MEASURING MODEL–BRAIN ALIGNMENT IN LATENT SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating neural network models as hypotheses of brain computation requires accurately quantifying the alignment between artificial and biological representations. Here, we propose measuring representational alignment within a shared latent space learned via a cross-modal multi-encoder–decoder architecture. By mapping network and brain representations into a common low-dimensional space, our framework enables direct comparison of both overall latent distributions and per-stimulus correspondences. To demonstrate the properties of our method, we construct two diagnostic benchmarks using fMRI data paired with reference and manipulated network representations: (1) structural "foil" models, which are mechanistically dissimilar to brain computation by design yet match its coarse high-level structure, and (2) reference model representations with systematically corrupted information. We show that linear ridge regression is nearly flat across these structural transformations, whereas commonly used similarity metrics – Representational Similarity Analysis (RSA) and Centered Kernel Alignment (CKA) – score the foil models close to, or above, the reference network. In contrast, metrics derived from our latent space (i) accurately capture progressive information loss under controlled corruption and (ii) reliably identify structural discrepancies between the foil models' representations and biological neural data, providing a new robust, structure-sensitive tool for brain–AI alignment.

1 INTRODUCTION

Artificial neural networks have become a dominant framework for formalizing computational hypotheses of biological information processing (Golan et al., 2023). Evaluating the plausibility of candidate models typically relies on quantifying the alignment between their internal representations and neural or behavioral data (Golan et al., 2026; Cichy et al., 2016; Yamins et al., 2014). However, how to best measure this alignment remains an open challenge. Current benchmarks and large-scale model comparisons (Schrimpf et al., 2020; Conwell et al., 2024; Storrs et al., 2021) generally rank models along two core axes of neural predictivity: *encoding-model accuracy* (the degree to which individual brain responses can be predicted from neural network representations (Yamins et al., 2014; Naselaris et al., 2011) and *geometric similarity* (the extent to which both systems preserve similar representational geometries across a shared set of stimuli (Kriegeskorte et al., 2008; Kornblith et al., 2019).

Despite recent advances, including the introduction of generalized shape metrics (Williams et al., 2021; Khosla & Williams, 2024), key issues remain. One central issue is saturation: hundreds of architecturally distinct networks reach near-identical predictivity and geometric similarity (Conwell et al., 2024; Storrs et al., 2021). The high parametric capacity of predictive encoding models makes them flexible enough to dramatically reorient a network's tuning basis (Prince et al., 2024); therefore benchmarks based on linear predictivity saturate (Höfling et al., 2026; Feather et al., 2025); and rankings can change when the representational alignment measure changes (Soni et al., 2024). Another issue is attribution, by optimising synthetic activity to maximise standard similarity scores against brain recordings, Cloos et al. (2025) demonstrated that a high similarity score does not guarantee that models encode task-relevant information in a manner consistent with neural data.

Here we propose a new approach to this problem: to measure alignment between model and subjects in a learned latent space. We make three key contributions:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1. We introduce two complementary type of metrics computed after mapping model and brain representations into a shared latent space via a joint cross-modal variational autoencoder (VAE): (i) **distributional metrics** – test whether the model's and the brain's latent representations for given stimuli are distributed alike, occupying the same region of the latent space (ii) **per-stimulus correspondence** – tests whether the model's representation of image i lands closer to the brain's representation of that same image than to the brain's representation of other images.
2. To validate these metrics, we test them against two negative controls that an ideal alignment measure should penalize relative to a reference network: (i) **foil models** – mechanistically dissimilar to the brain, yet matching its high-level structure. (ii) **synthetic categorical representations** – progressively corrupted categorical representations of the reference network.
3. Our results show that our latent metrics (i) reliably identify the foil models' representations as dissimilar to the neural data and (ii) accurately capture progressive information loss in the synthetic categorical representations. Ridge regression is compressed, i.e., is nearly flat across all constructions, and RSA and CKA reward representations that follow the brain's coarse high-level structure despite being mechanistically dissimilar to it.

2 Related work

An ideal representational alignment measure should be insensitive to variations that are of no interest, such as noise in the data and interindividual idiosyncrasies that do not reflect computational function, while being sensitive to the representational signatures of different computational models. Generalized shape metrics make explicit what transformations we consider nuisance variation (Williams et al., 2021). These metrics measure the discrepancy that remains after alignment using the transformations that we wish to discard. The approach of linear encoding models provides high parametric flexibility. While it enables us to find the best fitting representational dimensions to account for a brain representation, it can render computationally distinct models empirically indistinguishable (Avitan & Golan, 2026). On the other end of the spectrum, rigid measures of alignment such as fixed RSA (Khaligh-Razavi et al., 2017; Schütt et al., 2023) or the softmatching distance (Khosla & Williams, 2024) often reveal significant differences but may lead us to reject models that could be salvaged by reweighting of their features.

Unlike traditional metrics computed directly from raw stimulus-response matrices, here we explore the benefits of learning a joint latent space across a model and all subjects' neural data. Although training introduces context-dependence, it promises to provide a powerful shared representation that captures both what is consistent among the representations and what varies. Unlike encoding models, comparisons in the shared latent space penalize a model for having irrelevant features. Unlike RSA, the latent space can emphasize dimensions explaining less representational variance when this improves encoder-decoder performance. Like encoding models and weighted RSA, our framework allows reweighting of representational dimensions and more general affine transformation, while still penalizing features not represented in the brain. When adjudicating among multiple candidate computational models, we show that the lingua franca of the shared latent space can help us recognize the correct model.

3 The framework

Figure 1 illustrates our framework: a multimodal variational autoencoder (VAE) jointly embeds model representations and subject fMRI responses into a shared latent space, where model–brain alignment is directly evaluated.

Joint embedding. We embed model and brain in a common latent space with a multimodal variational autoencoder (Kingma & Welling, 2013; Papathanasiou et al., 2026) in which every structure (model and brain) has its own encoder and decoder and all pathways are active, i.e., each structure is reconstructed from itself and from the other. In the case of brain data coming from multiple subjects, each subject also has their own encoder and decoder. Formally, the model representation has encoder E_{m} and decoder D_{m}, and subject $s = 1, \ldots, S$ has encoder E_s and decoder D_s. Encoders map to a shared 32-dimensional latent, and decoders map latent representations back to their original structure spaces, and also cross-reconstruct to the other modality. More details on the VAE architecture and training and evaluation specifics are given in Appendix B.

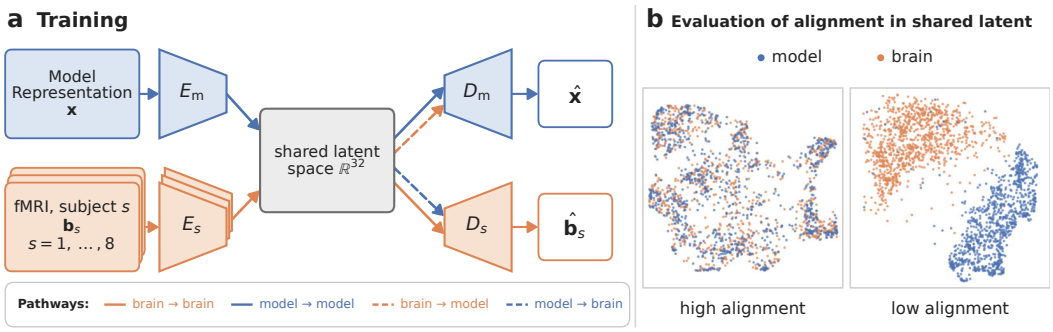


Figure 1: **Framework for measuring model-brain alignment in latent spaces. (a)** Training: the model representation $\mathbf{x}$ and each subject's fMRI response $\mathbf{b}_s$ are encoded in one shared 32-d latent space. Each structure is reconstructed both from its own and from the other's latent representation. **(b)** Examples of model-brain alignment during evaluation. Model and brain latents of the test images are compared inside the 32-d latent space. The clouds are UMAP-2d projections, with model in blue and brain in orange, for a reference trained network (left) and for the shuffled null (right).

Objective. Each training image was seen by exactly one subject s. With model features $\mathbf{x}$, brain response $\mathbf{b}_s$, and single reparameterised samples $\mathbf{z}_\mathrm{m} \sim q_\mathrm{m}(\mathbf{z} \mid \mathbf{x})$ and $\mathbf{z}_s \sim q_s(\mathbf{z} \mid \mathbf{b}_s)$ from the diagonal-Gaussian posteriors output by E_m and E_s, the loss of the image is:

$$\mathcal{L}_s = \underbrace{\big[\ell(D_s(\mathbf{z}_s), \mathbf{b}_s) + \ell(D_s(\mathbf{z}_\mathrm{m}), \mathbf{b}_s)\big]}_{\text{brain}\to\text{brain} + \text{model}\to\text{brain}} + \underbrace{\big[\ell(D_\mathrm{m}(\mathbf{z}_\mathrm{m}), \mathbf{x}) + \ell(D_\mathrm{m}(\mathbf{z}_s), \mathbf{x})\big]}_{\text{model}\to\text{model} + \text{brain}\to\text{model}} + \big[\mathrm{KL}_\mathrm{m} + \mathrm{KL}_s\big], \quad (1)$$

where $\ell(\hat{\mathbf{y}}, \mathbf{y}) = \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2$ is the squared error summed over all output dimensions (51,048 z-scored model features; $\approx$ 20,000 voxels), $\mathrm{KL}_\mathrm{m} = \mathrm{KL}\big[q_\mathrm{m}(\mathbf{z} \mid \mathbf{x}) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big]$ and likewise for KL_s. The training objective averages Eq. 1 over the training images of all subjects, i.e. it sums over every (model, subject) pair; it has no subject–subject terms, as no training image was seen by two subjects (full form in Appendix A).

We note that no term in the loss pulls the model and brain latent representations together, this is implicitly encouraged through the cross-modal reconstruction. Thus, it is not guaranteed that there exist low-dimensional projections of the two structures (brain and model), which occupy the same region of the latent space, and can reconstruct both its own and the other original structures. A low-dimensional representation whose dominant structure is brain-alien must spend shared capacity on it (because of the self-reconstruction term), whereas an encoding model keeps the brain-predictive projection of a model and can mask large sets of brain-inaccessible units as shown in (Muzellec & Kar, 2026). The framework is *bidirectional* and, importantly, encodes the two structures in one common space, enabling the direct comparison on a distributional level, as *first-order* point clouds in one space — whether the clouds interpenetrate, whether the model and brain latent representations for stimulus i's coincide.

Metrics. In the common latent space, we derive the following metrics to compare the model's with the brain's representations:

- **modality integration LISI** (iLISI)) (Korsunsky et al., 2019; Sun et al., 2026): a local distributional metric, which for each latent representation takes the inverse Simpson index of the modality labels (label:model/brain) among its k nearest neighbours, averaged and shifted to run from 0 (fully segregated) to 1 (fully mixed).
- **Energy distance** (Székely & Rizzo, 2013): $2\,\mathbb{E}\|\mathbf{z}_\mathrm{m} - \mathbf{z}_s\| - \mathbb{E}\|\mathbf{z}_\mathrm{m} - \mathbf{z}'_\mathrm{m}\| - \mathbb{E}\|\mathbf{z}_s - \mathbf{z}'_s\|$ (zero only when the two distributions coincide) quantifies the global distributional difference.
- **Identification**: i) **Top-k**: the fraction of images whose model latent point for stimulus i has, as one of its k nearest points, the brain latent point for the same stimulus (top-1 chance 0.001). ii) **Median rank**: for each test image, the brain latent points of all 872 test images are ranked by their distance to that image's model latent point; we report the median over

images of the rank of each image's own brain point (1 = nearest; lower is better; chance 436.5).

3.1 DATA AND FEATURES

Brain. All subjects from Natural Scenes Dataset (NSD) (Allen et al., 2022), and ROIs of ventral, lateral, midventral and midlateral streams of the streams parcellation ($\approx$ 20k voxels per subject) were used from the Algonauts 2023 (Gifford et al., 2023). For measures that require training, e.g., ridge and the VAE, the training images are the $\approx$ 9k images unique to each subject; evaluation for all measures is on the 872 images seen by all eight subjects, held out from training.

Model. ResNet-50 (ImageNet-1K supervised weights) (He et al., 2016; Deng et al., 2009) was used as the reference model. Sixteen layers sampled at fixed intervals to get full coverage of the representations (seven convolutional, seven ReLU and average-pool) are projected by a fixed sparse random projection to a 51k-dimensional feature vector per image. Sampling at fixed intervals covers the network uniformly, every stage and every kind of operation, without presupposing where brain-relevant computation lies. This is the model feature basis used in the main results, but we run the same analysis taking a second basis, the layers where RN-50's computation is organised: the output of each of the sixteen residual blocks and the average pool, i.e., the representations each stage passes on to the next (Appendix G).

Labels. NSD contains 80 fine-grained *categories* in 12 *supercategories*. Where a construction needs a single category per image we use the rarest label present, as this is the most informative choice.

4 CONSTRUCTED REPRESENTATIONS

To characterise what each metric reads, we build two types of representations: *Synthetic Categorical Representations* that discard or corrupt categorical information (in a reference network's representations) in known steps (ladders), and *Foil models Represnetations* that keep the brain's high-level supercategory structure, but are mechanistically dissimilar to the brain computation. We hypothesize that these constructed *"inferior"* model representations, should align with the brain data less that the reference network and explore how different metrics score them. Below, the details of the constructions follow and a visual schematic describing them can be found in Fig. 2.

4.1 SYNTHETIC CATEGORICAL REPRESENTATIONS

The first rung of *Category* and *Supercategory Corruption ladders* starts from the reference representations, i.e., pretrained on ImageNet RN50, and replaces each image's representation vector by the average representation of its group (group=category or supercategory set, respectively). In that way, parts of the representation that encode additional information to the group identity are averaged out. For example, averaging the network's activations over all images that contain a cat, discards the within-category detail those activations carried — the colour, pose and setting of each cat — and keeps only what corresponded to the category itself, the presence of a cat. We take the derived representations to simulate an inferior network: one that recognizes categories but not the finer detail within them, which we expect to align with the brain less. To make the constructions realistic, we add back the leading components ($k = 10$ PCs) of variance to each representation, as a network would not respond identically to every image of a cat even when nothing finer than "cat" is encoded.

For the subsequent rungs in the ladder, on top of discarding finer detail information in representations, we also actively corrupt them with wrong information. Rungs within ladders differ in two ways. The first is *whose average* an image receives: its own group's, or another group's chosen by a swap rule, e.g. swapping cats with pizzas. Thus, when we compute any alignment metric, we will try to align the brain activations for the cat images to the model's representations for the pizza. We take this to simulate a model that mixes categories up. Importantly, a swap does not alter the network's geometry: the group means and the distances among them are unchanged. Only the image-to-mean assignment is corrupted, so that, e.g., cat and dog images, which both the network and the brain place close together, are mapped to the distant pizzas and chair positions. So, with this mechanism, relations among images that the brain preserves are corrupted in the model's representations. The second is *whose variation* is added back: the image's own, or another image's.

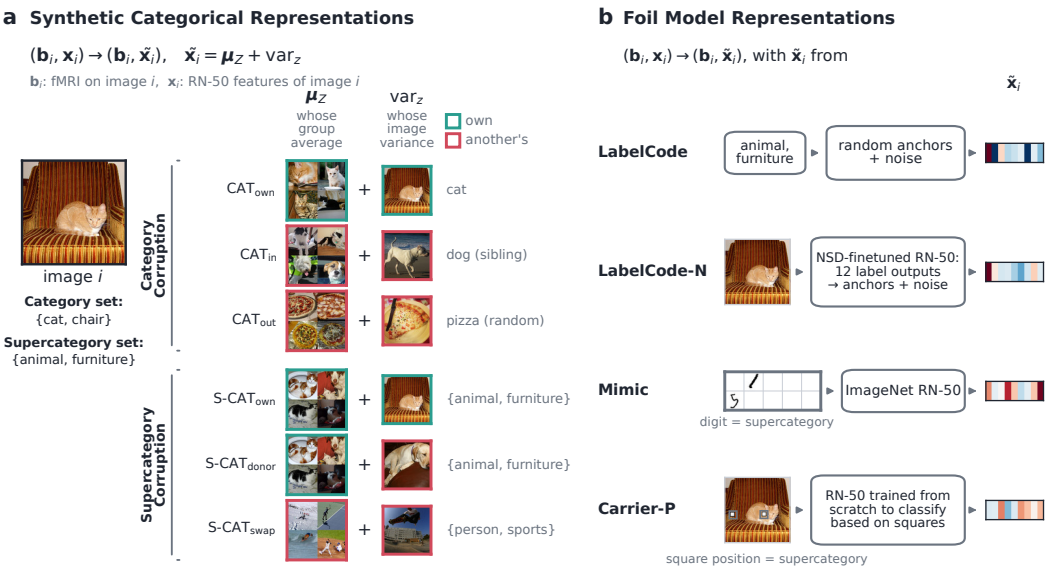


Figure 2: **Illustration of constructed representations used to test metrics.** Each construction replaces the RN50 features ($\mathbf{x}_i$) paired with the fMRI response ($\mathbf{b}_i$) of image i, with $\tilde{\mathbf{x}}_i$: $(\mathbf{x}_i, \mathbf{b}_i) \to (\tilde{\mathbf{x}}_i, \mathbf{b}_i)$. **(a)** *Synthetic Categorical Representations*: $\tilde{\mathbf{x}}_i = \boldsymbol{\mu}_Z + z$, replaces the model representations with a *group* average plus one image's variation along the top 10 principal components. Category Corruption ladder (*group* = category): CAT_{own} replaces the representation with the average of the image's own category group; CAT_{in}, with the average of a sibling category in the same supercategory (dog); CAT_{out}, with a random category (e.g. pizza). For Supercategory Corruption ladder, *group* = supercategory set. **(b)** *Foil Models* are built to retain the supercategory structure in their representations but by a different mechanism: LABELCODE: category label random anchors plus noise; MIMIC: RN50 inference on MNIST-digit canvas of the supercategories; CARRIER-P: RN50 trained to classify from planted squares.

We hypothesise that these constructed representations align with the brain less than the reference RN50 they are built from, in an order fixed by construction; each rung containing less, or less correct, information than the one before, and we ask how each metric scores that order. The brain side is never altered: every instrument sees the pairing of the brain response $\mathbf{b}_i$ with the constructed, corrupted, vector $\tilde{\mathbf{x}}_i$. The constructions are representations, not networks. However, they are meant to simulate network behavior, i.e., what a network of a given (lower) competence would emit.

Formally, and as shown in Fig. 2, let I be a group—a category on the **Category Corruption ladder**, the exact set of supercategories present in the image on the **Supercategory Corruption ladder**—with images $i_1, \ldots, i_n$, and let J be another group chosen by the swap rule. For image i_ℓ, the brain response $\mathbf{b}_{i_\ell}$ is paired with the corrupted $\tilde{\mathbf{x}}_{i_\ell}$:

$$\tilde{\mathbf{x}}_{i_\ell} \;=\; \boldsymbol{\mu}_Z \;+\; \text{variance}_z, \qquad Z \in \{I, J\}, \quad z \in \{i_\ell, i_m, j_p\}, \tag{2}$$

where $\boldsymbol{\mu}_Z$ is the average of the network's activations over the images of Z, and $\text{variance}_z = \text{PC}_k(\mathbf{x}_z - \boldsymbol{\mu}_Z)$ is the variance of image z about its group average, kept to its top-k principal components. Generally, $k = 10$, except if stated otherwise. The two choices are whose average and whose variation image i_ℓ receives.

- $Z = I$ keeps the image's own group average; $Z = J$ swaps it for another's. On the Category Corruption ladder, J is I's partner in a fixed pairing of the categories: a sibling in the same supercategory (CAT_{in}); or a randomly paired category (CAT_{out}). On the Supercategory Corruption ladder the only swap is a random swap of supercategory-sets.
- If $Z = J$, then $z = j_p$: the variation of an image of the swapped group, drawn at random with replacement (as groups differ in size). If $Z = I$, then $z = i_\ell$ adds back the image's own variation, or $z = i_m$, $m \neq \ell$, adds another image's (belonging to I, drawn at random).

Table 1, in Appendix C summarizes the rungs for Category and Supercategory Corruption Ladders.

4.2 FOIL MODEL REPRESENTATIONS

The foil models are built to retain the supercategory block structure in their representations, but are mechanistically dissimilar with both the reference network and the brain representations, by construction. They differ in whether they result from a real network or are artificially constructed, on what data training/inference is performed on, etc. Here, their details follow and Appendix C specifies each implementation in full.

LABELCODE: is built directly from each image's supercategory labels — the image itself is never processed — so whatever alignment a metric assigns rewards supercategory membership alone rather than any visual processing computations a network would make. We construct artificial representations in a way that mirrors Eq. 2, but with both terms replaced: instead of averaging RN50's representations we use random anchors, and instead of an image's own variation we add noise along the network's residual directions:

$$\tilde{\mathbf{x}}_i = c\, L_i B + \sum_{j \le k} \varepsilon_{ij}\, \sigma \sqrt{\lambda_j}\, \mathbf{v}_j, \qquad B \in \mathbb{R}^{12\times D},\ B_{ab} \overset{\text{iid}}{\sim} \mathcal{N}(0,1),\ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1),\ k = 10, \quad (3)$$

where L_i is the multi-hot supercategory vector, $\mathbf{v}_j$ and λ_j are the leading directions and variances of the trained network's within-category residuals, c scales the centred signal LB to the per-image signal variance of the additive ceiling code, and σ sets the signal-to-noise ratio to ≈ 0.30, so the noise carries about three times the variance of the signal. The anchors B are twelve independent random directions in RN50 feature space.

LABELCODE-N: is Eq. 3 with the annotation vector L_i replaced by the output of N_{ft}, a ResNet-50 fine-tuned on the twelve-way supercategory task, with the same anchors B and the same noise. NSD images are fed to the N_{ft} model, and the 12 value output is what we take as L_i in Eq. 3, instead of the multihot supercategory labels in LABELCODE.

MIMIC — RN-50 with artificial stimulus. Instead of artificially constructing the network's representations, here, we artificially construct the stimulus. For each NSD image, a canvas of MNIST digits is assembled from its supercategory labels, as in Fig. 5 in Appendix, and passed through the same RN-50. The mimic therefore has genuine network features and genuine variance, but no content of the image whose brain response it is paired with, thus, the two representations cannot be mechanistically similar.

CARRIER-B, CARRIER-P — RN-50 on the natural images, trained to ignore them. Following Dujmović et al. (2023), every NSD image receives one fixed-position white square per supercategory in its multihot label-set, and a RN50 is trained from scratch on these images. The network learns to classify from the squares alone (mAP 1.00 with them, 0.24–0.26 without). In CARRIER-B the twelve positions are chosen so that the squares' pixel distances mirror the brain's supercategory geometry ($r = 0.93$ between the pixel and brain RDMs); in CARRIER-P the same positions are permuted across supercategories ($r = -0.08$). These networks do see the NSD image, and their early layers possibly encode some NSD image information, but their primary encoded information (based on which they classify) is the square positions.

4.3 CONTROLS

Degradation ladder RN50$_\alpha$. RN50$_\alpha$ interpolates the weights of the ImageNet-trained RN50 toward a random initialisation, $W(\alpha) = (1-\alpha)\, W_{\text{trained}} + \alpha\, W_{\text{init}}$, for $\alpha = 1$, we get an untrained network, and the ladder is the reference for genuine loss of quality. $RN50_{\text{sc}}$ is a ResNet-50 trained from scratch on the NSD images with the twelve supercategory labels as its only supervision (test mAP 0.81), and the honest counterpart of the CARRIER-B and CARRIER-P networks, which were given the same labels and a way to circumvent them.

Shuffled null. The ImageNet pretrained RN50's features with the image–brain 1:1 correspondence destroyed. The feature distribution is identical, only the correspondence is removed. Every floor-corrected percentage in the paper is $(x - \text{null})/(\text{RN-50} - \text{null})$ with this row as the null.

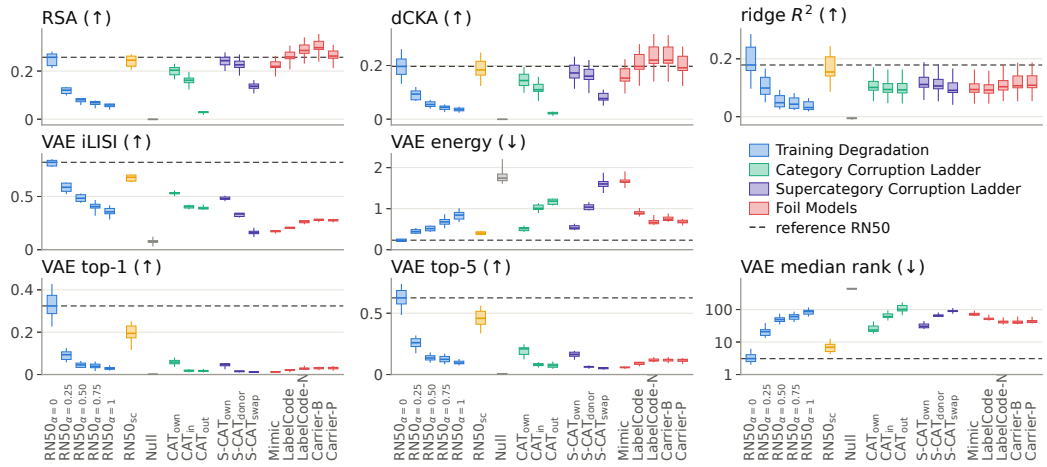


Figure 3: **Alignment scores.** Each box summarises the eight subjects; arrows mark the direction of better alignment for each metric. Ridge, RSA and dCKA are computed on the features; the latent read-outs (iLISI, energy, top-1, top-5, median rank) in the VAE's 32-d co-embeddings, averaged over three VAE training seeds for each subject. Identification is model-to-brain over the 872 shared test images (chance 0.001, 0.006 and 436.5; median rank on a logarithmic axis). Dark line: the median of the reference RN50 (RN50$_{\alpha=0}$). Means over subjects are in Table 3 (Appendix E).

5 Results

Fig. 3 reports the three standard metrics we are evaluating—ridge regression (Naselaris et al., 2011), RSA (Kriegeskorte et al., 2008) and debiased CKA (dCKA), i.e., CKA (Kornblith et al., 2019) computed from the unbiased HSIC estimator of Song et al. (2012), as in Nguyen et al. (2020)—and the framework's read-outs from the co-embedded 32-d latent (Section 3): occupancy (iLISI), energy distance, and identification—top-1/top-5 retrieval and the median rank of the true image—as described in Section 3. The *Training Degradation* and Null rows anchor the scale for each metric and assess **sensitivity**; the *Foil models* relative to the trained reference network test **specificity**; the *Synthetic Categorical Representations* test **expressivity** and **ordering**. Numbers in the text are means over the eight subjects (Table 3, Appendix E); Fig. 3 shows their spread across subjects. Statistical tests are in Appendix F. We note that the same patterns as in the results below hold when the experiments are done in a second RN-50 feature basis (Appendix G). Along with the agreement across three training seeds (Table 3), this provides further robustness to the results.

5.1 Foil Models

In Fig. 3 we show the results for the foil models—Carrier-B, Carrier-P, Mimic, LabelCode and LabelCode-N, described in Section 4.2 and Appendix C.

Sensitivity. Every metric and read-out falls monotonically along the *Training Degradation* ladder—ridge from 0.410 to 0.193, RSA from 0.249 to 0.056, dCKA from 0.196 to 0.036, iLISI from 0.816 to 0.358, top-5 from 0.623 to 0.098, while energy rises from 0.23 to 0.84 and the median rank from 3.4 to 85.8 (for both, lower indicates higher alignment)—so all seven are sensitive to the degradation of encoded information and each metric's scale is shown.

Specificity. RSA and dCKA score foil models close to, or above, the trained network; VAE metrics correctly flag them as dissimilar with the brain data. Four foil models, i.e., LabelCode, LabelCode-N, Carrier-B, Carrier-P, score at or above the reference ImageNet pretrained RN-50's RSA 0.249 and dCKA 0.196, ($p \leq 0.008$ only for Carrier-B and LabelCode-N). LabelCode is an artificial construction built only from each image's supercategory labels—the image itself is never processed—so whatever alignment the metric assigns rewards supercategory membership alone rather than any visual processing computations a network would make, and Carrier-B and Carrier-P, networks that classify merely from planted squares (Fig. 6). The Mimic, which

sees digit canvases (Fig. 5) rather than images, falls below, but is close to the reference model's RSA/dCKA. Two further RSA variants, correlation distance compared by Spearman's ρ_a (Schütt et al., 2023; Nili et al., 2014) and crossnobis distance (Walther et al., 2016) compared by a whitened cosine (Diedrichsen et al., 2020) are shown in Appendix I. For reference, in addition to the Degradation ladder, we include $RN50_{\mathrm{sc}}$, a model trained from scratch on NSD images with twelve multi-hot supercategory labels as its only supervision. Notably, for RSA and dCKA, $RN50_{\mathrm{sc}}$ (0.241, 0.188) is ranked below four of the five foil models. Ridge regression holds all five between 0.285 and 0.316, below the trained network's 0.410, but well above the untrained network's 0.193. VAE latent read-outs **floor the foil models**: iLISI places all five between 0.172 and 0.279, against the network's 0.816. Identification separates them most strongly from the reference RN-50: top-1 0.013–0.030 against 0.328 for the trained and 0.029 for the untrained one. When we perform the alignment per layer, and taking each network's best layer (Appendix H), RSA and dCKA do rank every foil model below the reference network, as choosing the best layer raises the trained RN-50's RSA from 0.249 to 0.382 while the foil models' barely move, yet, they still score the foil models at 65–88% of the reference's RSA (55–89% of its dCKA), and, notably, CARRIER-B as much as $RN50_{\mathrm{sc}}$.

5.2 SYNTHETIC CATEGORICAL REPRESENTATIONS

The *Synthetic Categorical Representations* grade expressivity and ordering on rungs that are ordered by construction (Section 4.1, Eq. 2, Fig 2). In Fig. 3, on the *Category Corruption*, $\mathrm{CAT}_{\mathrm{own}}$, $\mathrm{CAT}_{\mathrm{in}}$ and $\mathrm{CAT}_{\mathrm{out}}$ denote whose mean the image receives: its own category's, a sibling's within its supercategory, or a random category's outside its supercategory. On the *Supercategory Corruption*, $\text{S-CAT}_{\mathrm{own}}$ and $\text{S-CAT}_{\mathrm{donor}}$ have the correct supercategory-set mean and differ in whose variance is added back ($\text{S-CAT}_{\mathrm{own}}$ the image's own and $\text{S-CAT}_{\mathrm{donor}}$ another image's), and $\text{S-CAT}_{\mathrm{swap}}$ takes the mean from a random supercategory set (and variance from a random image of the swapped set).

Ridge is inexpressive and nearly flat throughout. Every construction scores 0.284–0.317, against 0.410 for the trained network and 0.193 for the untrained one: ridge falls by 0.217 on genuine degradation ($\alpha : 0 \to 1$) and by at most 0.033 (significant, but small) on the constructed representations. The *Category Corruption* ladder reads $0.308 \to 0.293$ and the *Supercategory Corruption* ladder $0.317 \to 0.284$. Whether the replacement category is a sibling within the supercategory or a randomly paired one, i.e., $\mathrm{CAT}_{\mathrm{in}}$ and $\mathrm{CAT}_{\mathrm{out}}$, the two category swaps score alike (0.294 and 0.293).

RSA and dCKA score the supercategory means close to the trained network, and barely separate own from donor variance. The supercategory mean with own variance, $\text{S-CAT}_{\mathrm{own}}$, reaches 97% and 90% of the reference on the RSA/dCKA, very similar to the $RN50_{\mathrm{sc}}$ (97%, 96% of reference), which was trained from scratch on NSD images with twelve multi-hot supercategory labels as its only supervision. Averaging within a supercategory label-set removes every within-group difference and keeps only the between-group structure, and that block structure is the part of the brain's geometry to which RSA/dCKA are most sensitive as also shown in Section 5.1. At each network's best layer (Appendix H), where RSA and dCKA do rank every foil model below the reference, they score $\mathrm{CAT}_{\mathrm{own}}$ and $\text{S-CAT}_{\mathrm{own}}$ above it (0.427 and 0.397 against 0.382 on RSA; 0.334 and 0.321 against 0.295 on dCKA), whereas the latent read-outs keep both well below it (iLISI 0.502 and 0.403 against 0.709; top-1 0.090 and 0.059 against 0.312). Replacing the image's own variation with a donor's ($\text{S-CAT}_{\mathrm{donor}}$) costs RSA only 0.015 and dCKA 0.012 (still at 91% and 84% of the reference): both metrics register this only weakly.

VAE metrics are expressive and ordered in the direction of construction Occupancy distinguishes a correct mean ($\mathrm{CAT}_{\mathrm{own}}$ 0.532) from a swapped one ($\mathrm{CAT}_{\mathrm{in}}$ 0.403, $\mathrm{CAT}_{\mathrm{out}}$ 0.393), and the image's own variance ($\text{S-CAT}_{\mathrm{own}}$ 0.483) from a donor's ($\text{S-CAT}_{\mathrm{donor}}$ 0.329) and from a swapped set ($\text{S-CAT}_{\mathrm{swap}}$ 0.162); identification does the same (top-5 0.196 against 0.080 and 0.073; 0.162 against 0.062 and 0.052). Occupancy barely separates $\mathrm{CAT}_{\mathrm{in}}$ from $\mathrm{CAT}_{\mathrm{out}}$, however, the median rank of the true image does, and orders every rung of both ladders ($27 < 65 < 111$ on the category ladder, $32 < 66 < 91$ on the supercategory ladder; chance 436.5).

The mechanism is visible in the VAE latent visualisation (UMAP-2d, (McInnes et al., 2018)), Figure 4. Because both systems are placed in one space, every difference the numbers summarise is visualised, as where each cloud sits and how much the two interpenetrate. The training objective rewards co-location wherever it supports cross-reconstruction, so the network's latents interpenetrate

the brain's, whereas the foil models land as compact, modular clusters, largely outside the brain's cloud. Along the *Corruption ladders*, in the order of construction, the model's points thin inside the brain's cloud and, e.g., for S-CAT$_{\text{swap}}$, begin to separate from it. The ordering the latent's metrics report is also the ordering the visualisations hint towards. Figure 9 in Appendix J recolours the same panels by the image's supercategory rather than by modality.

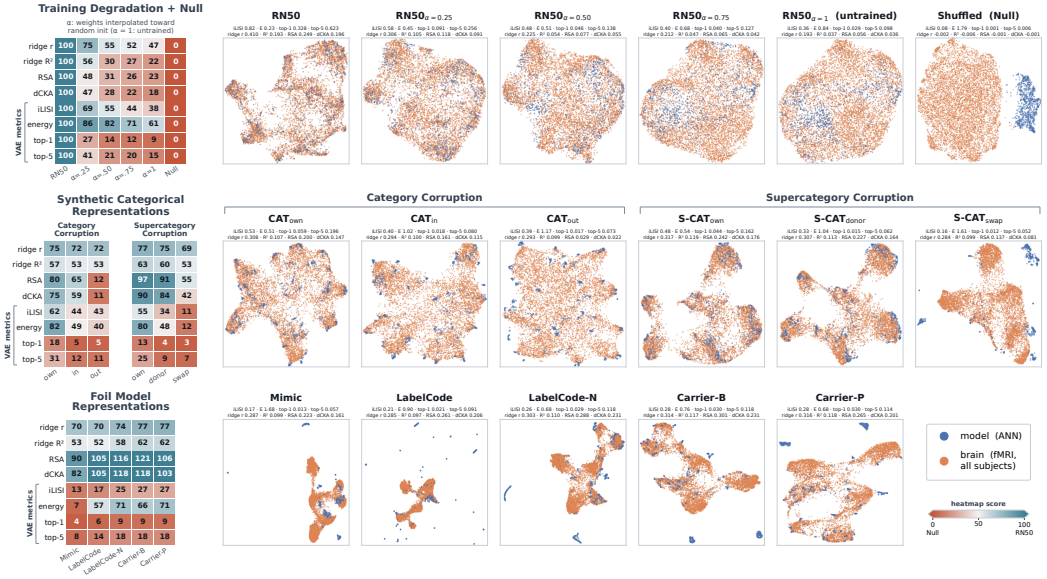


Figure 4: VAE latent visualisation (UMAP-2d). Panel titles carry distributional (iLISI, energy) and pairing-aware measures (top-1, top-5), computed in the 32-d latent, and ridge and RSA/dCKA on the raw data. The more brain and model representations interpenetrate, the higher their alignment. Foil models land as compact, modular clusters, largely outside the brain's cloud. Along the ladders, in the order of construction, the model's points thin inside the brain's cloud and, for the swapped supercategory set, begin to separate from it. The heatmaps give every metric as a percentage of the trained network, floor-corrected, so that 100 is RN50 and 0 the shuffled null.

6 Discussion

The measures read different properties of a representation. Ridge regression reads the partition of the stimuli that a representation induces, and it is nearly flat across them. In our experiments, RSA and dCKA are dominated by the highest-variance components and read coarse geometry, which in ventral and lateral visual cortex is dominated by category structure, so representations that keep the supercategory blocks (even through foil models, i.e., computations dissimilar to the visual processing computations the brain would make) are credited with most or all of the reference trained network's score. The *bidirectionally* shaped latent read-outs compare the two systems *at the level of distributional* and *stimulus-specific* differences, aspects of alignment overlooked by frequently used metrics. We suggest that the latent read-outs are informative as a set and should be read together, as they describe different aspects of alignment in the latent space. VAE metrics distinguished every *Foil model* and correctly ordered the coarse steps of the *Synthetic Categorical Representations*.

In our framework, each comparison requires training a joint embedding, and because the brain's latent representation is trained together with the model's, a score is a property of the model–brain pair rather than of a model against a fixed brain reference. Additionally, our experiments use one dataset (NSD) and one set of visual areas. Nonetheless, the latent read-outs offer a unique way of comparison and provide a powerful shared representation that captures both what is consistent among the representations and what varies. Model and brain are placed in one space, so their relation is observed in ways that the two separate structure spaces do not allow, enabling the exploration of different aspects of representational alignment. When adjudicating among multiple candidate computational models, i.e., foil and reference model, we show that the lingua franca of the shared latent space can help us recognize the more correct model, i.e., the reference.

Reproducibility statement

We have uploaded our code, anonymized in a zip file on Supplementary Material in Openreview.

AI use statement

We used a general-purpose LLM for literature triage, and for editing text and prose. All methods and experiments were conceptualized by the authors. LLMs were used for assistance in coding and LLM generated code was tested by the authors. AI-assisted text was reviewed and rewritten by the authors. We take responsibility for the final content of this work.

Ethics statement

This work uses the publicly available Natural Scenes Dataset (Allen et al., 2022), collected under its own ethical approvals; no new human-subject data were collected. We foresee no harmful applications; the work's aim is to improve the measurements used to compare models with brains.

References

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Itamar Avitan and Tal Golan. Model–behavior alignment under flexible evaluation: When the best-fitting model isn't the right one. *Advances in Neural Information Processing Systems*, 38:12081–12120, 2026.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):27755, 2016.

Nathan Cloos, Moufan Li, Markus Siegel, Scott Brincat, Earl Miller, Guangyu Robert Yang, and Christopher Cueva. Differentiable optimization of similarity scores between models and brains. In *International Conference on Learning Representations*, volume 2025, pp. 63438–63457, 2025.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1):9383, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis, 2017.

Jörn Diedrichsen, Eva Berlot, Marieke Mur, Heiko H Schütt, Mahdiyar Shahbazi, and Nikolaus Kriegeskorte. Comparing representational geometries using whitened unbiased-distance-matrix similarity. *arXiv preprint arXiv:2007.02789*, 2020.

Marin Dujmović, Jeffrey S Bowers, Federico Adolfi, and Gaurav Malhotra. *Obstacles to inferring mechanistic similarity using Representational Similarity Analysis*. Universitätsbibliothek Johann Christian Senckenberg, 2023.

Jenelle Feather, Meenakshi Khosla, N Murty, and Aran Nayebi. Brain-model evaluations need the neuroai turing test. *arXiv preprint arXiv:2502.16238*, 2025.

Alessandro T Gifford, Benjamin Lahner, Sari Saba-Sadiya, Martina G Vilas, Alex Lascelles, Aude Oliva, Kendrick Kay, Gemma Roig, and Radoslaw Martin Cichy. The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. *arXiv preprint arXiv:2301.03198*, 2023.

Tal Golan, JohnMark Taylor, Heiko Schütt, Benjamin Peters, Rowan P Sommers, Katja Seeliger, Adrien Doerig, Paul Linton, Talia Konkle, Marcel Van Gerven, et al. Deep neural networks are not a single hypothesis but a language for expressing computational hypotheses. *Behavioral and Brain Sciences*, 46, 2023.

Tal Golan, Heiko H Schütt, and Nikolaus Kriegeskorte. Making models disagree to learn how brains compute. *Nature Reviews Neuroscience*, pp. 1–22, 2026.

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*, pp. 770–778, 2016.

Larissa Höfling, Matthias Tangemann, Lotta Piefke, Susanne Keller, Matthias Bethge, and Katrin Franke. Only brains align with brains: Cross-region alignment patterns expose limits of normative models. In *International Conference on Learning Representations*, volume 2026, pp. 112924–112956, 2026.

Seyed-Mahdi Khaligh-Razavi, Linda Henriksson, Kendrick Kay, and Nikolaus Kriegeskorte. Fixed versus mixed rsa: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *Journal of Mathematical Psychology*, 76:184–197, 2017.

Meenakshi Khosla and Alex H Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. In *Proceedings of unireps: the first workshop on unifying representations in neural models*, pp. 326–341. PMLR, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. URL `https://arxiv.org/abs/1802.03426`.

Sabine Muzellec and Kohitij Kar. Reverse predictivity for bidirectional comparison of neural networks and biological brains. *Nature Machine Intelligence*, 8(3):474–488, 2026.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things. *Uncovering How Neural Network Representations Vary with Width and Depth*, pp. 1–25, 2020.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014.

Angeliki Papathanasiou, Jascha Achterberg, Thomas E Nichols, and Rui Ponte Costa. Task-guided cross-subject latent alignment: a multi-encoder-decoder VAE. arXiv:2606.15989, 2026. In Proceedings of the 9th Conference on Cognitive Computational Neuroscience (CCN), New York, NY, USA, 2026.

Jacob S Prince, Colin Conwell, George A Alvarez, and Talia Konkle. A case for sparse positive alignment of neural systems. In *ICLR 2024 Workshop on Representational Alignment*, 2024.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020.

Heiko H Schütt, Alexander D Kipnis, Jörn Diedrichsen, and Nikolaus Kriegeskorte. Statistical inference on representational geometries. *Elife*, 12:e82566, 2023.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(5), 2012.

Ansh Soni, Sudhanshu Srivastava, Meenakshi Khosla, and Konrad Paul Kording. Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. biorxiv, 2024.

Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of cognitive neuroscience*, 33(10):2044–2064, 2021.

Jianle Sun, Chaoqi Liang, Ran Wei, Peng Zheng, Lei Bai, Wanli Ouyang, Hongliang Yan, and Peng Ye. scmrdr: A scalable and flexible framework for unpaired single-cell multi-omics data integration. *Advances in Neural Information Processing Systems*, 38:154538–154565, 2026.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

Alexander Walther, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, and Jörn Diedrichsen. Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage*, 137:188–200, 2016.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in neural information processing systems*, 34:4738–4750, 2021.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

A Full training objective

Parameters. The model representation has encoder E_{ϕ_m} and decoder D_{θ_m}, and subject $s = 1, \ldots, S$ ($S = 8$) has encoder E_{ϕ_s} and decoder D_{θ_s}; $\phi = (\phi_\mathrm{m}, \phi_1, \ldots, \phi_S)$ collects the parameters of all encoders and $\theta = (\theta_\mathrm{m}, \theta_1, \ldots, \theta_S)$ those of all decoders. Each encoder outputs the mean and log-variance of a diagonal-Gaussian posterior over the shared 32-dimensional latent, $q_{\phi_\mathrm{m}}(\mathbf{z} \mid \mathbf{x})$ for the model and $q_{\phi_s}(\mathbf{z} \mid \mathbf{b}_s)$ for subject s.

Data. Let $\mathcal{T}_s$ be the training images of subject s and $\mathcal{T} = \bigcup_s \mathcal{T}_s$ (62,551 training images). Only images seen by a single subject are used for training, so image $i \in \mathcal{T}_s$ comes with the model features $\mathbf{x}_i$ and one brain response, $\mathbf{b}_{s,i}$.

No subject–subject pairs. A term pairing two subjects would need images with a response from both. The only such images are NSD's 1,000 shared images, which are held out of training (the 872 of them seen by all eight subjects are the test images). A subject's decoder therefore never reconstructs from another subject's latent, and the subjects share the latent space only through the model's pathways, which are trained against every subject.

B Training Architecture and read-out

Data. Every VAE is trained on all eight NSD subjects (Allen et al., 2022), using the vertices of the ventral, lateral, midventral and midlateral streams in the Algonauts 2023 release (20,726–20,736 per subject). Training uses only the images each subject saw alone (69,502 images, each paired with the

one subject who saw it), so no pathway, on either side, sees any of the $1{,}000$ shared images. A fixed random 10% of the training images ($6{,}951$; 828–919 per subject; the same for every representation and seed) is held out for validation, and the remaining $62{,}551$ train the network. Every read-out is computed on the 872 shared images seen by all eight subjects. Model features are z-scored per feature over all training images.

Architecture. The model representation and each subject have their own encoder and decoder (nine of each). An encoder has two fully connected hidden layers of 512 and 256 units, each followed by LayerNorm, ReLU and dropout (0.3), and two linear heads for the mean and log-variance of a diagonal Gaussian over the shared 32-dimensional latent. A decoder mirrors it (256 and 512 units) and ends in a linear layer to the output dimension, bounded by a $\tanh$ scaled by a learned factor.

Objective and optimisation. The loss objective is given in Eq. equation 1 and Appendix A. We use Adam with learning rate 10^{-4} and no weight decay, mini-batches of 128 images and gradient-norm clipping at 1, for up to 60 epochs. A separate VAE is trained for every representation and feature basis: three seeds (1, 10, 100) in the $51{,}048$-dimensional basis, one in the semantic and best-layer bases. The seed sets the initialisation, dropout, latent samples and batch order.

Choosing the read-out epoch. After every epoch we store the encoder means of the validation images. The epoch at which a run is read out is chosen from validation data alone, by one rule applied to every run, seed and basis. The rule tracks two signals: the validation objective (the full loss on the validation split) and validation identification (the top-5 accuracy and median rank of each validation image's brain latent among the latents of that subject's own validation images, model to brain). A run counts as trained up to S, the first epoch at which both signals have gone 10 epochs without improving: the objective without a new minimum, and identification without top-5 rising by more than 0.005 or the median rank falling by more than 2%. With t_{obj} the epoch of the lowest validation objective, and t_{id} the last epoch at which identification improved, the run is read at

$$t^* = \max\left(t_{\mathrm{obj}},\, t_{\mathrm{id}}\right). \tag{4}$$

The objective alone can reach its minimum while the correspondence between model and brain latents is still improving; taking the later of the two epochs reads a run only once both have settled.

Read-outs. All latent read-outs use the encoder means at t^*. They are computed per subject, between the 872 model latents and that subject's 872 brain latents, and averaged over the eight subjects. iLISI uses the $k = 30$ nearest neighbours of each point among the $1{,}744$ pooled points; energy distance uses all pairwise Euclidean distances; identification ranks, for every image, the true brain latent among all 872 by Euclidean distance to the model latent (top-1: rank 1; top-5: rank ≤ 5; median rank over images, chance 436.5). The UMAP projections in the figures (30 neighbours, minimum distance 0.1) are fitted jointly on the model latents and all eight subjects' brain latents.

C CATALOGUE OF REPRESENTATIONS

Table 1, summarizes the rungs for (i) Category and (ii) Supercategory Corruption Ladders. Table 2 lists every constructed representation/model in the paper, what enters its computation and how it is built. In the main basis each is a $70{,}502 \times 51{,}048$ matrix, one row per image seen by any subject; the semantic and best-layer bases (Appendices G and H) build them from their own features.

Table 1: The rungs of the two ladders as instances of Eq. equation 2, written as (group average received, image whose variation is added).

Corruption	name = $(Z, z(k = 10))$
Category	$\mathrm{CAT}_{\mathrm{own}}$ = (own, i_ℓ), $\mathrm{CAT}_{\mathrm{in}}$ = (sibling–within supercategory, j_p), $\mathrm{CAT}_{\mathrm{out}}$ = (random, j_p)
Supercategory	S-$\mathrm{CAT}_{\mathrm{own}}$ = (own, i_ℓ), S-$\mathrm{CAT}_{\mathrm{donor}}$ = (own, i_m), S-$\mathrm{CAT}_{\mathrm{swap}}$ = (random, j_p)

Table 2: Every constructed representation/model in the main results. *Input*: what enters the computation. Ladder rungs are built from RN50's features as a group mean plus one image's coefficients on the top k principal components of the within-group residuals, pooled over all training images ($k = 10$ in the main basis, $k = 3$ in the semantic basis). An image's category is its rarest COCO label, and "its label-set" is the set of supercategories it contains (829 distinct sets).

Name	Input	Construction
Networks		
RN50	image	torchvision ResNet-50 with ImageNet-1K weights; the reference.
RN50_α	image	weights $(1-\alpha)W_{\text{trained}} + \alpha W_{\text{init}}$, BatchNorm statistics recomputed; $\alpha = 1$ is untrained.
$RN50_{\text{sc}}$	image	ResNet-50 trained from scratch on NSD with the 12 supercategory labels as its only supervision (mAP 0.81).
CARRIER-B	image + squares	ResNet-50 trained from scratch on images carrying one fixed-position square per supercategory present, placed so that their distances mirror the brain's supercategory geometry ($r = 0.93$); mAP 1.00, 0.24 without the squares.
CARRIER-P	image + squares	as CARRIER-B, with the positions permuted across supercategories ($r = -0.08$); mAP 1.00, 0.26 without the squares.
MIMIC	label canvas	RN50 on canvas of MNIST digits, one per supercategory present (indoor and outdoor excluded; Fig. 5).
Codes		
LABELCODE	labels	$s\, L_i B$ plus noise along RN50's residual directions (Eq. 3), with $L_i \in \{0,1\}^{12}$ the supercategory labels and B fixed Gaussian anchors.
LABELCODE-N	image	as LABELCODE, with L_i the 12 outputs of a ResNet-50 fine-tuned from ImageNet on the supercategory task (mAP 0.89).
Category Corruption		
CAT_{own}	RN50 features	own category mean + own variance from 10 PC components.
CAT_{in}	RN50 features	the mean of a sibling category in the same supercategory (fixed pairs) + the variance of a random image of that category.
CAT_{out}	RN50 features	as CAT_{in}, with the 80 categories paired at random (38 of the 40 pairs cross supercategories).
Supercategory Corruption		
$\text{S-CAT}_{\text{own}}$	RN50 features	own label-set mean + own variance.
$\text{S-CAT}_{\text{donor}}$	RN50 features	own label-set mean + the variance of another image of the same label-set (an image alone in its set keeps its own).
$\text{S-CAT}_{\text{swap}}$	RN50 features	the mean of the label-set obtained by relabelling the supercategories with one fixed 12-cycle (the nearest observed set where that one does not occur) + the variance of a random image of that set.
NULL	RN50 features	RN50's rows permuted across images by a derangement, so that no image keeps its own.

D CARRIER AND MIMIC EXPERIMENT

Fig. 5 and Fig. 6 show examples of how the CARRIER-P and MIMIC inputs were built.

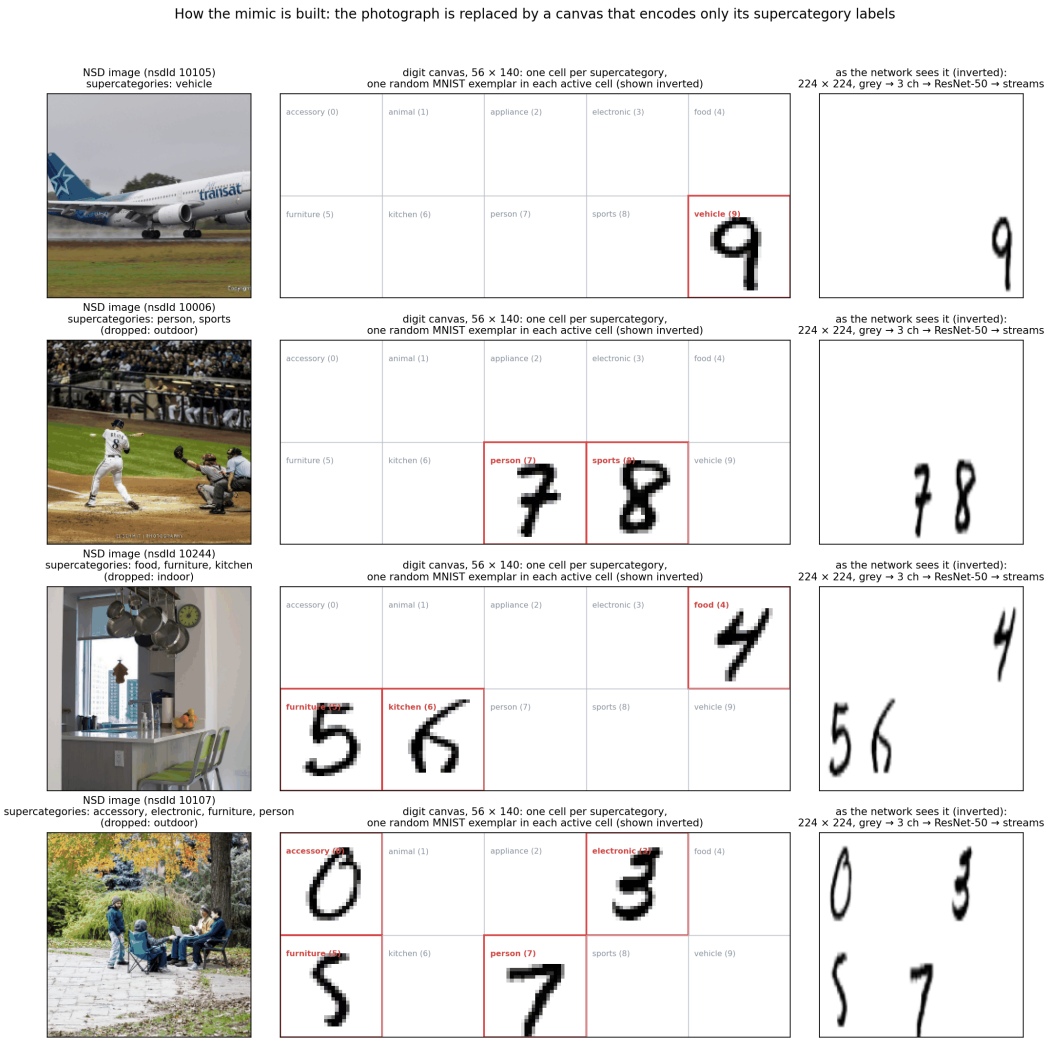


Figure 5: How the MIMIC input is built, for the same four NSD images as Fig. 6. Left: the NSD image and its COCO supercategory labels; the scene-level labels *indoor* and *outdoor* are dropped. Middle: a 56×140 canvas with one cell per remaining supercategory, in alphabetical order and assigned the digits 0–9 (accessory 0, . . . , vehicle 9); each supercategory present in the image places a random MNIST sample of its digit in its cell, and every other cell stays blank. Right: the canvas as passed to the network, resized to 224×224 and replicated to three channels. The canvas carries the image's supercategory multilabels and nothing else from the image.



Figure 6: How the planted-carrier network is fed. Left: NSD images with their multihot supercategory labels (the same four images as Fig. 5). Middle: the CARRIER-P layout—one fixed 16×16-pixel position per supercategory, the same for every image; here the positions are chosen randomly, so the layout carries none of the brain's supercategory geometry. Right: the image as the network sees it, with one carrier (black square, white interior) per supercategory it contains, planted before ImageNet normalisation exactly as in training. Both carrier networks reach mAP 1.00 with the squares and 0.24–0.26 without them, so they classify the images from the carriers, i.e, planted pixels, alone.

E ALIGNMENT SCORES TABLE

Figure 3 shows the alignment scores of every representation per subject. Table 3 gives the means over the eight subjects behind each box, with the latent read-outs as mean ± s.d. over the three VAE training seeds.

Table 3: **Alignment scores of every representation.** Ridge, RSA and dCKA are computed on the features; the latent read-outs in the VAE's 32-d co-embeddings, as mean ± s.d. over three VAE training seeds. Identification (top-1, top-5, median rank) is model-to-brain over the 872 shared test images seen by all subjects (chance 0.001, 0.006 and 436.5). Bold: at or above the trained network.

Condition	ridge r	ridge R^2	RSA	dCKA	iLISI (↑)	energy (↓)	top-1 (↑)	top-5 (↑)	med. rank (↓)
Training Degradation									
$\text{RN50}_{\alpha=0}$	**0.410**	**0.193**	0.249	0.196	**0.816 ± .015**	**0.23 ± .03**	**0.328 ± .006**	**0.623 ± .007**	**3.4 ± 0.1**
$\text{RN50}_{\alpha=0.25}$	0.306	0.105	0.118	0.091	0.584 ± .022	0.45 ± .04	0.091 ± .003	0.256 ± .003	21.9 ± 0.9
$\text{RN50}_{\alpha=0.50}$	0.225	0.054	0.077	0.055	0.484 ± .006	0.51 ± .03	0.046 ± .002	0.138 ± .003	50.9 ± 0.6
$\text{RN50}_{\alpha=0.75}$	0.212	0.047	0.065	0.042	0.405 ± .008	0.68 ± .04	0.040 ± .001	0.127 ± .004	61.1 ± 0.8
$\text{RN50}_{\alpha=1}$	0.193	0.037	0.056	0.036	0.358 ± .016	0.84 ± .04	0.029 ± .002	0.098 ± .004	85.8 ± 1.1
$RN50_{\text{sc}}$	0.386	0.167	0.241	0.188	0.674 ± .013	0.40 ± .05	0.196 ± .003	0.458 ± .007	7.3 ± 0.2
Category Corruption									
CAT_{own}	0.308	0.107	0.200	0.147	0.532 ± .010	0.51 ± .05	0.059 ± .002	0.196 ± .002	26.6 ± 0.9
CAT_{in}	0.294	0.100	0.161	0.115	0.403 ± .029	1.02 ± .16	0.018 ± .001	0.080 ± .004	65.4 ± 1.2
CAT_{out}	0.293	0.099	0.029	0.022	0.393 ± .012	1.17 ± .04	0.017 ± .001	0.073 ± .002	110.9 ± 2.4
Supercategory Corruption									
$\text{S-CAT}_{\text{own}}$	0.317	0.119	0.242	0.176	0.483 ± .020	0.54 ± .03	0.044 ± .002	0.162 ± .004	32.4 ± 0.4
$\text{S-CAT}_{\text{donor}}$	0.307	0.113	0.227	0.164	0.329 ± .065	1.04 ± .17	0.015 ± .002	0.062 ± .002	66.4 ± 1.2
$\text{S-CAT}_{\text{swap}}$	0.284	0.099	0.137	0.081	0.162 ± .012	1.61 ± .10	0.012 ± .001	0.052 ± .004	90.6 ± 2.3
Foil Models									
LABELCODE	0.285	0.097	**0.261**	**0.206**	0.206 ± .007	0.90 ± .06	0.021 ± .001	0.091 ± .001	53.4 ± 0.7
LABELCODE-N	0.303	0.110	**0.288**	**0.231**	0.261 ± .012	0.68 ± .08	0.029 ± .002	0.118 ± .002	42.8 ± 0.1
MIMIC	0.287	0.099	0.223	0.161	0.172 ± .018	1.68 ± .06	0.013 ± .001	0.057 ± .002	73.0 ± 1.5
CARRIER-B	0.314	0.117	**0.301**	**0.231**	0.279 ± .012	0.76 ± .02	0.030 ± .001	0.118 ± .003	43.1 ± 0.5
CARRIER-P	0.316	0.118	**0.265**	**0.201**	0.275 ± .038	0.68 ± .04	0.030 ± .004	0.114 ± .005	44.3 ± 1.0
NULL	−0.002	−0.006	−0.001	−0.001	0.077 ± .038	1.79 ± .29	0.001 ± .001	0.006 ± .001	436.2 ± 3.2

F STATISTICAL TESTS

Procedure. A *contrast* compares two representations, a and b, on one metric; its value is the difference between their scores, $a - b$ (e.g. LABELCODE − RN50 on RSA). We test 16 primary contrasts on seven metrics: each foil model against each of the two reference networks, RN50 and $RN50_{\text{sc}}$, and each step between adjacent rungs of the two Corruption ladders (Tables 4 and 5). *Across subjects*, each contrast gives one difference per subject (for the latent read-outs, after averaging over the three VAE seeds). We test whether the mean of the eight differences is zero with a two-sided paired t-test, and correct the p-values over the 16 contrasts of each metric (Holm). p-values are in parentheses in Table 4. We also report how many of the eight subjects show a difference in the direction of the mean and, because all subjects share one VAE per seed, whether the three seeds do. *Across images*, we ask whether the difference would also hold for a new sample of images: we resample the 872 test images with replacement, with the same draws for every representation so that the differences are paired, and report 95% percentile intervals (Table 5; for RSA we follow Schütt et al. (2023)). We do this for RSA and the four latent read-outs; ridge and dCKA are tested across subjects only.

Results. Every latent read-out scores every foil model below both RN50 and $RN50_{\text{sc}}$ (Table 4; corrected $p < 0.001$ in all 40 entries, in all eight subjects and all three seeds), and so does ridge ($p < 0.001$). RSA and dCKA score CARRIER-B and LABELCODE-N above RN50 (RSA $p = 0.001$ and 0.006; dCKA $p = 0.008$ for both), LABELCODE and CARRIER-P level with it (RSA $p = 0.21$ and 0.11; dCKA $p = 0.22$ and 0.34), and MIMIC below it. Against $RN50_{\text{sc}}$, CARRIER-B, CARRIER-P, LABELCODE-N and LABELCODE score higher on RSA ($p \leq 0.025$) and dCKA ($p \leq 0.032$). Along both Corruption ladders, every step worsens every metric (corrected $p \leq 0.021$), with two exceptions: RSA does not separate RN50 from $\text{S-CAT}_{\text{own}}$ ($p = 0.21$), and top-5 does not separate CAT_{in} from CAT_{out} ($p = 0.14$). Ridge registers every step, although after

the first step of each ladder the steps are small (0.002–0.023 in r). The relabelling from CAT_{in} to CAT_{out}, which changes only which category mean an image receives, is registered strongly by RSA, dCKA, energy and median rank ($p < 0.001$, all eight subjects), weakly by iLISI ($+0.010$, $p = 0.021$; seven of eight subjects, two of three seeds) and ridge ($+0.002$, $p = 0.018$; seven of eight subjects), and not by top-5. Every step of the RN50_{α} ladder (not tabulated) worsens all seven metrics in all eight subjects (uncorrected $p < 0.001$). The image bootstrap gives the same verdict as the corrected subject-level test in all 80 entries it covers (Table 5), but for RSA the image-level uncertainty is large compared with the subject-level one: two RSA differences that all eight subjects show only just exclude zero across images, $\text{S-CAT}_{\text{own}} - \text{S-CAT}_{\text{donor}}$ ($+0.015$, $[+0.0004, +0.028]$) and $\text{LABELCODE} - RN50_{\text{sc}}$ ($+0.020$, $[+0.001, +0.038]$).

Table 4: **Differences across subjects** for the 16 primary contrasts in the main basis. Each row compares two representations a and b: a foil model and a reference network, or two adjacent rungs of a Corruption ladder (a is the less corrupted one). Each entry is the mean over the eight subjects of the score of a minus the score of b (the scores themselves are in Table 3): a positive entry means that a scores higher, i.e. better on the metrics marked ↑ and worse on those marked ↓. In parentheses: the p-value of a two-sided paired t-test over the eight subjects, Holm-corrected over the 16 contrasts of each column; bold: corrected $p < 0.05$. Superscript $k/8$: the number of subjects whose difference has the sign of the mean, shown only when it is not $8/8$. The latent read-outs (iLISI, energy, top-5, median rank) are averaged over the three VAE training seeds before testing; †: the three seed-level differences do not all have the sign of the mean.

Contrast $a - b$	ridge r (↑)	RSA (↑)	dCKA (↑)	iLISI (↑)	energy (↓)	top-5 (↑)	med. rank (↓)
Foil models vs. RN50							
LABELCODE − RN50	**−0.125** (< 0.001)	+0.012$^{5/8}$ (0.21)	+0.010$^{5/8}$ (0.22)	**−0.610** (< 0.001)	**+0.67** (< 0.001)	**−0.532** (< 0.001)	**+50.0** (< 0.001)
LABELCODE-N − RN50	**−0.106** (< 0.001)	**+0.039** (0.006)	**+0.035** (0.008)	**−0.555** (< 0.001)	**+0.45** (< 0.001)	**−0.505** (< 0.001)	**+39.4** (< 0.001)
MIMIC − RN50	**−0.123** (< 0.001)	**−0.026** (0.025)	**−0.035** (0.003)	**−0.644** (< 0.001)	**+1.45** (< 0.001)	**−0.566** (< 0.001)	**+69.6** (< 0.001)
CARRIER-B − RN50	**−0.096** (< 0.001)	**+0.052** (0.001)	**+0.035** (0.008)	**−0.537** (< 0.001)	**+0.53** (< 0.001)	**−0.505** (< 0.001)	**+39.7** (< 0.001)
CARRIER-P − RN50	**−0.094** (< 0.001)	+0.016$^{5/8}$ (0.11)	+0.005$^{4/8}$ (0.34)	**−0.541** (< 0.001)	**+0.45** (< 0.001)	**−0.508** (< 0.001)	**+40.9** (< 0.001)
Foil models vs. $RN50_{\text{sc}}$							
LABELCODE − $RN50_{\text{sc}}$	**−0.101** (< 0.001)	**+0.020** (0.025)	**+0.019** (0.021)	**−0.468** (< 0.001)	**+0.50** (< 0.001)	**−0.367** (< 0.001)	**+46.0** (< 0.001)
LABELCODE-N − $RN50_{\text{sc}}$	**−0.083** (< 0.001)	**+0.047** (< 0.001)	**+0.044** (0.003)	**−0.413** (< 0.001)	**+0.28** (< 0.001)	**−0.340** (< 0.001)	**+35.5** (< 0.001)
MIMIC − $RN50_{\text{sc}}$	**−0.099** (< 0.001)	**−0.018** (0.025)	**−0.027** (0.003)	**−0.502** (< 0.001)	**+1.28** (< 0.001)	**−0.401** (< 0.001)	**+65.7** (< 0.001)
CARRIER-B − $RN50_{\text{sc}}$	**−0.072** (< 0.001)	**+0.060** (< 0.001)	**+0.044** (0.002)	**−0.396** (< 0.001)	**+0.35** (< 0.001)	**−0.340** (< 0.001)	**+35.8** (< 0.001)
CARRIER-P − $RN50_{\text{sc}}$	**−0.071** (< 0.001)	**+0.024** (0.008)	**+0.013**$^{7/8}$ (0.032)	**−0.399** (< 0.001)	**+0.27** (< 0.001)	**−0.343** (< 0.001)	**+37.0** (< 0.001)
Category Corruption							
RN50 − CAT_{own}	**+0.102** (< 0.001)	**+0.049** (< 0.001)	**+0.049** (< 0.001)	**+0.284** (< 0.001)	**−0.28** (< 0.001)	**+0.427** (< 0.001)	**−23.2** (< 0.001)
$\text{CAT}_{\text{own}} - \text{CAT}_{\text{in}}$	**+0.014** (< 0.001)	**+0.039** (< 0.001)	**+0.032** (< 0.001)	**+0.129** (< 0.001)	**−0.51** (< 0.001)	**+0.116** (< 0.001)	**−38.8** (< 0.001)
$\text{CAT}_{\text{in}} - \text{CAT}_{\text{out}}$	**+0.002**$^{7/8}$ (0.018)	**+0.132** (< 0.001)	**+0.093** (< 0.001)	**+0.010**$^{7/8\,\dagger}$ (0.021)	**−0.15** (< 0.001)	+0.006$^{4/8}$ (0.14)	**−45.6** (< 0.001)
Supercategory Corruption							
RN50 − $\text{S-CAT}_{\text{own}}$	**+0.093** (< 0.001)	+0.007$^{5/8}$ (0.21)	**+0.020** (0.001)	**+0.333** (< 0.001)	**−0.31** (< 0.001)	**+0.461** (< 0.001)	**−29.0** (< 0.001)
$\text{S-CAT}_{\text{own}} - \text{S-CAT}_{\text{donor}}$	**+0.010** (< 0.001)	**+0.015** (< 0.001)	**+0.012** (0.002)	**+0.154** (< 0.001)	**−0.50** (< 0.001)	**+0.100** (< 0.001)	**−34.0** (< 0.001)
$\text{S-CAT}_{\text{donor}} - \text{S-CAT}_{\text{swap}}$	**+0.023** (< 0.001)	**+0.090** (< 0.001)	**+0.083** (< 0.001)	**+0.167** (< 0.001)	**−0.56** (< 0.001)	**+0.010** (0.003)	**−24.3** (< 0.001)

Table 5: **Differences across images**: 95% bootstrap intervals of the differences in Table 4 (same rows and signs). The 872 test images are resampled with replacement, with the same draws for every representation, subject and VAE seed, so that the differences are paired: 2,000 draws for the latent read-outs, averaged over the eight subjects and then over the three seeds, and a separate set of 2,000 draws for RSA, averaged over the eight subjects. On each draw, each image's rank among the 872 candidates and its iLISI neighbourhood ($K = 30$) keep their full-set values, energy distance is recomputed as a weighted V-statistic, and RSA leaves out pairs of the same image. Ridge r and dCKA are not bootstrapped. In every entry, the interval excludes zero exactly when the corrected p in Table 4 is below 0.05.

Contrast $a - b$	RSA (↑)	iLISI (↑)	energy (↓)	top-5 (↑)	med. rank (↓)
Foil models vs. RN50					
LABELCODE − RN50	$[-0.010, +0.031]$	$[-0.625, -0.595]$	$[+0.64, +0.72]$	$[-0.548, -0.516]$	$[+46.9, +54.0]$
LABELCODE-N − RN50	$[+0.017, +0.058]$	$[-0.571, -0.538]$	$[+0.43, +0.48]$	$[-0.521, -0.490]$	$[+36.4, +42.5]$
MIMIC − RN50	$[-0.045, -0.007]$	$[-0.658, -0.630]$	$[+1.39, +1.53]$	$[-0.582, -0.550]$	$[+65.7, +73.4]$
CARRIER-B − RN50	$[+0.031, +0.072]$	$[-0.553, -0.522]$	$[+0.51, +0.55]$	$[-0.521, -0.489]$	$[+37.1, +42.5]$
CARRIER-P − RN50	$[-0.005, +0.034]$	$[-0.557, -0.525]$	$[+0.43, +0.47]$	$[-0.524, -0.493]$	$[+38.3, +43.8]$
Foil models vs. $RN50_{\mathrm{sc}}$					
LABELCODE − $RN50_{\mathrm{sc}}$	$[+0.001, +0.038]$	$[-0.485, -0.453]$	$[+0.46, +0.54]$	$[-0.383, -0.351]$	$[+42.9, +49.9]$
LABELCODE-N − $RN50_{\mathrm{sc}}$	$[+0.028, +0.064]$	$[-0.431, -0.396]$	$[+0.25, +0.31]$	$[-0.356, -0.324]$	$[+32.5, +38.3]$
MIMIC − $RN50_{\mathrm{sc}}$	$[-0.035, -0.001]$	$[-0.517, -0.487]$	$[+1.21, +1.36]$	$[-0.417, -0.386]$	$[+61.7, +69.3]$
CARRIER-B − $RN50_{\mathrm{sc}}$	$[+0.041, +0.079]$	$[-0.412, -0.380]$	$[+0.33, +0.38]$	$[-0.357, -0.325]$	$[+33.2, +38.4]$
CARRIER-P − $RN50_{\mathrm{sc}}$	$[+0.006, +0.041]$	$[-0.416, -0.383]$	$[+0.26, +0.30]$	$[-0.359, -0.328]$	$[+34.3, +39.8]$
Category Corruption					
RN50 − $\mathrm{CAT}_{\mathrm{own}}$	$[+0.036, +0.063]$	$[+0.278, +0.290]$	$[-0.30, -0.27]$	$[+0.411, +0.443]$	$[-25.1, -21.4]$
$\mathrm{CAT}_{\mathrm{own}}$ − $\mathrm{CAT}_{\mathrm{in}}$	$[+0.024, +0.055]$	$[+0.121, +0.137]$	$[-0.54, -0.48]$	$[+0.105, +0.128]$	$[-42.7, -35.4]$
$\mathrm{CAT}_{\mathrm{in}}$ − $\mathrm{CAT}_{\mathrm{out}}$	$[+0.111, +0.151]$	$[+0.001, +0.019]$	$[-0.19, -0.11]$	$[-0.001, +0.013]$	$[-52.9, -37.7]$
Supercategory Corruption					
RN50 − $\text{S-CAT}_{\mathrm{own}}$	$[-0.006, +0.021]$	$[+0.327, +0.339]$	$[-0.33, -0.30]$	$[+0.445, +0.475]$	$[-31.4, -27.0]$
$\text{S-CAT}_{\mathrm{own}}$ − $\text{S-CAT}_{\mathrm{donor}}$	$[+0.0004, +0.028]$	$[+0.144, +0.164]$	$[-0.54, -0.47]$	$[+0.091, +0.110]$	$[-36.8, -31.0]$
$\text{S-CAT}_{\mathrm{donor}}$ − $\text{S-CAT}_{\mathrm{swap}}$	$[+0.072, +0.107]$	$[+0.156, +0.177]$	$[-0.61, -0.52]$	$[+0.003, +0.017]$	$[-28.5, -20.1]$

G SEMANTICALLY SAMPLED LAYERS

The main analyses read each RN-50 representations through sixteen modules sampled at fixed intervals (Section 3). Here we read it at semantically meaningful positions instead: the output of each of the sixteen residual (bottleneck) blocks and the global average pool. Each block is projected by the same fixed sparse random projection to $3{,}500$ dimensions and all of them are concatenated, giving a $61{,}548$-dimensional feature vector per image. Every representation and construction is rebuilt in this basis, and a separate VAE is trained on it with the settings of Appendix B.

Matched ladders. A rung's corruption strength is set by the share of the (z-scored) feature variance that is group identity rather than image-level variation, and the top k within-group components retain a different share in different bases. We therefore choose k so that each ladder keeps the share of the $51{,}048$-dimensional ladders at $k = 10$, so that the resuts are comparable to the main $51{,}048$-dimensional feature base in the main section of the paper. Computed on the features alone: $k = 3$ gives 65.4% of the variance between supercategory sets for S-CAT_{own} (against 65.6%) and 62.5% between categories for CAT_{own} (against 61.7%), whereas $k = 10$ would leave 45% and 41%, making every rung a stronger corruption than in the main analyses.

Results. The results replicate (Table 6, Figure 7). Ridge is flat across every construction (0.284–0.318, against 0.406 for the network and 0.196 untrained). RSA and dCKA again score four of the five foil models above the trained network (LABELCODE $0.249/0.192$, LABELCODE-N $0.273/0.214$, CARRIER-B $0.307/0.236$ and CARRIER-P $0.265/0.201$, against $0.234/0.184$) and the MIMIC close to it ($0.224/0.159$), and they barely separate own from donor variance (0.203 against 0.192 RSA). The latent read-outs place every foil model below the untrained network on occupancy (0.166–0.254, against 0.352) and at it on identification (top-1 0.017–0.029, against 0.030), while $RN50_{\text{sc}}$, which RSA and dCKA place below four of the five foil models ($0.222/0.172$), is second only to the trained network on every latent read-out (iLISI 0.664, top-1 0.175, median rank 8.3).

Table 6: **Semantically sampled basis** ($61{,}548$-d: stem, the sixteen residual blocks and the average pool). One training seed, each run read at the epoch selected by the stopping protocol (Appendix B). Ladders at the matched $k = 3$. Columns as in Table 3; identification over the 872 shared test images (chance 0.001, 0.006 and 436.5). Bold: at or above the trained network.

Condition	ridge r	ridge R^2	RSA	dCKA	iLISI (↑)	energy (↓)	top-1 (↑)	top-5 (↑)	med. rank (↓)
RN50	**0.406**	**0.183**	0.234	0.184	**0.772**	**0.30**	**0.308**	**0.609**	**3.9**
$\text{RN50}_{\alpha=1}$	0.196	0.036	0.052	0.032	0.352	0.92	0.030	0.104	83.9
$RN50_{\text{sc}}$	0.384	0.164	0.222	0.172	0.664	0.40	0.175	0.428	8.3
Category Corruption									
CAT_{own}	0.305	0.105	0.175	0.126	0.554	0.53	0.058	0.195	26.5
CAT_{in}	0.295	0.100	0.135	0.092	0.411	0.86	0.018	0.073	77.2
CAT_{out}	0.294	0.100	0.026	0.018	0.399	0.87	0.016	0.068	132.1
Supercategory Corruption									
S-CAT_{own}	0.315	0.117	0.203	0.144	0.466	0.47	0.041	0.162	31.0
S-$\text{CAT}_{\text{donor}}$	0.308	0.113	0.192	0.134	0.260	0.87	0.013	0.061	71.1
S-CAT_{swap}	0.286	0.100	0.118	0.070	0.277	0.74	0.012	0.050	97.9
Foil Models									
MIMIC	0.287	0.100	0.224	0.159	0.166	1.47	0.017	0.067	62.9
LABELCODE	0.284	0.097	**0.249**	**0.192**	0.195	1.96	0.019	0.078	59.1
LABELCODE-N	0.303	0.109	**0.273**	**0.214**	0.218	1.72	0.021	0.092	49.0
CARRIER-B	0.316	0.119	**0.307**	**0.236**	0.240	1.08	0.027	0.106	45.8
CARRIER-P	0.318	0.119	**0.265**	**0.201**	0.254	0.86	0.029	0.113	46.0
NULL	-0.001	-0.006	0.000	-0.001	0.105	1.68	0.002	0.007	433.9

H BEST LAYER

Here every network is read at a single layer: the block with the highest RSA to the brain among all of its convolutional, rectification and pooling blocks, projected by the same fixed sparse random projection to $3{,}500$ dimensions. This is `layer4.1.conv2` for RN50 and

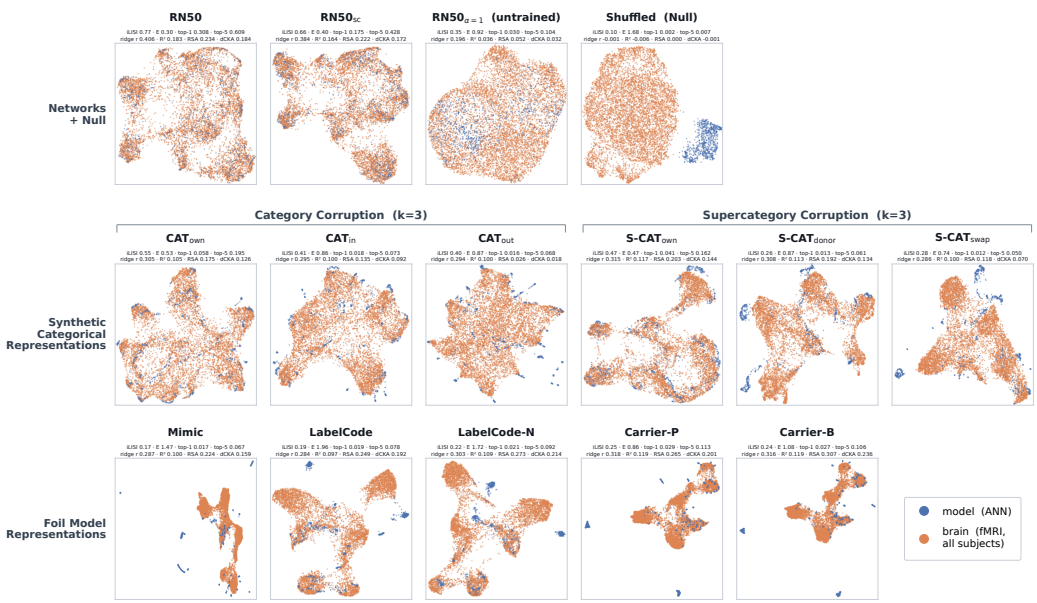


Figure 7: VAE latent visualisation (UMAP-2d) in the semantically sampled basis, one seed, each run at its selected epoch. Rows: training degradation and null; category ladder; supercategory ladder (both at $k = 3$); adversaries. Colours and panel titles as in Figure 4.

$RN50_{\text{sc}}$, `layer1.0.downsample.0` for the untrained network, `layer4.0.downsample.0` for CARRIER-B, `layer2.2.relu` for CARRIER-P and `layer3.1.conv3` for the MIMIC. LABELCODE and LABELCODE-N are built in RN50's best-layer basis (the constructions of Section 4.2 applied to its 3,500 features), and the null shuffles it. A separate VAE is trained on each representation with the settings of Appendix B.

Results. Read at its best layer, the trained network outscores every adversary on RSA and dCKA (0.382/0.295; Table 7, Figure 8), but the foil models still reach 65–88% of it on RSA and 55–89% on dCKA, CARRIER-B (88%, 89%) as much as $RN50_{\text{sc}}$ (87%, 85%), a network actually trained on the images. Still, the corrupted representations CAT_{own} and $\text{S-CAT}_{\text{own}}$ are scored by RSA and dCKA higher than the reference ImageNet pretrained RN-50.

Ridge again holds every foil model between 0.285 and 0.306, against 0.407 for the network and 0.196 untrained. The latent read-outs flag them as dissimilar with the brain representations and place all five below the untrained network on occupancy (0.065–0.256, against 0.279) and top-1 (0.016–0.025, against 0.040), where the trained network reaches 0.709 and 0.312; on median rank they sit at or just ahead of the untrained network (47–61, against 63.5), an order of magnitude behind the trained network (3.8) and $RN50_{\text{sc}}$ (7.1).

I RSA ESTIMATORS AND COMPARATORS

An RSA score depends on how the dissimilarity matrices are estimated and how they are compared (Nili et al., 2014; Diedrichsen & Kriegeskorte, 2017). Table 8 reports three combinations. *Correlation distance with Pearson correlation* is the RSA of the main text. *Correlation distance with Spearman's ρ_a*, the rank RDM comparator of Schütt et al. (2023). *Crossnobis with whitened cosine* estimates brain RDMs as squared Euclidean distances between voxel patterns cross-validated over the NSD repeats, with each voxel divided by its residual standard deviation (Walther et al., 2016); the estimates are neither square-rooted nor clipped at zero. Model RDMs are squared Euclidean distances between z-scored features, which need no cross-validation because the features carry no measurement noise, and the two are compared by the whitened cosine (Diedrichsen et al., 2020). The different versions of RSA order the representations alike and differ only on near-ties.

Table 7: **Best layer** (3,500-d, each network's highest-RSA block). One training seed, each run read at the epoch selected by the stopping protocol (Appendix B). The ladders are built from RN50's best-layer features, at $k = 3$. Columns as in Table 3; identification over the 872 shared test images (chance 0.001, 0.006 and 436.5). Bold: at or above the trained network.

Condition	ridge r	ridge R^2	RSA	dCKA	iLISI (↑)	energy (↓)	top-1 (↑)	top-5 (↑)	med. rank (↓)
RN50	**0.407**	**0.188**	0.382	0.295	**0.709**	**0.25**	**0.312**	**0.598**	**3.8**
$RN50_{\mathrm{sc}}$	0.382	0.166	0.331	0.251	0.592	0.42	0.219	0.470	7.1
$\mathrm{RN50}_{\alpha=1}$	0.196	0.040	0.073	0.056	0.279	0.69	0.040	0.124	63.5
Category Corruption ladder ($k = 3$)									
$\mathrm{CAT}_{\mathrm{own}}$	0.367	0.157	**0.427**	**0.334**	0.502	0.68	0.090	0.302	13.8
$\mathrm{CAT}_{\mathrm{in}}$	0.296	0.100	0.150	0.110	0.203	1.16	0.023	0.089	62.1
$\mathrm{CAT}_{\mathrm{out}}$	0.294	0.100	0.019	0.015	0.182	1.28	0.017	0.069	92.7
Supercategory Corruption ladder ($k = 3$)									
$\text{S-CAT}_{\mathrm{own}}$	0.354	0.147	**0.397**	**0.321**	0.403	0.70	0.059	0.215	21.4
$\text{S-CAT}_{\mathrm{donor}}$	0.304	0.111	0.228	0.175	0.185	1.15	0.017	0.069	66.6
$\text{S-CAT}_{\mathrm{swap}}$	0.286	0.100	0.142	0.087	0.132	1.22	0.012	0.057	79.4
Foil Models									
LABELCODE	0.285	0.097	0.257	0.200	0.078	1.62	0.019	0.077	61.4
LABELCODE-N	0.303	0.109	0.280	0.221	0.065	1.64	0.025	0.105	48.9
MIMIC	0.286	0.098	0.247	0.162	0.097	1.42	0.016	0.076	61.3
CARRIER-B	0.306	0.112	0.336	0.264	0.182	1.10	0.023	0.094	52.3
CARRIER-P	0.302	0.108	0.279	0.215	0.256	0.83	0.025	0.103	46.9
NULL	−0.001	−0.006	0.000	0.000	0.004	2.02	0.001	0.005	428.8

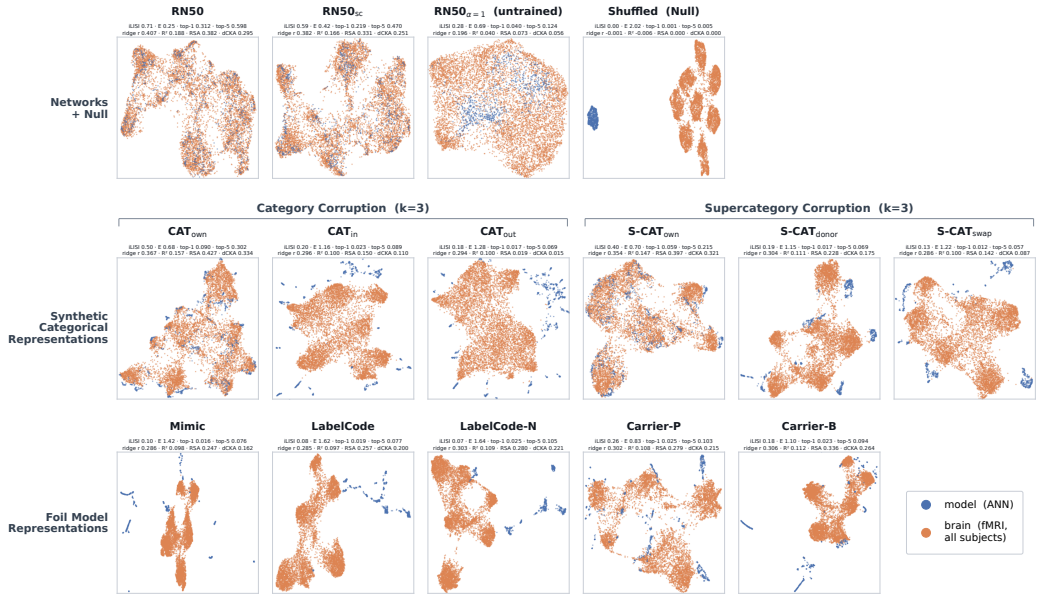


Figure 8: VAE latent visualisation (UMAP-2d) at each network's best layer (3,500-d), one seed, each run at its selected epoch; panel subtitles give the layer. Colours and panel titles as in Figure 4.

The three versions order the representations alike (Spearman correlation 0.96–1.00 between their scores), and all three barely separate own from donor variance (0.005–0.015).

J LATENT PANELS COLOURED BY STIMULUS CONTENT

Figure 9 recolours the latent panels of Figure 4 by supercategory (the rarest the image contains).

Table 8: RSA under three estimator/comparator combinations, averaged over the eight subjects and floor-corrected against a permutation null. Correlation distance on the 872 test images; crossnobis on the 727 of them with at least two repeats in every subject. The Pearson column is the RSA of Table 3. Bold: at or above the reference RN-50 trained network.

Condition	crossnobis whitened cosine	correlation distance Spearman ρ_a	correlation distance Pearson r
Training Degradation			
$\mathrm{RN50}_{\alpha=0}$	0.129	0.236	0.249
$\mathrm{RN50}_{\alpha=0.25}$	0.055	0.111	0.118
$\mathrm{RN50}_{\alpha=0.50}$	0.032	0.070	0.077
$\mathrm{RN50}_{\alpha=0.75}$	0.029	0.060	0.065
$\mathrm{RN50}_{\alpha=1}$	0.026	0.052	0.056
$RN50_{\mathrm{sc}}$	**0.135**	0.224	0.241
Foil Models			
LABELCODE	**0.191**	0.230	**0.261**
LABELCODE-N	**0.212**	**0.254**	**0.288**
MIMIC	**0.148**	0.193	0.223
CARRIER-B	**0.213**	**0.296**	**0.301**
CARRIER-P	**0.184**	**0.239**	**0.265**
Controls			
NULL	0.000	−0.001	−0.001
Category Corruption			
$\mathrm{CAT}_{\mathrm{own}}$	0.128	0.186	0.200
$\mathrm{CAT}_{\mathrm{in}}$	0.099	0.150	0.161
$\mathrm{CAT}_{\mathrm{out}}$	0.019	0.011	0.029
Supercategory Corruption			
$\mathrm{S\text{-}CAT}_{\mathrm{own}}$	**0.153**	0.226	0.242
$\mathrm{S\text{-}CAT}_{\mathrm{donor}}$	**0.148**	0.213	0.227
$\mathrm{S\text{-}CAT}_{\mathrm{swap}}$	0.070	0.113	0.137

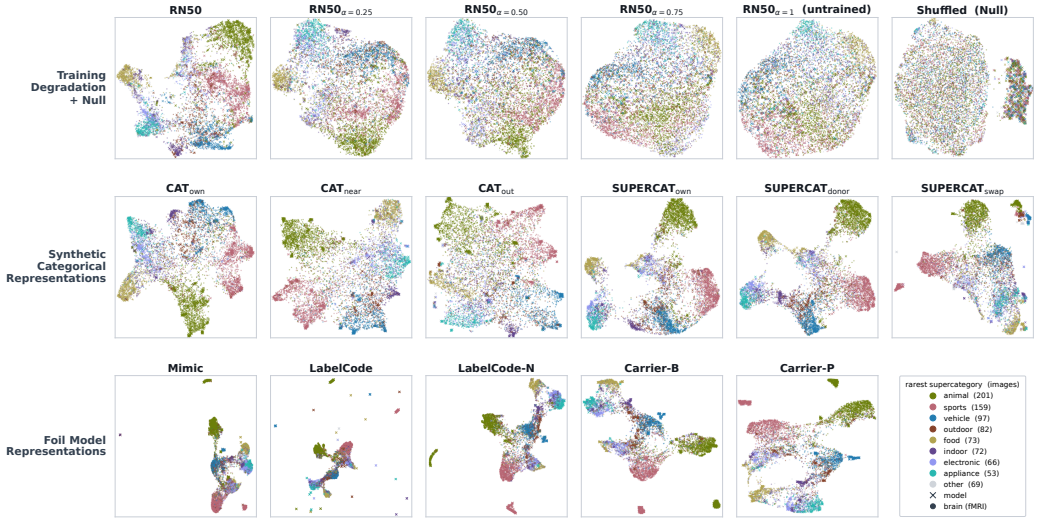


Figure 9: The panels of Figure 4, coloured by one supercategory per image: the rarest supercategory it contains (by frequency over all images, the rule used throughout for one label per image), so that, e.g., an image of a person playing tennis takes the colour of *sports*. The eight most frequent cover 92% of the 872 test images; the remaining points (accessory, furniture, kitchen, person only) are grey. Model points are drawn as × and brain points as ○.