showed substantially steeper within-block dependence for the Distortion-trained regime than for humans ($\Delta\mathrm{slope} = -7.89$, $p = 1.27 \times 10^{-7}$), with the same direction for Specialised and All-noise (Table 8). After accuracy control, however, only the Distortion-trained regime retains a residual association, and its sign reverses: the conditional association is *positive* (slope $= 1.02 \pm 0.30$, $t = 3.45$, $p_{\mathrm{BH}} = 0.010$), while humans and baseline models show no reliable accuracy-controlled slopes (all $p_{\mathrm{BH}} \geq 0.147$). This is the only association significant under every inference scheme and λ grid we tested. The negative pooled associations are therefore largely an accuracy effect, and the same pattern holds in both corpora (Fig. 5).

Slope (β). A similar pattern held for RD slope: asymmetry and β were negatively correlated across most regimes, (Distortion-trained $\rho = -0.57$; baseline CNNs -0.33 to -0.46), with steeper within-block dependence for the robustness-trained regimes than for humans (Distortion-trained $\Delta\mathrm{slope} = -1.21$, $p = 0.0085$), but these associations were largely accounted for by accuracy (Table 8; **Fig. 3B**). Under accuracy control, Distortion-trained models no longer showed a reliable asymmetry–β relationship (slope $= 0.07$, $p_{\mathrm{BH}} = 0.77$), and baseline CNNs showed none (all $p_{\mathrm{BH}} \geq 0.36$). Specialised models showed a negative association (slope $= -0.79 \pm 0.26$, $t = -3.08$,

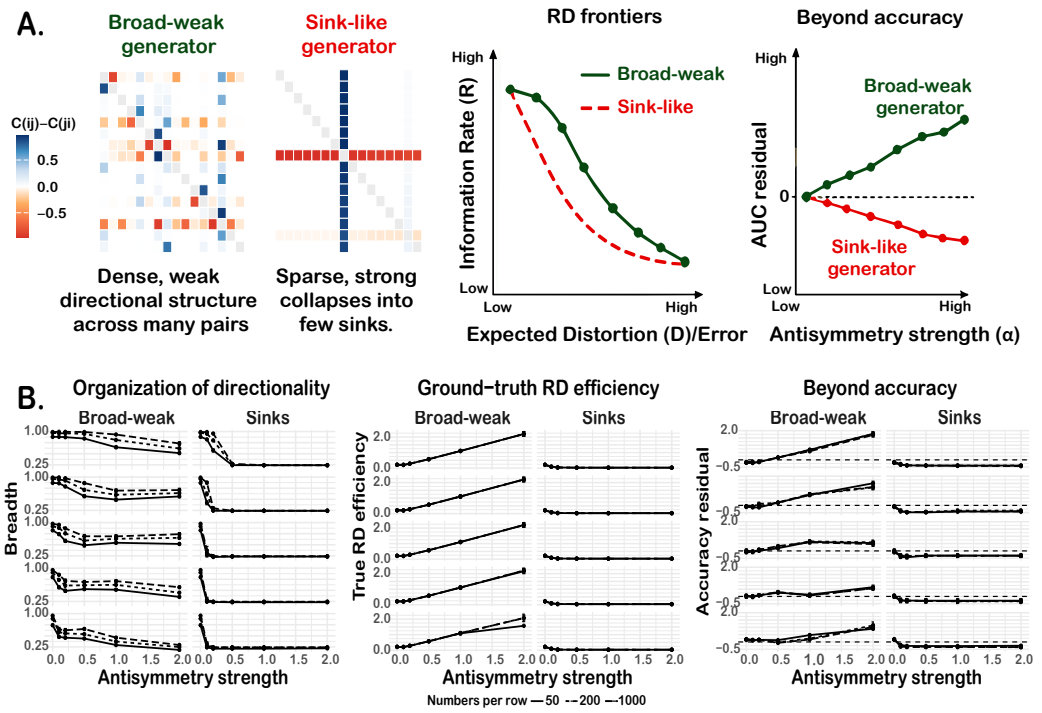

Figure 4: **Generative simulation: asymmetry organization controls RD geometry. A.** Left: simulated asymmetry matrices from one representative replicate per generator ($a = 1.0$, $\lambda_{\text{gen}} = 2.0$,$N = 200$). Right: schematic illustrations of the predicted RD frontier divergence and accuracy-residualized AUC, consistent with the simulation results in panel B. **B.** Full simulation grid. Columns compare generators; rows show five generalization regimes ($\lambda_{\text{gen}} \in \{0.2, 0.5, 1, 2, 5\}$, top to bottom).

$p_{\text{BH}} = 0.037$), but this is the one result in our analysis that is not robust to the λ grid: under a narrower grid it attenuates to approximately -0.13 and is no longer significant (Appendix A2.3). We therefore do not interpret it as an accuracy-independent effect.

Curvature (κ). Curvature showed the most accuracy-dependent pattern of the three signatures. The Distortion-trained regime showed a strong positive rank association ($\rho = 0.72$) and significantly steeper within-block dependence than humans ($\Delta\text{slope} = 3.96$, $p = 1.97 \times 10^{-5}$), as did All-noise and Specialised (Table 8), despite Specialised exhibiting an opposite pooled rank tendency **(Fig. 3C)**. None survived accuracy control: accuracy was strongly predictive of curvature (block-demeaned accuracy term $A_{\text{dm}} = -4.32 \pm 0.41$, $t = -10.63$, $p = 4.9 \times 10^{-11}$), whereas no group retained a significant accuracy-controlled asymmetry–curvature association (all $p_{\text{BH}} \geq 0.098$; Table 1), consistent with curvature tracking performance more than directional structure.

4.4 Generative Simulation

We implemented the mechanistic simulation to ask the following targeted question: *When directional confusions increase, does the underlying rate–distortion (RD) geometry expand in the same way for different forms of asymmetry?* We compared two generators matched on the same control parameters (generalization regime and sample size) but differing in how directionality is organized **(Fig. 4A)**. There was one *broad–weak* mechanism that distributes weak one-way biases across many class pairs, versus a second *sink-like* mechanism that concentrates probability mass into a small set of strong one-way errors.

Directional structure is recoverable but identifiability is mechanism-dependent. Across the full simulation grid ($n = 1800$ runs), numerical collapse was rare ($138/1800 = 7.7\%$), leaving $n = 1662$ non-collapsed runs for primary analyses. A stricter reliability screen, requiring that the recovered symmetric component of ρ aligns with ground truth (correlation > 0.2), removed an additional $509/1662 = 30.6\%$, yielding $n = 1153$ strictly-recovered runs (Table 9). Recovery was strongly *mechanism-dependent*, with broad–weak structure exceeding sinks in pass rate in

every $(\lambda_{\text{gen}}, N_{\text{per row}})$ slice and FDR significance in $10/15$ slices (BH–FDR; Table 9). The largest gaps occurred at moderate-to-high generalization sharpness and intermediate-to-large sample sizes (maximum pass-rate gap 0.417; e.g. 0.867 vs 0.450), indicating that sink-like directional structure is intrinsically harder to identify under the same MAP pipeline. Consequently, we report all main trends on the non-collapsed set and use strict-recovery only as a sensitivity check (Tables 9 and 12). This differential recoverability implies that our MAP pipeline may underestimate the prevalence of sink-like structure in empirical data. Results involving sink-like regimes should therefore be interpreted with this asymmetry in mind.

Broad–weak versus sinks produce opposite couplings between antisymmetry and RD geometry. Consistent with our prediction, the same increase in antisymmetry strength produces opposite changes in RD geometry depending on whether directionality is distributed broadly or concentrated into sinks. In the broad–weak generator, increasing antisymmetry systematically *expands* ground-truth RD geometry, whereas in the sink-like generator it *collapses* ground-truth RD geometry toward a near-degenerate regime. This qualitative dissociation is visible in the true RD efficiency curves across regimes and remains consistent under strict-recovery filtering **(Fig. 4; Table 12)**. This provides a concrete generative explanation for the empirical dissociation in which humans and ANNs can exhibit comparable global asymmetry magnitude yet occupy different regions of RD space.

The RD–asymmetry coupling is not reducible to overall performance. A natural concern is that RD geometry might simply track overall performance. However, removing accuracy within each $(\lambda_{\text{gen}}, N_{\text{per row}})$ slice, the dissociation persisted: broad–weak generators show strongly positive accuracy-adjusted AUC slopes across all 15 slices (all $p_{\text{FDR}} \leq 7.9 \times 10^{-31}$), while sink slopes are negative in all 15 slices (Fig. 4; Appendix A6).

The breadth–strength decomposition is predictive. To connect mechanism to observable summaries, we decomposed directional structure into *breadth* (how many class pairs exhibit directionality) and *strength* (how large the directional deviation is among asymmetric pairs). After residualizing outcomes for accuracy *within* $(\lambda_{\text{gen}}, N_{\text{per row}})$ slices, we asked which aspect of asymmetry explains residual variation in RD signatures. For residual RD *efficiency* (AUC) the component model is strong and slice-consistent: breadth and strength are both typically negative (significant in $11/15$ and $12/15$ slices, BH–FDR) and their interaction typically positive ($5/15$). Residual AUC therefore depends on how directionality is distributed — broad-and-weak versus sparse-and-strong mixtures — and not merely on how much of it there is. A one-number global magnitude model is by contrast nearly null for AUC residuals, reaching significance in only $3/15$ slices. Global A_F^{off} thus obscures the organization that AUC is sensitive to, whereas slope and curvature track global magnitude in simulation (coefficients, ranges and per-slice counts in Table 10).

4.5 Ablations: symmetrizing a confusion matrix is not a neutral operation.

Confusion and similarity matrices are routinely symmetrized before analysis. We asked how much of the inferred RD geometry depends on the directional information this discards, applying two ablations — symmetrization, and a within-row shuffle that scrambles directional targets while holding accuracy fixed — to matched matrices from humans, a baseline CNN (ResNet-152) and a specialised model (Appendix A3). The dependence is substantial, and its size orders the three systems exactly as their error concentration does (Fig. 7). Symmetrization raised $\log_{10}\kappa$ in $\sim$90% of matrices and lowered AUC in every system (Appendix A3). Fragility, the per-matrix increase in $\log_{10}\kappa$, gives specialised (0.16) $<$ humans (0.24) $<$ ResNet-152 (0.59), with all pairwise Wilcoxon comparisons significant ($p = 1.7 \times 10^{-6}$ to 0.029) — the same ordering as top-1 error concentration (0.148, 0.178, 0.412), and as the disagreement between the two arbitrary mirror directions (0.07, 0.16, 0.22), which measures the directional information a symmetry assumption discards. The shuffle sharpens the split by moving ResNet-152's curvature by $+0.35$ ($p = 1.1 \times 10^{-5}$) while changing humans and the specialised model only slightly (-0.06 in both). This is not a magnitude effect, as change in A_F^{off} explains under 2% of the variance in curvature change across all 405 ablated matrices. Removing directional structure therefore perturbs frontier curvature most where directional mass is most concentrated, so symmetrization is not a neutral preprocessing step.

5 Discussion

Human and ANN vision systems can achieve similar categorization accuracy while relying on fundamentally different inductive biases, differing not in how often they err, but in how failures are organized. The breadth–strength dissociation we document reflects distinct priors about which

features and prototypes are privileged under uncertainty. Humans distribute errors broadly, consistent with graded similarity across many feature dimensions, while ANNs concentrate failures into a small number of dominant collapse directions. Crucially, this difference is invisible to global asymmetry magnitude. Robustness-oriented training can match scalar asymmetry metrics without producing the human organization. Because it suppresses directionality across the board (i.e., below the human level) two systems can appear aligned on a scalar metric while arriving there by entirely different routes. Treating confusion matrices as effective noisy channels and linking their directional structure to RD frontier geometry suggests why: in our simulations, breadth–strength organization specifically predicts how efficiently a system compresses stimulus–response information (AUC). Slope (β) and curvature (κ) by contrast primarily track the global magnitude of asymmetry rather than its organizational structure (Section A6.2). The same global asymmetry score can thus index fundamentally different generalization regimes depending on whether errors are distributed or concentrated — a distinction that accuracy-based evaluation cannot recover.

At the representational level, the breadth–strength dissociation we observe behaviorally would be consistent with differences in the geometry of learned feature spaces. Broad–weak asymmetries with manifolds in which categories lie along graded similarity gradients, so that many class pairs are weakly but meaningfully separated, and sink-like asymmetries with representations that collapse many inputs onto a few dominant attractor states. Establishing this link empirically is an important open direction. RD efficiency (AUC) has been shown to track penultimate-layer representational structure beyond accuracy [Caglar et al., 2026], suggesting that the organization documented here may be probed without direct access to activations.

This has a direct methodological implication. Symmetric similarity and dissimilarity structure is assumed by construction across much of the field — in representational similarity analysis, in multidimensional scaling, and wherever a confusion matrix is averaged with its transpose before analysis — and the cognitive literature has long noted that human similarity judgments violate it [Tversky, 1977, Shepard, 1964]. Our ablations show that this assumption is not a neutral preprocessing step for the quantities studied here. Instead, symmetrizing a confusion matrix measurably changes the shape of its inferred RD frontier, and does so unequally across systems, especially for those whose errors are most concentrated. A symmetric analysis therefore discards information that is both real and differentially distributed, which can make systems with very different directional organization appear more alike than they are. Curvature therefore indexes how gracefully a system degrades, and fragility to symmetrization tracks error concentration rather than error rate.

We distinguish two kinds of conclusion. The accuracy-independent conclusions are (i) the mechanistic dissociation between asymmetry organization and RD geometry, established in simulation at matched accuracy, and (ii) the link between AUC and internal representational geometry reported in prior work [Caglar et al., 2026]. The conclusions that depend on accuracy are the empirical couplings of scalar asymmetry to κ and, in large part, to β and AUC. These are largely accounted for by accuracy, so naive correlations can partially reflect shared dependence on performance rather than structure alone. Specifically, the generative simulations are what separate organization from magnitude, formalizing a link between the organization of directional confusions and the capacity–generalization trade-offs that shape behavior. The same increase in directional asymmetry can produce opposite RD behaviors depending on how it is organized. In our simulations, broad–weak asymmetries shift the RD frontier toward higher efficiency, whereas sink-like asymmetries shift it toward lower efficiency, and these effects persist under accuracy control.

On a practical level, matching human accuracy or aggregate asymmetry is not sufficient for a human-like error organization, as systems can concentrate failures into sink-like collapses while appearing aligned on scalar metrics. Breadth and strength can serve directly as an evaluation metric computable from any confusion matrix, but using them as a training objective would additionally require a differentiable surrogate, since confusion counts rest on argmax decisions.

Limitations and future directions. We note three scope conditions. First, no matched human psychophysics data exists for label spaces larger than $K = 16$, so the human–model contrast is untested beyond this granularity, and at large label spaces the full directional decomposition remains under-determined at achievable error densities (Appendix A5). Second, human estimates rest on $n = 81$ blocks against $n = 1569$ ANN blocks, and individual-observer breadth is trial-limited (Appendix A2.6). Third, the MAP inference of ρ is validated against ground truth only in simulation, where recovery is mechanism-dependent: sink-like structure is harder to identify than broad–weak structure, so our pipeline may understate its prevalence. Prior-rate sensitivity is bounded only indirectly (Appendix A2.3). Natural extensions include class-level asymmetry analysis, adversarially-trained and human-similarity-supervised models, and other task domains.

Acknowledgments and Disclosure of Funding

This work was supported in part through the Minerva computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences. B.L. is supported by the Brain & Behavior Research Foundation, the Sidney R. Baer, Jr. Foundation, the U.S. National Institutes of Health (NIH), the U.S. Department of Veterans Affairs (VA), the Hasso Plattner Foundation, and the Windreich Family Foundation. The authors declare no competing interests.

References

Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

Maria Attarian, Brett D Roads, and Michael C Mozer. Transforming neural network visual representations to predict human judgments of similarity. *arXiv preprint arXiv:2010.06512*, 2020.

R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972. doi: 10.1109/TIT.1972.1054855.

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

Leyla Roksan Caglar, Pedro A. M. Mediano, and Baihan Lin. Rate-distortion signatures of generalization and information trade-offs, 2026. URL `https://arxiv.org/abs/2603.01568`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Leo D'Amato, Gian Luca Lancia, and Giovanni Pezzulo. The geometry of efficient codes: How rate-distortion trade-offs distort the latent representations of generative models. *PLOS Computational Biology*, 21(5):e1012952, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL `https://arxiv.org/abs/2010.11929`.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018a.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018b.

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

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

David J Getty, John A Swets, Joel B Swets, and David M Green. On the prediction of confusion matrices from similarity judgments. *Perception & Psychophysics*, 26(1):1–19, 1979.

Shashi Kant Gupta, Mengmi Zhang, Chia-Chien Wu, Jeremy Wolfe, and Gabriel Kreiman. Visual search asymmetry: Deep nets and humans share similar inherent biases. *Advances in neural information processing systems*, 34:6946–6959, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Anthony MV Jakob and Samuel J Gershman. Rate-distortion theory of neural coding and its implications for working memory. *Elife*, 12:e79450, 2023.

Michael J Kahana and Robert Sekuler. Recognizing spatial patterns: A noisy exemplar approach. *Vision research*, 42(18):2177–2192, 2002.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507. Springer, 2020.

Minghao Liu, Jiaheng Wei, Yang Liu, and James Davis. Human and AI perceptual differences in image classification errors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14318–14326, 2025.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pages 7721–7735. PMLR, 2021.

Robert M Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1):39, 1986.

Robert M Nosofsky. Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, 23(1):94–140, 1991.

James E Pustejovsky and Elizabeth Tipton. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4):672–683, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

Eleanor Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192–233, 1975.

Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605, 1975.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on ImageNet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR, 2020.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 (3):379–423, 1948.

Roger N Shepard. Attention and the metric structure of the stimulus space. *Journal of mathematical psychology*, 1(1):54–87, 1964.

Roger N Shepard. Toward a universal law of generalization for psychological science. *Science*, 237 (4820):1317–1323, 1987.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Chris R Sims. Efficient coding explains the universal law of generalization in human perception. *Science*, 360(6389):652–656, 2018.

Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens Van Der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 804–814, 2022.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43): e2200800119, 2022.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

Amos Tversky and Itamar Gati. Similarity, separability, and the triangle inequality. *Psychological review*, 89(2):123, 1982.

Xue-Xin Wei and Michael Woodford. Representational geometry explains puzzling error distributions in behavioral tasks. *Proceedings of the National Academy of Sciences*, 122(4):e2407540122, 2025.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

A1 Compute Resources and Data Availability

Computational resources. The pipeline has three steps with distinct cost profiles. Step 1 — MAP inference of the latent distortion matrix ρ — is the computational bottleneck, scaling as $O(K^2)$ in both parameter space and the inner Blahut–Arimoto loop. For $K = 16$ categories, MAP inference takes approximately 50 minutes per confusion matrix on 2 CPU cores; the full empirical analysis covers 1,650 confusion matrices (all models and conditions in the GEN benchmark repository), corresponding to approximately 1,375 core-hours in total. Steps 2 and 3 — RD frontier tracing and signature extraction (β, κ, AUC) — run in seconds given ρ and add negligible overhead. The generative simulation grid (1,800 runs across two asymmetry structures) requires approximately 10–20 CPU-hours. All jobs were parallelized across an HPC cluster, substantially reducing wall-clock time. The ImageNet-1k evaluation (Appendix A5) is the only component requiring GPU resources: three pretrained models were evaluated on the 50,000-image ImageNet validation set under six corruption conditions each, for 18 inference runs totalling 900,000 forward passes and approximately 1 GPU-hour in total on a single GPU node of the same HPC cluster. No model training was performed in this work.

Data and model licenses. The GEN benchmark [Geirhos et al., 2018b] is available at `https://github.com/rgeirhos/generalisation-humans-DNNs` under CC BY 4.0. The ModelZoo benchmark [Geirhos et al., 2021] is available at `https://github.com/bethgelab/model-vs-human`. Its code is released under the MIT License, its image datasets derive from ImageNet and remain subject to the ImageNet license, and the human responses it distributes were collected as reported in Geirhos et al. [2018b]. All pre-trained models (GoogLeNet, ResNet-152, VGG-19, the robustness-trained variants, and the modern architectures of Appendix A4) were accessed through ModelZoo and remain under their original licenses, listed in that repository's `MODEL_LICENSES` file. The SWSL ResNeXt-101 weights in particular are released under CC BY-NC 4.0. MAP inference follows the pipeline of Sims [2018], available at `https://osf.io/x5ckn/`. The ImageNet-1k evaluation additionally uses the ImageNet validation set under the ImageNet terms of access, the ImageNet-C corruption benchmark [Hendrycks and Dietterich, 2019] under CC BY 4.0, and pretrained weights obtained through `timm` under the Apache License 2.0. No new models are introduced in this work and no new human data were collected.

Code and data availability. All analysis code is available at `https://github.com/leylaroksan/rdt_asymmetry`, including the rate–distortion inference pipeline, the asymmetry metrics, the directional-structure ablations, the generative simulation, and scripts reproducing every figure and table in this paper. The MAP inference and Blahut–Arimoto optimization were originally developed by Sims [2018]. The RD signature extraction follows the implementation in Caglar et al. [2026], available at `https://github.com/leylaroksan/rdt_geometry`. Derived data sufficient to reproduce every figure and table without re-running MAP inference — fitted distortion matrices, RD signature summaries, collapse flags, the simulation grid outputs, and the $K = 1000$ confusion matrices — are deposited at `https://doi.org/10.6084/m9.figshare.34059993`. The underlying behavioral and model data remain available from the GEN and ModelZoo repositories.

A2 Statistical and Analytical Procedures

A2.1 Collapsed Channel Filtering

We exclude confusion matrices with collapsed responses based on diagnostics computed from the row-normalized conditional distribution $C(y|x)$ (i.e., each row of the confusion matrix normalized to sum to 1):

- Mean row entropy $< 10^{-3}$, or
- Mean row-maximum probability > 0.999.

Only a small subset of system–condition matrices were excluded, but the exclusion is consequential for β, the signature most sensitive to near-degenerate channels (the affected estimates are reported in Appendix A2.4).

A2.2 Asymmetry–RD Regression Models

We used two linear modeling approaches to assess how asymmetry covaries with RD signatures. First, models with fixed effects per (experiment, condition) block and group-specific slopes. Second, within-block demeaning of predictor and outcome variables, followed by regression with interaction terms. Concretely, asymmetry–RD relationships are estimated via:

- *(i) Rank correlations:* Spearman correlations within group between $A_{\mathrm{F}}^{\mathrm{off}}$ and each RD signature.
- *(ii) Block-controlled models:* Linear models with experiment/condition fixed effects and group-specific slopes.
- *(iii) Accuracy-controlled models:* Because asymmetry and RD geometry both covary with overall accuracy, we additionally test whether asymmetry–RD associations persist after controlling for accuracy within matched (experiment, condition) blocks, isolating the contribution of directional structure independent of performance level.

A2.3 MAP Prior and λ Grid

The latent distortion matrix ρ is inferred by maximum-a-posteriori estimation under a structured, component-wise prior. We decompose ρ into its symmetric off-diagonal part, its diagonal, and its asymmetric part, and place an independent exponential prior on each, with rates 0.1, 10 and 10 respectively (equivalently, prior means 10, 0.1 and 0.1). The prior is deliberately asymmetric across components: the large prior mean on the symmetric off-diagonal component reflects that confusions between distinct classes are costly, while the small means on the diagonal and asymmetric components shrink toward zero self-distortion and toward symmetry, so that directional structure must be supported by the data rather than assumed by the prior. These shared hyperparameters place all inferred ρ matrices on a common scale, so that slope and distortion range are comparable across systems without rescaling ρ after fitting.

Optimization is carried out on the log scale, and the posterior is approximated locally by a Laplace expansion around the MAP estimate using the numerical Hessian; the resulting mean and covariance are retained with each fit. After optimization the minimum entry of ρ is subtracted to enforce non-negativity ($\rho \leftarrow \rho - \min(\rho)$), which translates but does not rescale the cost matrix. The RD frontier is then traced by scaling ρ over 50 log-spaced values of the inverse temperature $\lambda \in [10^{-3}, 10^{3}]$ (Section 3.2), computing the rate $R_k = I(X;Y)$ in bits and the expected distortion $D_k = \mathbb{E}[\rho(X,Y)]$ under an empirical class prior, which is uniform for the balanced GEN design. Local slopes s_k are finite differences taken after sorting by D and removing duplicate D values, with $\beta \equiv \mathrm{median}_k\, s_k$ and $\kappa \equiv \mathrm{Var}_k(s_k)$.

The adequacy of this specification is bounded by the recoverability analysis in simulation (Appendix A6), which compares the recovered symmetric and antisymmetric components of ρ against known ground truth — the antisymmetric component is recovered only under broad–weak structure — and reports strict-recovery pass rates across generation regimes. Sensitivity to the λ grid was assessed by re-deriving all three signatures over four alternative grids (ranges $[10^{-2}, 10^{2}]$ and $[10^{-4}, 10^{4}]$; 25 and 100 grid points) on a stratified subsample of 196 blocks. AUC is essentially invariant: its rank ordering across blocks is identical to the canonical grid in every case (Spearman $\rho = 1.000$) and group means shift by less than 4%. β and κ are rank-stable but range-dependent in magnitude (Spearman $\rho = 0.72$–0.96 and 0.92–0.98 respectively), with κ varying by up to four orders of magnitude across grid ranges. This is expected rather than anomalous: β is the median and κ the variance of local slopes along the traced segment, so both are defined relative to the swept range, whereas AUC integrates the full curve. Comparisons involving β and κ magnitudes should therefore be read as conditional on the stated grid. All comparisons reported are made at the single canonical grid and are internally consistent. What does not transfer is the absolute scale of β and κ, which should not be compared against values computed under a different grid specification.

Even though the prior rates themselves were not varied by re-fitting, the λ sweep bounds the dominant part of this sensitivity. Blahut–Arimoto depends on ρ only through $\exp(-\lambda\rho)$, so rescaling $\hat{\rho}$ is exactly equivalent to shifting the λ range. Because the prior is common to every fit, any such rescaling applies identically to all blocks and leaves the rank- and slope-based between-block comparisons we report unchanged, consistent with the invariance observed above. What remains untested is the *relative* weighting of the three prior components, which alters the shape of $\hat{\rho}$ rather than its scale.

A2.4 Inference and Collapsed-Channel Exclusion

Within-block demeaning removes block-level means, but rows within a block remain non-independent, because a block contains many model instances evaluated on the same stimuli, and the Distortion-trained regime alone contributes roughly ten rows per block. Ordinary least squares treats these as independent observations and understates the standard errors for the groups with many instances per block, while the pooled residual variance – dominated by the largest group – overstates them for the smallest. We therefore report cluster-robust standard errors with (experiment, condition) blocks as clusters, using the bias-reduced CR2 estimator with Satterthwaite degrees of freedom [Pustejovsky

Table 1: **Accuracy-controlled asymmetry–RD slopes (GEN).** Per-group within-block slopes of each RD signature on off-diagonal Frobenius asymmetry, controlling for accuracy within (experiment, condition) blocks ($n = 1610$ blocks after collapsed-channel exclusion; 115 demeaning blocks). Slopes are identical under both inference schemes; only the standard errors differ. OLS treats blocks' rows as independent; CR2 is the bias-reduced cluster-robust estimator with blocks as clusters and Satterthwaite degrees of freedom. p_{BH} is corrected across the seven groups within each signature. Conclusions in the main text follow the CR2 column. Of the two associations significant under CR2, only the Distortion-trained AUC coupling is also robust to the λ grid; the Specialised β association is not (Appendix A2.3).

		OLS		Cluster-robust (CR2)		
Group	Slope	SE	p_{BH}	SE	df	p_{BH}
Efficiency (AUC)						
Humans	-0.886	0.479	0.225	0.478	22.4	0.147
GoogLeNet	0.722	0.770	0.418	0.558	20.2	0.292
ResNet-152	-0.458	0.498	0.418	0.368	7.6	0.292
VGG-19	0.087	0.642	0.892	0.532	8.5	0.874
Distortion-trained	**1.024**	0.117	4.2×10^{-17}	0.297	37.9	**0.010**
Specialised	0.585	0.501	0.418	0.324	22.9	0.147
All-noise	0.694	0.468	0.323	0.324	16.1	0.147
Slope (β)						
Humans	-0.495	0.371	0.640	0.239	22.4	0.154
GoogLeNet	0.141	0.598	0.814	0.465	20.2	0.765
ResNet-152	0.123	0.386	0.814	0.133	7.6	0.535
VGG-19	-0.334	0.498	0.704	0.244	8.5	0.360
Distortion-trained	0.067	0.091	0.704	0.188	37.9	0.765
Specialised	-0.785	0.389	0.306	0.255	22.9	0.037
All-noise	-0.347	0.363	0.704	0.176	16.1	0.154
Curvature ($\log_{10}(\kappa{+}1)$)						
Humans	0.833	0.614	0.614	0.382	22.4	0.132
GoogLeNet	0.588	0.989	0.674	0.674	20.2	0.496
ResNet-152	-0.278	0.639	0.674	0.241	7.6	0.496
VGG-19	0.346	0.824	0.464	0.413	8.5	0.496
Distortion-trained	0.107	0.150	0.674	0.287	37.9	0.712
Specialised	1.322	0.643	0.279	0.497	22.9	0.098
All-noise	0.627	0.600	0.674	0.305	16.1	0.132

and Tipton, 2018], and draw the conclusions in the main text from these. OLS values are given alongside in Tables 1 and 2 for comparison. The two schemes agree on every qualitative conclusion except the Specialised β association, which survives correction under CR2 ($p_{\mathrm{BH}} = 0.037$) but not under OLS ($p_{\mathrm{BH}} = 0.31$), but that association is in turn not robust to the λ grid (Appendix A2.3), so we do not rely on it. Satterthwaite degrees of freedom range from 7.6 to 39.2, so inference is weakest for the groups with fewest independent blocks (ResNet-152, VGG-19) regardless of scheme. Figure 5 shows the same relationships graphically for both corpora.

Collapsed channels are excluded as described in Appendix A2.1, computed from the row-normalized conditional distribution of each stored confusion matrix: 40 of 1650 GEN blocks are excluded (Distortion-trained 3.11%, ResNet-152 and VGG-19 1.23% each, all other groups 0%), leaving $n = 1610$. This exclusion matters most for β, which is the signature most sensitive to near-degenerate channels. The excluded blocks occur at near-chance distortion levels, where a system's responses collapse onto a single class. In 24 of the 40 this produces a numerically non-monotone frontier segment and a positive β (Appendix A6.1). Retaining them changes the Distortion-trained β slope from 0.07 to 0.46

The two corpora are fit separately. Demeaning operates within (experiment, condition), so fitting GEN and ModelZoo jointly pools their rows into shared blocks and alters the GEN estimates. The GEN analysis reported here is the one that uses GEN blocks alone. The distinction is consequential. Fitted separately, local models (BagNet-9/33) retain a significant accuracy-controlled association between asymmetry and RD efficiency (slope $+2.53$, $p_{\mathrm{BH}} = 0.002$ under CR2; Table 2), which the joint fit does not recover. No other modern family retains an association on any signature.

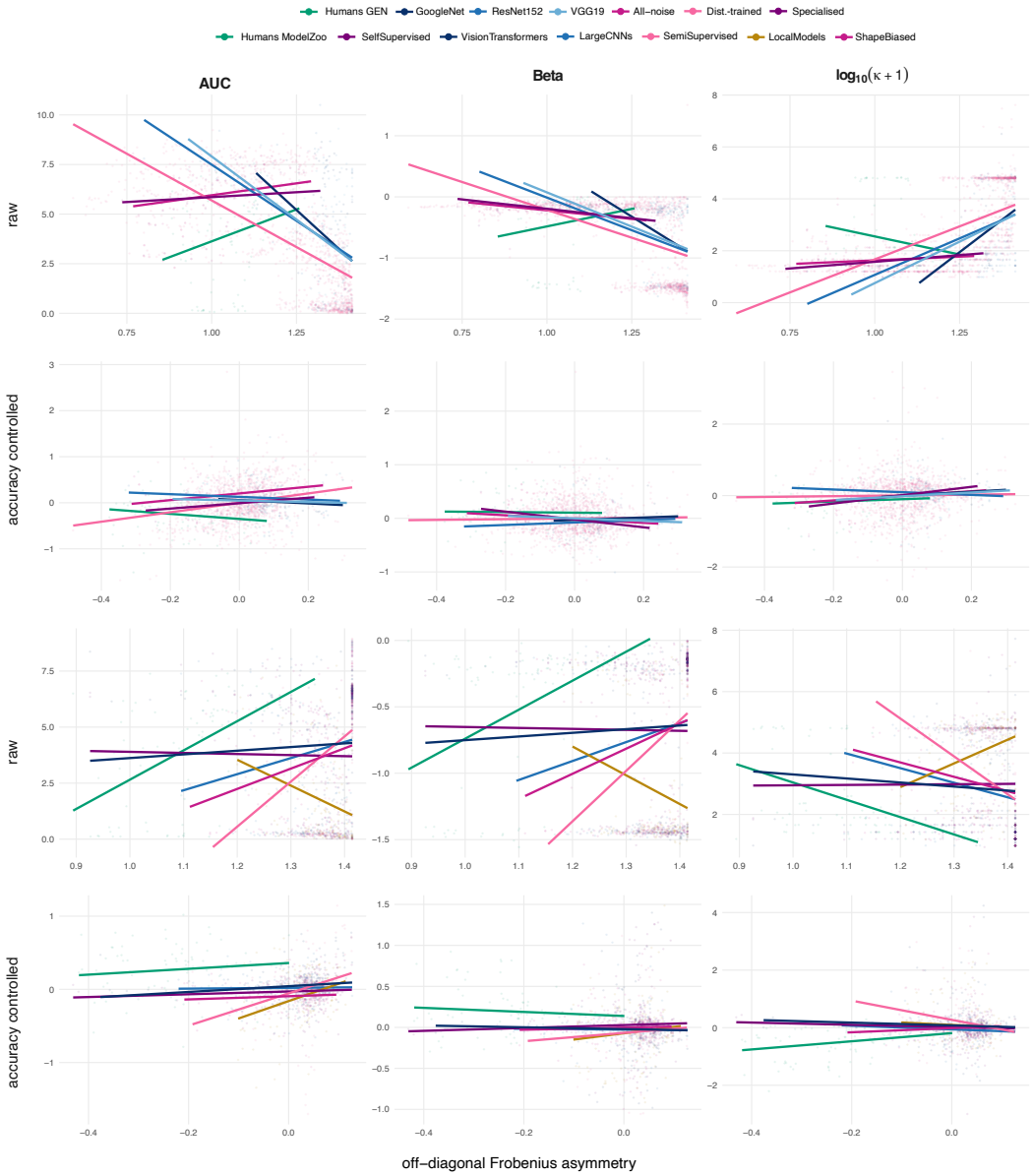

Figure 5: **Asymmetry–RD coupling before and after accuracy control, in both corpora.** Each point is one block (experiment×condition×model); lines are per-group linear fits. The x-axis is off-diagonal Frobenius asymmetry A_F^{off}. **Rows 1–2:** GEN benchmark ($n = 1610$ blocks across 115 experiment×condition cells); the *Raw* row reproduces Fig. 3 with the *Accuracy-controlled* row added. **Rows 3–4:** ModelZoo architectures ($n = 1135$ blocks across 73 cells). In the *Raw* rows both axes are untransformed. In the *Accuracy-controlled* rows both axes are demeaned within (experiment, condition) and residualized on accuracy, so the fitted slope in each panel is the slope reported in Tables 1 and 2. Panel scales are free, since the raw and residualized ranges differ. Before accuracy control, greater asymmetry coincides with lower efficiency, steeper slope and higher curvature in most model families, but not in humans or in the robustness-trained regimes, whose couplings run the other way (Table 8A). After accuracy control most couplings flatten, and those that remain are positive for AUC: the Distortion-trained regime in GEN and the local-receptive-field models in ModelZoo. The two corpora are fit separately, because demeaning within (experiment, condition) across a pooled corpus alters the GEN estimates.

Table 2: **Accuracy-controlled asymmetry–RD slopes (ModelZoo).** As Table 1, fit separately on the ModelZoo corpus ($n = 1135$ blocks, 73 demeaning blocks) with ModelZoo humans as reference. The corpora are fit separately because within-block demeaning across a pooled corpus alters the GEN estimates.

		OLS		Cluster-robust (CR2)		
Group	Slope	SE	p_{BH}	SE	df	p_{BH}
Efficiency (AUC)						
Humans (ModelZoo)	0.370	0.331	0.461	0.518	27.4	0.673
Vision Transformers	0.400	0.247	0.245	0.298	12.2	0.409
Large CNNs (BiT-M)	0.057	0.333	0.865	0.255	27.6	0.825
Semi-supervised	2.075	0.457	2.2×10^{-5}	0.981	14.2	0.184
Self-supervised	0.201	0.228	0.528	0.161	13.4	0.409
Shape-biased (SIN)	0.114	0.337	0.858	0.272	28.7	0.792
Local models	**2.525**	0.414	1.0×10^{-8}	0.635	39.2	**0.002**
Slope (β)						
Humans (ModelZoo)	−0.286	0.378	0.937	0.369	27.4	0.780
Vision Transformers	−0.083	0.282	0.937	0.185	12.2	0.903
Large CNNs (BiT-M)	−0.049	0.381	0.937	0.317	27.6	0.903
Semi-supervised	0.456	0.523	0.937	0.580	14.2	0.780
Self-supervised	0.154	0.260	0.937	0.126	13.4	0.780
Shape-biased (SIN)	0.031	0.386	0.937	0.248	28.7	0.903
Local models	0.627	0.473	0.937	0.554	39.2	0.780
Curvature ($\log_{10}(\kappa+1)$)						
Humans (ModelZoo)	1.518	0.743	0.145	0.684	27.4	0.206
Vision Transformers	−0.521	0.554	0.394	0.406	12.2	0.307
Large CNNs (BiT-M)	−0.637	0.748	0.394	0.502	27.6	0.307
Semi-supervised	−3.177	1.028	0.014	1.591	14.2	0.206
Self-supervised	−0.505	0.511	0.394	0.275	13.4	0.206
Shape-biased (SIN)	0.660	0.758	0.394	0.595	28.7	0.307
Local models	−0.827	0.930	0.394	0.799	39.2	0.307

A2.5 Asymmetry Metric Implementation

Breadth, strength, and the Frobenius index are computed on the row-normalized confusion matrix C (Section 3.3). A class pair $i < j$ is counted as asymmetric when $|C_{ij} - C_{ji}| > \varepsilon$, with $\varepsilon = 10^{-12}$ for all empirical analyses and $\varepsilon = 10^{-6}$ for the simulation summaries (Appendix A6.2). Because f_{pairs} counts pairs exceeding ε, it increases with the number of trials per cell, and the subsampling control in Appendix A2.6 bounds that contribution. A_F^{off} carries no threshold and is the asymmetry predictor in every RD analysis reported here, so none of the asymmetry–RD results inherits this sensitivity.

A2.6 Per-Observer Analysis

To test whether the human profile is an artifact of aggregating observers, we analyzed per-condition individual-observer confusion matrices from the 5–6 observers with adequate coverage in the GEN human data, computed over the 4–6 perturbation families each observer completed. The strength half of the dissociation is robustly individual: every observer's mean strength (0.024–0.043) falls entirely below the baseline CNN and distortion-trained range in the same corpus (0.121–0.154), and every observer's Frobenius asymmetry sits at the low, human end. The specialised and all-noise regimes are the exceptions (0.030 and 0.025), and they are also the two groups statistically indistinguishable from humans on global asymmetry (Table 7). Individual breadth is lower than the pooled human value. A subsampling control indicates this is a trial-count effect, as thinning the pooled human matrix to a single observer's trial count reduces its breadth to the individual level (median 0.33 vs. 0.32 at matched per-condition density). Aggregate breadth should therefore be read as a population-level property, with the per-observer signature carried mainly by low strength. We find no evidence of distinct strategy subgroups, since cross-observer correlations of per-experiment breadth are uniformly high (0.90–0.99), and inter-observer breadth SD within an experiment ($\approx$0.07) is small relative to the GEN human–model breadth gap (0.712 vs. 0.535; Table 6).

Sparsity does not separate the two populations at per-condition GEN density. Individual observers occupy 0.46–0.56 in fraction of empty off-diagonal cells, against 0.438 for the pooled human matrix

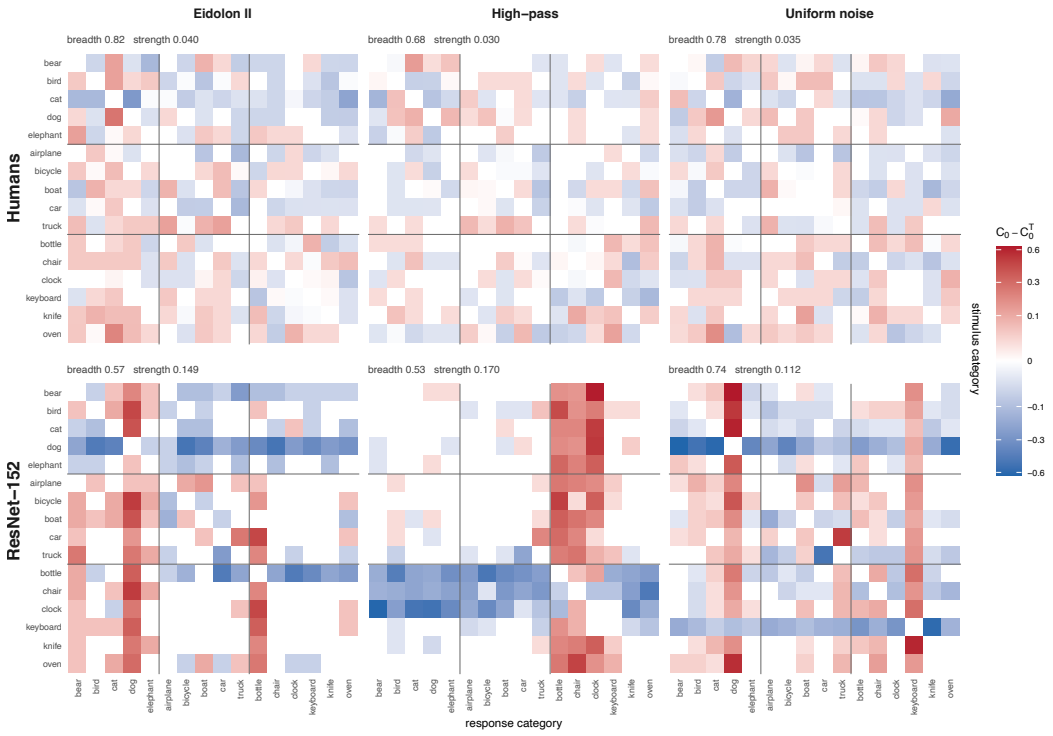

Figure 6: **Representative directional confusion maps.** Off-diagonal asymmetry maps $C_0 - C_0^{\mathsf{T}}$ for humans (top) and ResNet-152 (bottom) under three perturbation families, where C_0 is the row-normalized confusion matrix with the diagonal removed. Cell (i, j) is positive (red) when category i is misread as j more often than the reverse, and negative (blue) when i receives more than it sends, so a predominantly blue row marks a class acting as an error sink. Classes are ordered by superordinate category (animals, vehicles, objects; grey lines). Breadth and strength are printed above each panel, computed as in Appendix A2.5. The color scale is signed-square-root so that the weak human structure stays visible alongside the strong ANN structure. Humans show pale asymmetries distributed across nearly all pairs. ResNet-152 resolves into a few strong receiving columns whose identity changes across conditions — *dog* under Eidolon II, *bottle*/*chair*/*clock* under high-pass, *dog* and *keyboard* under uniform noise — the per-matrix form of the modal-receiver instability reported in Section 4.2. Because the two rows are matched on stimuli and conditions, the human panels serve directly as the diffuse reference for the concentrated ANN panels.

and 0.54–0.71 for the GEN ANN groups. We therefore do not interpret individual-level sparsity as human-like, and only strength and Frobenius asymmetry are cleanly individual.

A2.7 Error-Mass Concentration and Receiver Stability

If ANN sparsity reflected a single sink category, the number of classes receiving above-chance error mass would be close to one and the identity of the dominant receiver would be constant across perturbation conditions. Neither holds. Averaged over models, error mass above chance falls on 4.95 classes (humans: 6.07), and each model has 7–18 distinct top receivers across conditions, with modal-receiver consistency of 0.21–0.46 (humans: 0.73). Concentration is nonetheless higher than in humans: the single largest receiving class absorbs an excess of 0.281 over a density-matched null (humans: 0.086), and the top three classes absorb 0.63 of all error mass (humans: 0.41). Ranked by regime, top-1 concentration is 0.145 (all-noise), 0.148 (specialised), 0.178 (humans), 0.378 (Distortion-trained) and 0.412–0.446 (baseline CNNs).

A2.8 Label-Grouping Robustness

We recomputed all asymmetry metrics under two relabelings of the same trials. The semantic grouping collapses the 16 GEN categories into 8 superordinate categories, but the direction and significance of all three effects are preserved (breadth: humans 0.92 vs. ANNs 0.79, $p = 6 \times 10^{-13}$; strength: 0.048 vs. 0.16, $p = 7 \times 10^{-5}$; Frobenius: 0.89 vs. 1.15, $p = 5 \times 10^{-25}$). Because any single semantic grouping is one particular choice of partition, we additionally drew 50 random partitions of the 16 classes into 8 groups and recomputed the metrics for each; humans showed higher breadth and lower strength than ANNs in 50 of 50. Both analyses re-partition the same images, trials, responses and models, so they test robustness to how labels are grouped and not to the size of the label space, which is addressed in Appendix A5. Breadth and strength both increase under coarsening relative to their $K = 16$ values (Table 6), so what is preserved is the direction of the human–ANN difference rather than the absolute magnitudes.

A3 Directional-Structure Ablations

Design. To quantify how much the RD signatures depend on directional structure, we apply two ablations to a matched subset of GEN matrices — humans, a baseline CNN (ResNet-152), and a specialised training-regime model, across four perturbation experiments (high-pass, low-pass, phase scrambling, rotation), 27 matrices each. Each ablated matrice was passed through the same MAP and frontier pipeline used for the originals. *Symmetrization* mirrors one triangle of C onto the other, the operation implicit whenever behavioral data are forced into a symmetric similarity space. Because the choice of triangle (upper vs. lower) is arbitrary, we apply both directions (mirror-lower and mirror-upper) and treat their disagreement as a diagnostic. The *within-row off-diagonal shuffle* permutes the off-diagonal entries within each row, scrambling which classes receive a system's errors while leaving the diagonal — and therefore accuracy and each class's total error mass — unchanged. This is applied with three independent seeds per matrix. The five ablated variants per original matrix — two mirror directions and three shuffle seeds — give $81 \times 5 = 405$ ablated matrices in total. Comparisons are paired within matrix.

Both manipulations act on integer counts, so neither is as clean as its name suggests. Mirroring symmetrizes the counts but leaves row totals unequal, so the row-normalized channel retains part of its asymmetry: 49% of A_F^{off} is removed for ResNet-152, against 73% for humans and 82% for the specialised model. The shuffle is close to magnitude-preserving for concentrated systems, where little symmetric structure remains for a permutation to destroy (ResNet-152: A_F^{off} $+1.9\%$ on average), but raises magnitude for broader ones (humans $+10.5\%$, specialised $+17.6\%$).

Signature normalization. The ablation signatures are computed on min–max normalized frontiers, with R and D each rescaled to $[0, 1]$ within each curve, rather than on the raw frontiers used elsewhere in the paper. Symmetrization alters both the scale of the inferred frontier and its shape. Normalizing within curve removes the scale change, so the reported $\Delta \log_{10}(\kappa{+}1)$ isolates the change in trade-off shape, which is the quantity the ablation is about. Signature values in this section are therefore not comparable to those in Table 1 or Fig. 3. Computed on raw frontiers the same comparisons are not significant ($p = 0.16$–0.50). As expected, at raw scale, κ spans nine orders of magnitude, so a change in frontier shape produces a negligible change in $\log_{10} \kappa$.

Control: the effect is not driven by asymmetry magnitude. Across all 405 ablated matrices, the change in A_F^{off} accounts for under 2% of the variance in $\Delta \log_{10}(\kappa{+}1)$ (adjusted $R^2 = 0.015$), and

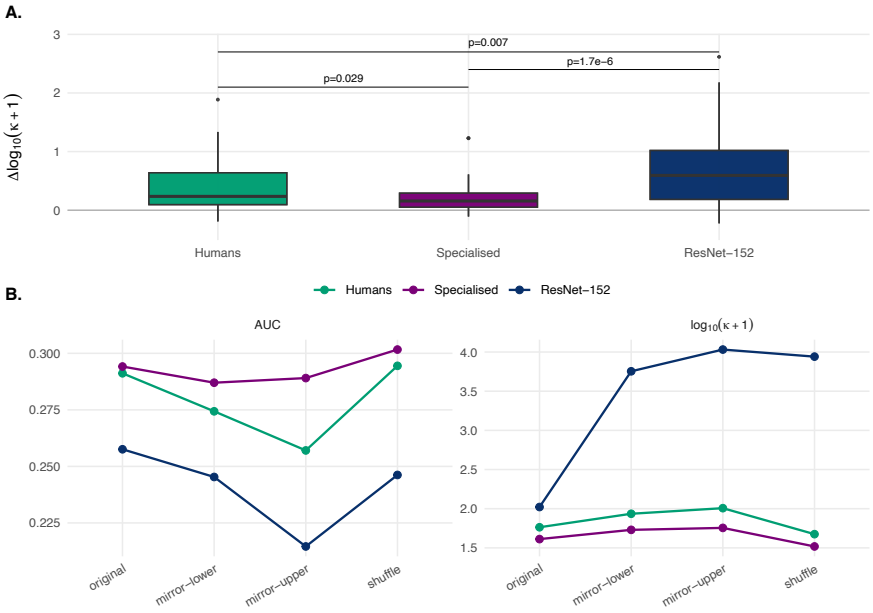

Figure 7: **Removing directional structure reshapes RD geometry most for the system with the most concentrated error mass.** Humans, a specialised-regime model and a baseline CNN (ResNet-152) under the original matrices and three ablations: mirroring the lower triangle onto the upper (*mirror-lower*), the reverse (*mirror-upper*), and a within-row off-diagonal shuffle that scrambles the directional target of errors while holding the diagonal — and hence accuracy and per-class error magnitude — fixed (*shuffle*). **(A)** Curvature fragility: the per-matrix increase in $\log_{10}(\kappa{+}1)$ under symmetrization, with two-sided Wilcoxon comparisons. The ordering is specialised $<$ humans $<$ ResNet-152. **(B)** Median signature value per system at each manipulation, for efficiency (AUC, left) and curvature ($\log_{10}\kappa$, right). The two panels report different statistics: (A) is the median of per-matrix *differences*, whereas (B) plots the median signature at each condition, so the vertical rise visible in (B) is a difference of medians and exceeds the median difference in (A) under the skewed κ distribution. All points aggregate four perturbation experiments (high-pass, low-pass, phase scrambling, rotation; 27 matrices each). The baseline CNN separates from humans and the specialised model under every ablation, and the two arbitrary mirror directions disagree with each other most for it: symmetrization is least well-defined for the systems carrying the most directional structure.

the ordering runs against a magnitude account: the shuffle changes ResNet-152's magnitude least and its curvature most, while symmetrization removes the smallest share of its asymmetry and produces its largest curvature shift. Because the 27 matrices per system span multiple distortion levels within four experiments, they are not independent samples. The reported Wilcoxon statistics are paired comparisons across matched matrices and should be read as descriptive effect sizes rather than tests over independent observations.

Results. Symmetrization raises curvature in close to every matrix (fraction with positive $\Delta\log_{10}(\kappa{+}1)$: 0.91, 0.89, 0.93), and the size of the rise orders the three systems as their error concentration does: specialised 0.16, humans 0.24, ResNet-152 0.59 (medians; all pairwise Wilcoxon $p = 1.7 \times 10^{-6}$ to 0.029). The disagreement between the two arbitrary mirror directions — the directional information a symmetry assumption discards — follows the same ordering (0.07, 0.16, 0.22). However, the shuffle separates the systems more sharply still. It changes curvature only slightly for humans and the specialised model (-0.06 in both) while shifting ResNet-152 by $+0.35$ ($p = 1.1 \times 10^{-5}$). Efficiency moves in the opposite pattern — symmetrization costs humans the most AUC (-6.9 to -10.3%) and ResNet-152 the least (-1.1 to -6.4%). Full values are in Table 3.

Table 3: **Directional-structure ablations.** Effects of removing directional information from GEN confusion matrices, for humans, a specialised single-distortion model, and a baseline CNN, across four perturbation experiments (27 matrices per system). Rows are ordered by fragility. **(A)** $\log_{10}\kappa$ by variant, median [IQR] over matrices. **(B)** Ablation effects. Inflation is the per-matrix increase in $\log_{10}\kappa$ under symmetrization, pooled over both mirror directions ($n = 54$); "frac. pos." is the fraction of matrices in which it increases; "mirror disagr." is the median absolute difference in $\log_{10}\kappa$ between the two arbitrary symmetrizations. Shuffle columns give the paired change relative to the original matrix, with Wilcoxon p-values. Changes in β are consistent in direction (steeper under symmetrization in every system) but are not reported numerically, because β is a median of local slopes that can approach zero, making relative changes unstable.

(A) $\log_{10}\kappa$ by variant

System	Original	Mirror-lower	Mirror-upper	Shuffle
Specialised	1.61 [1.50–1.83]	1.73 [1.69–1.94]	1.75 [1.69–1.95]	1.52 [1.46–1.72]
Humans	1.76 [1.60–2.26]	1.94 [1.73–2.84]	2.01 [1.73–3.24]	1.67 [1.51–2.21]
ResNet-152	2.02 [1.77–3.94]	3.75 [1.87–4.66]	4.03 [2.07–4.54]	3.94 [1.80–4.66]

(B) Ablation effects

	Symmetrization				Shuffle	
System	κ inflation median (IQR)	frac. pos.	mirror disagr.	AUC change	$\Delta\log_{10}\kappa$ (p)	AUC change (p)
Specialised	0.16 (0.24)	0.91	0.07	-2.8 to -4.6%	-0.06 (0.0055)	$+0.5\%$ (0.20)
Humans	0.24 (0.54)	0.89	0.16	-6.9 to -10.3%	-0.06 (0.90)	$+3.1\%$ (0.017)
ResNet-152	**0.59** (0.83)	0.93	**0.22**	-1.1 to -6.4%	$+\mathbf{0.35}$ (1.1×10^{-5})	$+3.0\%$ (0.059)

A4 Extension to Modern Architectures

A4.1 Model selection and data.

The main analyses use matched human psychophysics and ANN behavior from the GEN benchmark repository [Geirhos et al., 2018b], which provides ∼83k human trials across 12 perturbation families at graded distortion levels for $K = 16$ ImageNet-derived categories. Architecture choice in the main analysis was constrained by the requirement for matched human psychophysics data at sufficient trial density. Only models for which per-cell confusion counts are dense enough to reliably estimate directional asymmetry at the class-pair level are included. This limits the main analysis to three baseline CNNs (GoogLeNet, ResNet-152, VGG-19) and three robustness training regimes (Distortion-trained, Specialised, All-noise).

To assess whether the global asymmetry dissociation extends to modern architectures not included in the GEN repository, we additionally analyzed a broad selection of 15 architectures spanning six architecture families:

- **Large-scale supervised models**: BiT models (bitm-resnetv2-152x2, bitm-resnetv2-152x4 [Kolesnikov et al., 2020])
- **Transformers and vision–language models**: Vision Transformer ViT Large Patch16-224 [Dosovitskiy et al., 2021], SWAG-ViT-L/16-IN1K [Singh et al., 2022], CLIP Radford et al. [2021]
- **Architectural inductive-bias probes (LocalModels)**: BagNet-9 and BagNet-33 [Brendel and Bethge, 2019]
- **Robustness/shape-biased models**: standard ResNet-50 CNNs trained on Stylized ImageNet (ResNet50 SIN) and a mixture of Stylized and regular ImageNet ResNet50 (SIN+IN) [Geirhos et al., 2018a]
- **Data rich semi-supervised models**: semi-weakly supervised pretrained ResNeXt-101 32x16d SWSL [Mahajan et al., 2018], Noisy Student pretraining (efficientnet-l2-noisy-student-475 Xie et al. [2020])
- **Self-supervised / contrastive models**: Simple Framework for Contrastive Learning models (SimCLR resnet50x1, SimCLR resnet50x4 Chen et al. [2020a]), MoCo v2 Chen et al. [2020b], InfoMin Tian et al. [2020]

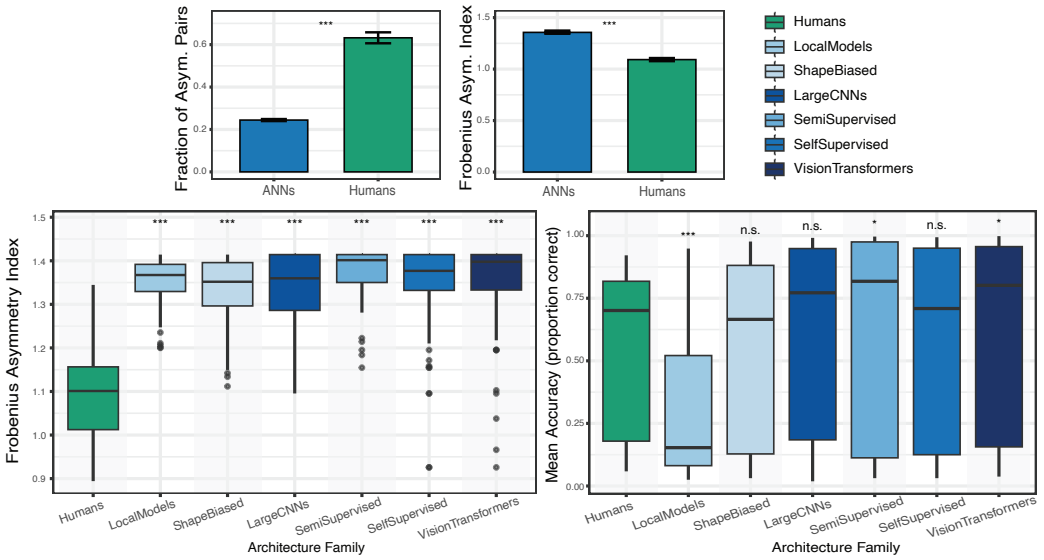

Figure 8: **Modern architectures are more asymmetric than humans at every accuracy level. Top row:** Fraction of asymmetric class pairs (left) and Frobenius asymmetry index (right) comparing humans against all 15 modern model architectures. Error bars show s.e.m. and significance marks correspond to two-sided Wilcoxon rank-sum tests on block-wise summaries. **Bottom row:** Frobenius asymmetry index (left) and mean classification accuracy (right) by architecture family. Each point summarizes one block (unique experiment×condition×model). Boxes show the distribution across blocks. Significance marks are based on planned Wilcoxon comparisons of each family against humans with BH–FDR correction.

Model outputs were obtained from the ModelZoo repository [Geirhos et al., 2021] using the same stimulus set, enabling direct comparison of asymmetry structure under identical perturbation conditions, with a second set of human data (∼85k psychophysics trials).

A4.2 Analytical approach and sparsity caveat

Asymmetry metrics are computed identically to the main analysis (Section 3.3), using block-wise summaries (one value per unique experiment×condition×model) on row-normalized confusion matrices. Group comparisons use two-sided Wilcoxon rank-sum tests with BH–FDR correction across planned comparisons against humans.

An important limitation of the ModelZoo repository data is that confusion matrices are sparser than in the GEN repository, since these models were not evaluated with matched per-stimulus trial counts, resulting in fewer observations per confusion matrix cell. The GEN repository confusion matrices contain on average 85 non-zero asymmetric class pairs out of 120 possible for humans ($K = 16$; 71%) and 64 for ANNs (54%), providing sufficient density for the breadth–strength decomposition. By contrast, the ModelZoo ANN confusion matrices contain on average only 29 pairs (24%), less than half the GEN ANN coverage despite identical $K = 16$ categories and stimulus conditions. ModelZoo human data achieves comparable coverage to the GEN benchmark (76 pairs; 63%) confirming that the drop in pair coverage is specific to the ModelZoo ANN confusion matrices. This sparsity simultaneously deflates estimates of asymmetry breadth (f_{pairs}), because many off-diagonal cells are exactly zero, and inflates conditional magnitude ($\bar{\Delta}$), because only the largest asymmetries survive the sparsity filter. These two metrics should therefore be interpreted with caution for the per-condition ModelZoo matrices. The Frobenius asymmetry index, which aggregates over all off-diagonal entries and is substantially less sensitive to the proportion of zero cells, is robust to this issue and is therefore the primary metric reported in the per-condition analysis below. Pooling each perturbation family across its severity levels raises density enough for the full breadth–strength decomposition (see Appendix A4.4 for this analysis).

A4.3 Results

All six ModelZoo architecture families (15 architectures) remain significantly more asymmetric than humans on the Frobenius index (all $p < 0.001$, BH–FDR corrected; **Appendix Fig. 8**, bottom-left),

Table 4: **Modern architectures on pooled ModelZoo matrices.** Mean breadth (f_{pairs}), strength ($\bar{\Delta}$), off-diagonal Frobenius asymmetry (A_F^{off}), fraction of empty off-diagonal cells, and accuracy by architecture family, with humans as the reference row. The metrics are defined in Appendix A2.5. Matrices are pooled per perturbation family across severity levels, giving $n = 12$ blocks per family, and collapsed channels are excluded. Values are therefore not comparable to the per-condition analyses in Table 6 and Fig. 8. Pooled accuracies in particular are trial-weighted and run roughly 0.07 above the per-condition means in that figure. Block-level distributions and significance against humans are in Fig. 9.

Family	Breadth	Strength	Frobenius	Off-diag. zeros	Acc.
Humans	0.923	0.022	0.882	0.140	0.62
Vision Transformers	0.469	0.059	1.325	0.691	0.71
Semi-supervised	0.451	0.060	1.333	0.716	0.70
Self-supervised	0.548	0.064	1.297	0.634	0.65
Large CNNs (BiT-M)	0.578	0.050	1.238	0.588	0.67
Shape-biased (SIN)	0.648	0.055	1.223	0.531	0.62
Local models (BagNet)	0.720	0.090	1.283	0.447	0.39

replicating the global asymmetry dissociation observed in the main analysis across a substantially broader range of architectures. This result holds regardless of training paradigm, including self-supervised (SimCLR, MoCoV2, InfoMin), semi-supervised (ResNeXt101-SWSL, EffNet-L2), shape-biased (ResNet50-SIN), and large-scale supervised models (BiT-M). Critically, it also holds for Vision Transformers and CLIP, which represent a qualitatively different architectural class from the CNNs evaluated in the main analysis. CLIP and SWAG-ViT-L/16 are trained with language or hashtag supervision rather than ImageNet labels alone, yet both fall within the transformer family's asymmetry profile — one of the two least broad families — giving no indication that multimodal supervision by itself reshapes directional organization.

The accuracy profile across architecture families is heterogeneous (**Appendix Fig. 8**, bottom-right). Local Models (BagNet-9, BagNet-33) are significantly less accurate than humans, while Vision Transformers and semi-supervised models are significantly more accurate. The remaining families do not differ from humans. Yet all families are significantly more asymmetric than humans, confirming a double dissociation between asymmetry structure and accuracy that extends beyond the main analysis to modern architectures.

The main analyses in Sections 4.1 and 4.3 use the GEN data because it provides directly matched human psychophysics under identical stimulus conditions at per-condition density. The ModelZoo systems are analyzed separately in Appendix A4.4, where pooling across severity levels supports the breadth–strength decomposition. The accuracy-controlled RD analysis on the extended model set is per-condition and appears in Appendix A2.4 (Table 2, Fig. 5).

A4.4 Breadth–strength decomposition on pooled matrices

To obtain the density required for the breadth–strength decomposition, we pooled each perturbation family across its severity levels, restricted to the 12 parametric experiments that humans and ModelZoo models share, yielding $\sim$77 trials per class for models and $\sim$308 for humans. Metrics are computed as in Section 3.3. Per-family Wilcoxon tests compare $n = 12$ family blocks against $n = 12$ human blocks, but the pooled comparison against the benchmark-human-like models ($n = 84$ vs. 12) is better powered and is the one emphasized in the main text. Family means are given in Table 4 and block-level distributions in Fig. 9.

A5 Error Concentration at $K = 1000$

Models and stimuli. We evaluated three pretrained architectures — ResNet-50, ViT-B/16, and ConvNeXt-V2 (FCMAE-pretrained) — on the ImageNet-1k validation set (50,000 images, $K = 1000$ classes). This analysis is model-only and requires no matched human data, and is therefore independent of the GEN benchmark in images, categories, and label-space size. Because clean-data confusion matrices at $K = 1000$ are prohibitively sparse ($\sim$8 errors per class), we evaluated each model under three ImageNet-C corruption families spanning distinct corruption types (Gaussian noise, defocus blur, contrast) at severities 3 and 5, raising error density to $\sim$20–44 errors per class. Conditions were only analyzed when they reached at least 20 errors per class (11/18 conditions met this criterion and all are listed in Table 5 for completeness).

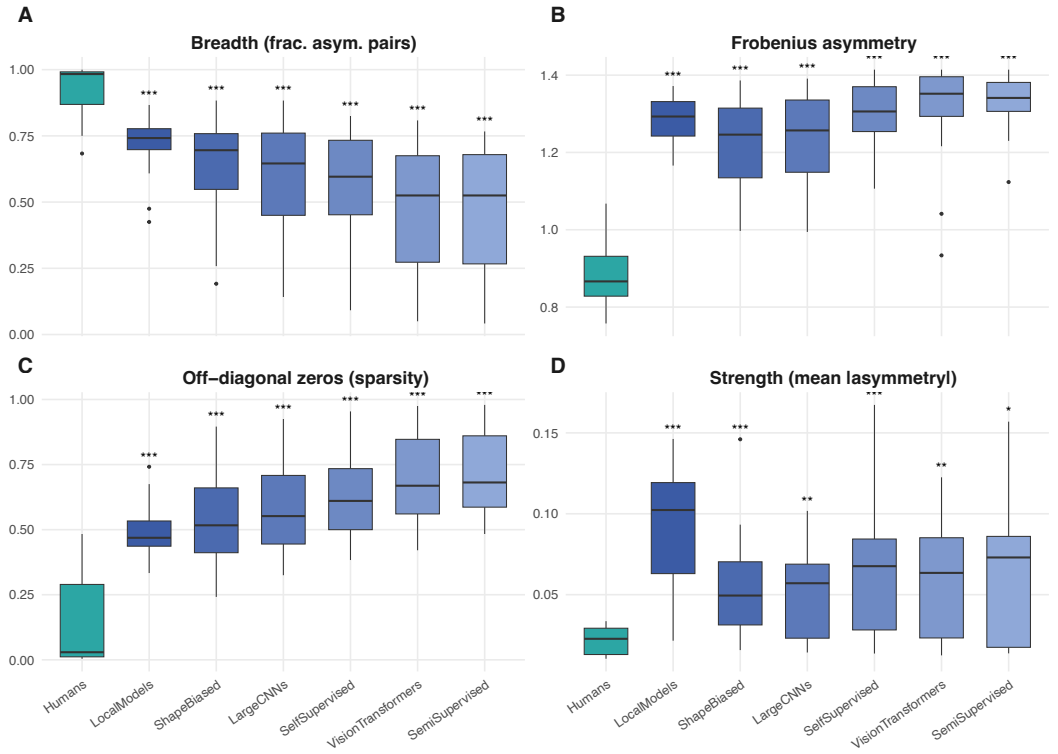

Figure 9: **Humans and modern architectures dissociate on all four asymmetry measures.** Block-level distributions for humans and six ModelZoo architecture families: **(A)** breadth (fraction of asymmetric class pairs), **(B)** Frobenius asymmetry A_F^{off}, **(C)** off-diagonal sparsity (fraction of empty off-diagonal cells), and **(D)** strength (mean $|C_{ij} - C_{ji}|$ among asymmetric pairs). Each box summarizes $n = 12$ blocks, one per shared parametric experiment. Matrices are pooled per perturbation family across severity levels, so values are not comparable to the per-condition analyses in Fig. 8 and Table 6. Every family is less broad, sparser, more Frobenius-asymmetric and higher in strength than humans. These are the 24 family×metric comparisons reported in Section 4.2. Stars denote planned two-sided Wilcoxon rank-sum comparisons against humans with BH–FDR correction ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). Family means are given in Table 4.

Concentration metric. Breadth and the pairwise asymmetry statistics are not estimable at this density, since the number of off-diagonal cells ($K(K-1) = 999{,}000$) vastly exceeds the number of errors. We therefore report the effective number of error targets per class, $n_1 = \exp(H)$, where H is the Shannon entropy of the row's off-diagonal error distribution, averaged over classes. This Hill number is well defined at low counts and is reported relative to density-matched references rather than interpreted in absolute terms.

Density-matched references. For each real confusion matrix we generated reference matrices at $K = 1000$ from the same broad–weak and sink cost geometries used in the generative simulation (Section 3.5), calibrated to the real matrix's per-class error count and trial count. Because the diffuse reference and the real matrix share an identical error density, finite-sample bias affects both equally and cancels in the comparison. The diffuse reference is stable across draws (46.1–46.5 effective targets), confirming that the comparison is not sensitive to the particular reference sample, but the concentration is not localizable to a subset of classes. The 50×50 block formed by the heaviest-receiving classes and the classes feeding them carries only $\sim$3.5% of total error mass, so no sub-block of the 1000×1000 matrix represents the structure and the matrix cannot be displayed directly at this scale. Concentration at $K = 1000$ is therefore per-class — each class spreads its errors over few targets — without a shared set of sink categories, the large-label-space counterpart of the shifting-receiver result in Section 4.2.

Scope. This experiment establishes that the concentration signature generalizes to a substantially larger label space and to transformer and self-supervised architectures. It does not establish that

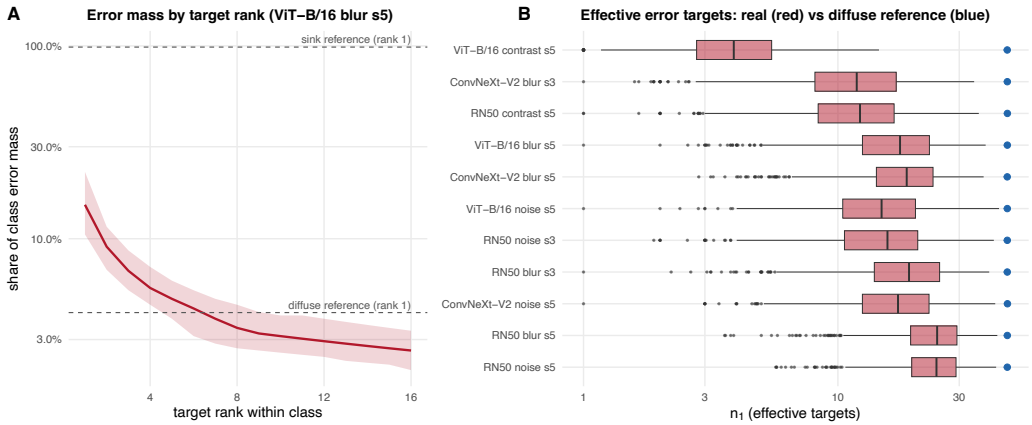

Figure 10: **Error concentration at $K = 1000$ against density-matched references. (A)** Share of a class's error mass falling on its k-th most-confused target for ViT-B/16 under defocus blur at severity 5, the condition closest to the median effective-target count across the 11 analyzed conditions (median, with interquartile range across classes), truncated at the last rank defined for at least three quarters of classes. Dashed lines mark the rank-1 share of the density-matched diffuse and sink references. The real profile sits between them, concentrated but not degenerate. **(B)** Effective number of error targets n_1 per class (red) against the diffuse reference (blue), for all 11 conditions meeting the density criterion, ordered by separation from the reference with the most concentrated (ViT-B/16, contrast, severity 5) at the top. The matrices themselves are not shown: at $K = 1000$ no compact sub-block carries the structure, so error structure is displayed as profiles rather than as a matrix. Reference values are those reported in Table 5 and the references are summarized rather than re-drawn here, so this figure and the reported statistics rest on identical quantities.

Table 5: **Error concentration at $K = 1000$, by condition.** The 11 of 18 model×corruption×severity conditions meeting the density criterion (≥ 20 errors per class), ordered by density. "Eff. targets" is the effective number of error targets per class, $\exp(H)$; "diffuse" is the density-matched broad–weak reference. Fold is the ratio of real to reference top-target share ($\sim$0.041 in all conditions).

			Errors	Top-target	Eff. targets		
Model	Corruption	Sev.	per class	share	Real	Diffuse	Fold
ResNet-50	Defocus blur	5	43.8	0.147	24.1	46.4	3.6×
ResNet-50	Gaussian noise	5	42.0	0.134	24.3	46.3	3.3×
ConvNeXt-V2	Defocus blur	5	36.1	0.183	18.9	46.3	4.4×
ResNet-50	Defocus blur	3	33.2	0.173	19.6	46.3	4.2×
ViT-B/16	Defocus blur	5	30.4	0.184	17.8	46.4	4.4×
ConvNeXt-V2	Gaussian noise	5	26.8	0.164	18.0	46.4	4.0×
ResNet-50	Gaussian noise	3	25.1	0.180	16.1	46.3	4.3×
ViT-B/16	Gaussian noise	5	24.1	0.182	15.7	46.3	4.4×
ViT-B/16	Contrast	5	24.1	0.576	4.3	46.3	13.9×
ConvNeXt-V2	Defocus blur	3	23.3	0.237	12.7	46.4	5.7×
ResNet-50	Contrast	5	21.2	0.219	12.7	46.4	5.3×

the full directional breadth–strength decomposition generalizes to $K = 1000$, which remains underdetermined at achievable error densities, nor does it extend to other task paradigms such as detection or segmentation, which do not yield a comparable n-way confusion matrix without further design choices.

A6 Secondary diagnostics

A6.1 Non-monotone frontier segments

We summarize RD geometry from the inferred costs using two complementary notions of *slope*:

1. **Global RD slope and curvature (primary).** For each cost matrix (ρ_{true} and $\hat{\rho}$) we compute an RD curve by sweeping an inverse-temperature parameter λ over a fixed grid and recording $(R(\lambda), D(\lambda))$. We define the global slope signature β as the median finite-difference derivative

along the frontier, $\beta = \text{median}\{\Delta R/\Delta D\}$. We define curvature κ as the variance of these finite-difference slopes, and efficiency (AUC) as the trapezoidal area under the RD curve. In 25 of the 1650 GEN blocks the Blahut–Arimoto sweep produces a numerically non-monotone frontier segment, yielding a positive β. These are near-deterministic channels — mean row entropy 2×10^{-4} and mean row-max 1.000, against 0.93 and 0.70 for the analysis set as a whole — so $\hat{\rho}$ is poorly identified over part of the range. Twenty-four are removed by the collapsed-channel filter (Appendix A2.1). The remaining block narrowly misses both criteria (mean row entropy 0.023 against the 10^{-3} cutoff; mean row-max 0.996 against 0.999) and is retained in all reported analyses; excluding it changes no group-wise Spearman correlation by more than 0.004.

2. **Local operating-point slope (secondary).** We compute the empirical operating rate $R^\star = I(X;Y)$ from the sampled confusion *counts* (using the empirical stimulus prior), then recover the local slope s by root-finding for the value of λ at which the optimal channel under cost ρ attains mutual information $R^\star$.

A6.2 Simulation Results

Channel Generation, Sampling, and RD Model Fitting. Given ρ_{true}, a generation inverse temperature λ_{gen}, and a uniform stimulus prior $p(x) = 1/K$, we generate channels via BA iterations and draw counts via

$$N_{i\cdot} \sim \text{Multinomial}(N_{\text{per row}}, q_{\lambda_{\text{gen}}}(\cdot \mid x = i)),$$

before recovering $\hat{\rho}$ by applying the same MAP RD pipeline used for empirical systems (see Section 3.2).

Simulation Grid and Diagnostics. We compute RD signatures from both ρ_{true} and $\hat{\rho}$, and compute asymmetry metrics from the corresponding induced channels/sampled confusions. Simulations are run over grids varying antisymmetry magnitude a, generation inverse temperature λ_{gen}, and per-class trial count $N_{\text{per row}}$, across both antisymmetry structures (broad–weak, sink) with multiple seeds. We reserve β for the empirical RD slope metric (generalization strength) and use λ to denote inverse-temperature parameters in Blahut–Arimoto and simulation generation. For pairwise breadth/strength summaries we threshold asymmetric pairs at $\varepsilon = 10^{-6}$, and for each replicate we collect RD metrics, asymmetry scores, and recovery diagnostics. We also fit replicate-level mixed-effects models and apply BH–FDR correction across trend tests (Table 11). Recovery is validated against known ground truth by the correlation between the symmetric components of ρ_{true} and $\hat{\rho}$, which defines the recovery criterion in Table 9. Slice-wise residual regressions are in Table 10.

Recovery sensitivity and regime-wise pass-rate tests. We report full recovery tables and per-slice two-sample proportion tests (BH–FDR), confirming broad–weak $>$ sinks pass rates in all 15 slices and FDR significance in $10/15$. We additionally summarize how recovery rates vary with λ_{gen} and $N_{\text{per row}}$ and emphasize that strict-recovery contrasts are conditional on identifiability, not unconditional mechanism differences.

Component-wise recovery of ρ is mechanism-dependent. We compared the symmetric and antisymmetric components of $\hat{\rho}$ against ground truth across the full grid. The antisymmetric correlation is undefined for the 300 runs at $a = 0$, where the true antisymmetric component vanishes. Among the remainder, recovery of the antisymmetric component is confined to the broad–weak generator, where the median correlation rises with antisymmetry magnitude (from 0.12 at $a = 0.1$ to 0.39 at $a = 2.0$) and with sample size (at $a = 2.0$, from 0.30 at $N_{\text{per row}} = 50$ to 0.44 at $N_{\text{per row}} = 1000$). For the sink generator it does not recover at any setting, with medians between -0.02 and -0.01 at every a and every $N_{\text{per row}}$. The symmetric component recovers moderately and declines as antisymmetry grows (0.60 at $a = 0$; 0.19 broad–weak and 0.14 sinks at $a = 2.0$) without improving at larger $N_{\text{per row}}$, indicating an identifiability rather than a sampling limit. This is the mechanism behind the pass-rate gap in Table 9: the directional component of sink-like costs is not identified by the MAP pipeline. Accordingly $\hat{\rho}$ should not be read as an estimate of directional cost structure for sink-like systems. The asymmetry metrics reported empirically are computed directly from the confusion matrices rather than from $\hat{\rho}$ (Section 3.3), so they are unaffected.

Slope and curvature behave differently from AUC: magnitude is sufficient. The residualized RD slope magnitude and curvature show a complementary pattern. For slope magnitude, global magnitude is highly predictive across all regimes (Frobenius coefficient negative in $15/15$ slices; median -0.144, range $[-0.202, -0.063]$, BH–FDR), whereas breadth is weaker and less reliable (median -0.783, significant in $3/15$ slices) and strength is more consistently negative (median -1.95,

Table 6: **Block-wise asymmetry summary.** One value per unique experiment×condition×model block. Values are mean±SE across blocks.

Metric	Blocks (A/H)	ANNs (mean±SE)	Humans (mean±SE)
Frobenius asymmetry index	1569 / 81	1.220 ± 0.0047	1.044 ± 0.0097
# asymmetric pairs (n_{pairs}), all ANNs	1569 / 81	64.16 ± 0.666	85.41 ± 2.76
# asymmetric pairs (n_{pairs}), baseline ANNs	243 / 81	53.95 ± 1.45	85.41 ± 2.76
Fraction of asymmetric pairs	1569 / 81	0.535 ± 0.0056	0.712 ± 0.023
Conditional strength $\bar{\Delta} = \mathbb{E}_{i<j}[\lvert C_{ij} - C_{ji}\rvert \mid \lvert C_{ij} - C_{ji}\rvert > \varepsilon]$	1569 / 81	0.1409 ± 0.0049	0.0422 ± 0.0022

Table 7: **Planned humans vs. group comparisons.** Frobenius asymmetry index A_F^{off} and accuracy, each group against humans (asymmetry 1.044, accuracy 0.564; $n = 81$ blocks). Δ is the group mean minus the human mean, with the Welch–Satterthwaite 95% CI. The rank column is the two-sided Wilcoxon rank-sum test, the parametric column Welch's t; both are Benjamini–Hochberg corrected across the six groups within each metric. Rank-based tests are primary (Section 3.3) and the two agree everywhere except Distortion-trained accuracy, where the parametric test is marginal. Blocks are collapse-filtered ($n = 1610$ retained); Wilcoxon W statistics are in the released analysis output.

					Welch		p_{BH}	
Group	n	Mean	Δ	95% CI	t	df	rank	param.
Frobenius asymmetry								
GoogLeNet	81	1.339	+0.295	[0.271, 0.319]	24.55	146.6	8.3×10^{-27}	5.2×10^{-53}
ResNet-152	80	1.315	+0.271	[0.239, 0.302]	16.92	148.1	2.7×10^{-22}	4.0×10^{-36}
VGG-19	80	1.329	+0.285	[0.256, 0.314]	19.42	156.2	4.4×10^{-24}	5.0×10^{-43}
Distortion-trained	1182	1.210	+0.166	[0.144, 0.188]	14.84	139.0	9.9×10^{-18}	2.8×10^{-30}
Specialised	53	1.022	−0.023	[−0.065, 0.020]	−1.06	79.0	0.136	0.291
All-noise	53	1.007	−0.037	[−0.078, 0.004]	−1.79	81.6	0.066	0.093
Accuracy								
GoogLeNet	81	0.512	−0.052	[−0.157, 0.052]	−0.99	157.4	0.923	0.371
ResNet-152	80	0.516	−0.048	[−0.154, 0.058]	−0.90	155.1	0.923	0.371
VGG-19	80	0.497	−0.067	[−0.170, 0.037]	−1.28	156.6	0.750	0.306
Distortion-trained	1182	0.482	−0.083	[−0.155, −0.011]	−2.28	93.9	0.612	0.050
Specialised	53	0.789	+0.225	[0.138, 0.311]	5.14	131.5	4.8×10^{-6}	6.0×10^{-6}
All-noise	53	0.775	+0.211	[0.119, 0.302]	4.54	131.0	4.8×10^{-6}	3.8×10^{-5}

significant in $10/15$ slices). For curvature (log-transformed), global magnitude is again uniformly predictive (Frobenius coefficient negative in $15/15$ slices; median -0.201, range $[-0.438, -0.089]$, BH–FDR), while breadth/strength terms are inconsistent. Thus, AUC seems to be the RD summary for which the *organization* of asymmetry (breadth vs strength) matters most, whereas slope and curvature primarily reflect overall directional magnitude.

Finite-sample error diagnostics for fitted versus true RD geometry. Recovery error is regime-dependent, where larger samples mainly stabilize the recovered AUC and reduce its dispersion, while slope-related quantities can fail sporadically in a mechanism-specific way. This motivates using non-collapsed data for primary trends and strict recovery as a conservative sensitivity analysis.

Saturation behavior for sink-like organization. Sink-like organization transitions rapidly into a sparse, high-strength regime, where breadth declines steeply and then saturates as antisymmetry grows. This supports the interpretation that sink-like directionality quickly concentrates into a small set of one-way confusions and then becomes insensitive to further increases in antisymmetry strength.

Strict-recovery robustness of residual effects. Finally, we repeat the key accuracy-residualized regressions under strict recovery and summarize sign/stability of the main effects in Table 12. This confirms that the qualitative pattern—AUC residuals requiring breadth–strength organization, and slope/curvature residuals tracking global magnitude—persists under conservative filtering.

Table 8: **Asymmetry–RD associations by group.** Spearman correlations between Frobenius asymmetry and RD metrics (AUC, β_{median}, $\log_{10}(\kappa{+}1)$), plus block-demeaned slope differences (Δslope) vs. humans. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

(A) Rank correlations

Group	n	%Coll.	ρ_{AUC}	ρ_{β}	ρ_{κ}
Humans	81	0.00	0.29	0.17	-0.22
GoogLeNet	81	0.00	-0.40	-0.33	0.43
ResNet-152	80	1.23	-0.39	-0.46	0.41
VGG-19	80	1.23	-0.43	-0.35	0.48
Distortion-trained	1182	3.11	-0.73	-0.57	0.72
Specialised	53	0.00	0.37	0.18	-0.31
All-noise	53	0.00	0.32	0.22	-0.36

(B) Δslope vs. humans

Group	AUC	β	κ
Humans		0 (ref.)	
GoogLeNet	2.67 (p=.34)	0.91 (.29)	-0.87 (.62)
ResNet-152	-0.69 (.74)	0.36 (.58)	-0.48 (.71)
VGG-19	-1.36 (.57)	-0.35 (.63)	0.61 (.68)
Distortion-trained	**-7.89** ($1.3{\times}10^{-7}$)***	**-1.21** (.0085)**	**3.96** ($2{\times}10^{-5}$)***
Specialised	**-4.77** (.014)*	**-1.22** (.042)*	**3.40** (.0049)**
All-noise	-3.55 (.055)	-0.83 (.14)	**2.56** (.026)*

Table 9: Recovery sensitivity. A run is counted as recovered if the correlation between the recovered and true symmetric components of ρ exceeds 0.2. Fractions are computed over all 60 runs per slice and structure, including numerically collapsed runs (26 collapsed runs pass this criterion). The table reports recovery fractions within each $(\lambda_{\text{gen}}, N_{\text{per row}})$ slice and BH–FDR adjusted p-values for differences in recovery rates (two-sample proportion tests).

λ_{gen}	$N_{\text{per row}}$	Frac. recovered (broad–weak)	Frac. recovered (sinks)	Δ frac.	p_{FDR}
0.2	50	0.567	0.433	0.133	0.232
0.2	200	0.700	0.583	0.117	0.271
0.2	1000	0.867	0.633	0.233	0.009
0.5	50	0.667	0.517	0.150	0.172
0.5	200	**0.867**	**0.450**	**0.417**	$<$0.001
0.5	1000	0.817	0.483	0.333	$<$0.001
1.0	50	0.817	0.500	0.317	0.001
1.0	200	0.783	0.433	0.350	$<$0.001
1.0	1000	0.817	0.517	0.300	0.002
2.0	50	0.817	0.533	0.283	0.003
2.0	200	0.817	0.417	0.400	$<$0.001
2.0	1000	0.767	0.400	0.367	$<$0.001
5.0	50	0.900	0.517	0.383	$<$0.001
5.0	200	0.833	0.750	0.083	0.369
5.0	1000	0.817	0.633	0.183	0.056

Table 10: **Residual regression summary.** Within each $(\lambda_{\text{gen}}, N_{\text{per row}})$ slice, we first residualize each RD signature by mean diagonal probability (accuracy proxy), then regress the residual on either (i) breadth/strength terms or (ii) a global asymmetry magnitude term. "Structure offset" refers to contrast between sink-like and broad–weak configurations. Coefficients are summarized across slices; significance counts use BH–FDR within-slice tests.

Outcome	Predictor set	Term	Median	Min	Max	Sig. slices	Slices
AUC	components	Breadth	-2.05	-5.40	1.52	11	15
AUC	components	Strength	-4.28	-11.56	2.45	12	15
AUC	components	Breadth × Strength	5.33	-4.65	16.27	5	15
AUC	components	Structure offset	-0.447	-0.657	-0.162	15	15
AUC	magnitude	Global asymmetry mag.	-0.0047	-0.0470	0.206	3	15
AUC	magnitude	Structure offset	-0.482	-0.691	-0.171	15	15
Global slope	components	Breadth	-0.78	-2.12	0.37	3	15
Global slope	components	Strength	-1.95	-3.70	-0.40	10	15
Global slope	components	Breadth × Strength	-5.02	-9.96	5.73	7	15
Global slope	components	Structure offset	-0.188	-0.357	0.066	9	15
Global slope	magnitude	Global asymmetry mag.	-0.144	-0.202	-0.063	15	15
Global slope	magnitude	Structure offset	-0.243	-0.386	-0.081	15	15
Curvature	components	Breadth	-0.048	-0.167	0.129	3	15
Curvature	components	Strength	-0.165	-0.540	0.084	6	15
Curvature	components	Breadth × Strength	0.066	-1.030	1.250	2	15
Curvature	components	Structure offset	0.058	-0.011	0.124	7	15
Curvature	magnitude	Global asymmetry mag.	-0.201	-0.438	-0.089	15	15
Curvature	magnitude	Structure offset	0.060	-0.009	0.126	8	15

Table 11: **Mixed-effects ANOVA key terms.** Selected fixed-effect tests from replicate-level mixed models $Y \sim \texttt{structure} \times a \times \lambda_{\text{gen}} \times \log_{10} N_{\text{per row}} + (1|\texttt{cell})$.

Outcome	Term	F	p
auc_true	structure	10.16	0.00146
auc_true	a_scale	1176	2.2e-16
auc_true	structure:a_scale	877.1	2.2e-16
beta_median_true_pos	structure	24.41	7.8e-07
beta_median_true_pos	a_scale	827.7	2.2e-16
beta_median_true_pos	structure:a_scale	613.4	2.2e-16
kappa_true	structure	5.34	0.0209
kappa_true	a_scale	964.7	2.2e-16
kappa_true	structure:a_scale	701.8	2.2e-16

Table 12: **Strict-recovery robustness of residual AUC effects.** For each $(\lambda_{\text{gen}}, N_{\text{per row}})$ slice and structure, we regress accuracy-residualized AUC on antisymmetry strength a under the no-collapse filter (all non-collapsed runs) and the strict-recovery filter (additionally requiring symmetric-component recovery correlation > 0.2). Columns show the slope sign and BH–FDR adjusted p-value under each filter. Signs agree in all 30 rows — broad–weak positive in 15/15 slices and sinks negative in 15/15 under both filters — so the qualitative pattern is robust to identifiability constraints. However, the two structures differ in what strict recovery costs them: broad–weak slopes remain at $p < 0.001$ in every slice, whereas sink slopes fall from $p < 0.001$ in 11/15 slices to 4/15, consistent with the recovery gap in Table 9.

			No-collapse		Strict-recovery	
λ_{gen}	$N_{\text{per row}}$	Structure	Sign	p_{FDR}	Sign	p_{FDR}
0.2	50	Broad–weak	+	<0.001	+	<0.001
0.2	50	Sinks	−	<0.001	−	0.003
0.2	200	Broad–weak	+	<0.001	+	<0.001
0.2	200	Sinks	−	<0.001	−	<0.001
0.2	1000	Broad–weak	+	<0.001	+	<0.001
0.2	1000	Sinks	−	0.002	−	0.002
0.5	50	Broad–weak	+	<0.001	+	<0.001
0.5	50	Sinks	−	<0.001	−	0.019
0.5	200	Broad–weak	+	<0.001	+	<0.001
0.5	200	Sinks	−	<0.001	−	0.042
0.5	1000	Broad–weak	+	<0.001	+	<0.001
0.5	1000	Sinks	−	<0.001	−	0.029
1.0	50	Broad–weak	+	<0.001	+	<0.001
1.0	50	Sinks	−	<0.001	−	<0.001
1.0	200	Broad–weak	+	<0.001	+	<0.001
1.0	200	Sinks	−	<0.001	−	0.006
1.0	1000	Broad–weak	+	<0.001	+	<0.001
1.0	1000	Sinks	−	<0.001	−	<0.001
2.0	50	Broad–weak	+	<0.001	+	<0.001
2.0	50	Sinks	−	<0.001	−	<0.001
2.0	200	Broad–weak	+	<0.001	+	<0.001
2.0	200	Sinks	−	<0.001	−	0.049
2.0	1000	Broad–weak	+	<0.001	+	<0.001
2.0	1000	Sinks	−	<0.001	−	0.012
5.0	50	Broad–weak	+	<0.001	+	<0.001
5.0	50	Sinks	−	0.005	−	0.012
5.0	200	Broad–weak	+	<0.001	+	<0.001
5.0	200	Sinks	−	0.004	−	0.011
5.0	1000	Broad–weak	+	<0.001	+	<0.001
5.0	1000	Sinks	−	0.012	−	0.013

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state all four main findings. The human–model comparison is bounded to $K = 16$ visual categorization under controlled perturbation, while the model-side generalization to $K = 1000$ is stated with its own scope limits in Appendix A5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The dedicated Limitations paragraph in the Discussion addresses label granularity, human sample size, MAP prior sensitivity, mechanism-dependent recoverability, and the representational interpretation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is empirical and simulation-based and does not contain any formal theorems or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The full pipeline is described in Section 2 and Section 3; the MAP prior, λ grid and signature definitions are specified in Appendix A2.3; the simulation grid is specified in Appendix A6.2; all details of the GEN and ModelZoo benchmarks are publicly available.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All analysis code is available at `https://github.com/leylaroksan/rdt_asymmetry`, with scripts reproducing every figure and table. The derived data needed to reproduce the results without re-running MAP inference are deposited at figshare under CC BY 4.0: `https://doi.org/10.6084/m9.figshare.34059993`. The underlying behavioral and model data are publicly available via the GEN and ModelZoo repositories [Geirhos et al., 2018b, 2021]. See Appendix A1.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All perturbation families, model architectures, training regimes, and the block-wise aggregation procedure are described in Sections 3.1–3.5 and Appendix A4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Group comparisons use Wilcoxon rank-sum tests with BH–FDR correction, reported alongside Welch's t-tests and 95% confidence intervals on the mean difference via the Welch–Satterthwaite approximation (Table 7). Error bars are s.e.m. as stated in

figure captions. Accuracy-controlled regressions report slopes with standard errors, t-statistics and BH-corrected p-values under both OLS and cluster-robust (CR2) inference with Satterthwaite degrees of freedom (Appendix A2.4).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute requirements are detailed in Appendix A1. The rate–distortion inference, asymmetry metrics and simulation analyses are CPU-based and were run on the Minerva HPC cluster. The ImageNet-1k evaluation additionally required GPU inference over the ImageNet validation set (Appendix A5).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available datasets with no human subjects recruited by the authors. There are no privacy or fairness concerns that arise.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This is foundational research on evaluation methodology. No direct path to harmful applications is identified, however, the work could improve alignment evaluation, which is broadly positive.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No new models, datasets, or high-risk assets are released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The GEN benchmark is used under CC BY 4.0 and the ModelZoo benchmark code under the MIT License, both cited with their original papers. Pre-trained weights accessed through ModelZoo remain under their own licenses, including CC BY-NC 4.0 for SWSL ResNeXt-101. The ImageNet-1k evaluation additionally uses the ImageNet validation set under its terms of access, the ImageNet-C corruption benchmark under CC BY 4.0, and pretrained weights obtained through `timm` under the Apache License 2.0. Licenses for each are stated in Appendix A1.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: No new models or human data are introduced. The derived data supporting the analyses — fitted distortion matrices, RD signature summaries, collapse flags, simulation grid outputs, and the $K = 1000$ confusion matrices — are deposited at `https://doi.org/10.6084/m9.figshare.34059993` under CC BY 4.0. Provenance and licensing of the source data are given in Appendix A1.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No new human subjects were recruited. The human psychophysics data is from published benchmark repositories.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new human subjects research was conducted. Only existing data from published repositories was used.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a component of the core methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.