# ROSETTA NEURONS ACROSS VISION AND LANGUAGE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Can independently trained vision and language models develop individual neurons that respond to the same concepts? We identify these *cross-modal rosetta neurons*: pairs of neurons in separate vision and language models whose activations correspond across images and text. We extend neuron matching to paired images and captions by correlating each neuron's mean activation over the corresponding inputs. We find such matches across CLIP, DINOv2, MAE, and a Stable-Diffusion UNet paired with language models, including vision encoders trained without text. We recover the same neuron pairs in two separately sampled sets of image–caption pairs. Blinded model judges and human raters support concept agreement among selected high-correlation pairs. The matched population grows with model size, following sublinear power-law scaling across language models and text-free vision encoders. These findings suggest that independently trained models can represent corresponding visual and linguistic concepts at the level of individual neurons. We use these correspondences for text-indexed image retrieval: a concept named in text selects a language neuron, whose matched vision neuron retrieves images of that concept without additional training.

## 1 INTRODUCTION

An image and its caption can describe the same objects, actions, and scenes through different forms of information. A baseball game appears in one as players, bats, and a field, and in the other as words describing the play. This shared content motivates the idea that vision and language models may develop corresponding representations, even when trained independently Huh et al. (2024). Studies of representational similarity and model stitching examine such correspondence through the geometry of activation spaces or mappings between them Kornblith et al. (2019); Raghu et al. (2017); Lenc & Vedaldi (2015). We ask whether it also appears at the level of individual neurons: can a neuron in a vision model and a neuron in a language model respond to the same concept?

Within vision models, Dravid et al. (2023) identified neurons with corresponding responses across models trained with different architectures, objectives, and supervision. They called these shared units *rosetta neurons* and used their correspondences to connect the representations of different models. Dravid et al. (2026) subsequently studied neuron correspondence within vision and within language, finding that the matched population grows with model size. Here we extend this investigation across the modality boundary. Such correspondences would let us use a language neuron to select a vision neuron and test whether shared concepts emerge when each modality is learned independently. We call the resulting pairs *cross-modal rosetta neurons*: individual units in independently trained vision and language models whose activations correspond across images and their descriptions. Figure 1 illustrates such pairs through their strongest activating images and captions, which depict or describe concepts such as baseball, food, and airplanes.

How can we compare neurons when one model receives an image and the other receives text? Within a modality, both models can process the same inputs, allowing their neurons' responses to be compared directly. Across modalities, we use the *(image, caption) pair* as the shared observation. For each pair, we average a vision neuron's activation over spatial locations and a language neuron's activation over caption tokens. Each neuron then has one response per example, and we correlate these responses across the paired dataset. We use the mutual top-$K$ matching rule of Dravid et al. (2023) to identify corresponding neurons, retain pairs above a correlation threshold, and require the same pair to be recovered in 2 separately sampled datasets. This procedure connects the models through



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.

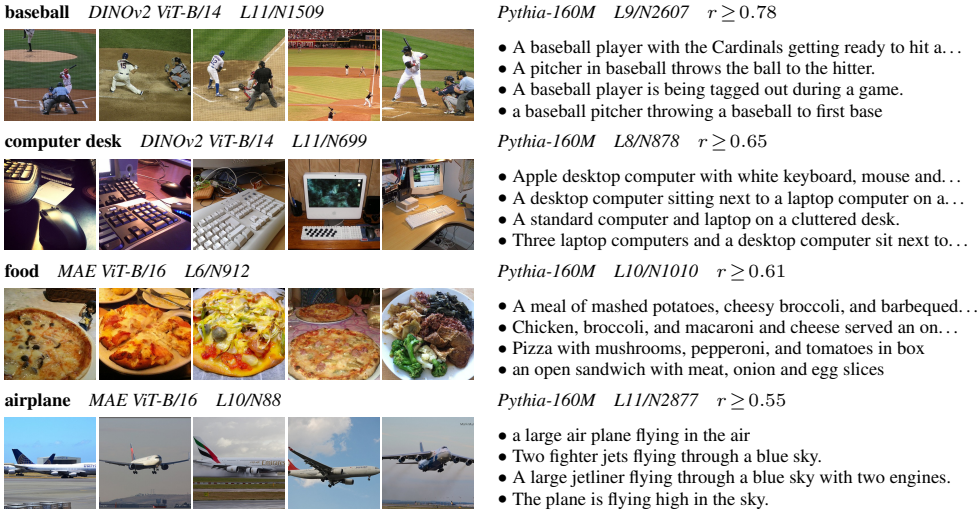


Figure 1: **Cross-modal rosetta neurons.** Each row is one matched pair: a single neuron in a vision encoder trained without any text (DINOv2 ViT-B/14 or MAE ViT-B/16 with its top-activating images on the left) and a single neuron in the Pythia-160M language model (with its top-activating captions), over the same 5,000 MS-COCO val2014 (image, caption) pairs. The two models were trained separately and share no weights and no input space. The pairs were found from activation statistics alone and with no concept labels. The bold word is our description and plays no part in the matching. $r$ is the Pearson correlation between the two neurons' mean-pooled activations, taking the smaller value across two independent draws of the data. Further examples can be seen in App. A.

their responses to corresponding content, using paired data without fitting a mapping between their representations.

We start by investigating whether this procedure finds a reproducible population of cross-modal matches. We evaluate 7 vision–language settings spanning different model pairs, training objectives, and paired datasets (Tab. 1). For our reference pair, CLIP ViT-B/16 and Pythia-160M, we find approximately 2,000 neuron pairs in each sampled set of 5,000 image–caption pairs, compared with approximately 20 under caption permutation. Of these, 1,103 neuron pairs are recovered in both sets. We further examine the population using caption-length controls, removal of dominant activation directions, and larger paired datasets. Together, these comparisons establish a population of matches that persists across data samples and several controls for shared input properties.

To examine what the matched neurons represent, we ask blinded model judges to label their strongest activating images and captions independently. Among the 100 pairs with the highest correlations, labels agree more closely for matched partners than for random or partner-permuted controls, and human ratings support agreement in the evaluated subset. Rescoring matched neurons with other captions of the same images also preserves their correlations. We then examine whether such correspondence occurs when the vision model was trained entirely without text, using DINOv2 and MAE. Both yield matched populations with concept agreement, showing that shared concepts can be identified across modalities even with text-free vision training.

Having identified these populations, we investigate how their size changes as the models grow. We vary the language model, the vision model, and both models together. Matched-pair counts increase across the tested model sizes. Scaling Pythia, DINOv2, MAE, or both models together meets our criterion for sublinear power-law scaling: the number of matched pairs grows more slowly than the number of analyzed neurons. The OpenCLIP vision series shows a flatter trend across model sizes. These results show that growing matched populations occur across both caption-supervised and text-free vision training.

Finally, we use the matches to retrieve images from a concept named in text. The text activates a language neuron, its matched vision neuron scores the candidate images, and those scores determine the ranking. For categories where retrieval works best, such as cat, giraffe, and cake, all 6 top-ranked

images contain the named concept. The application uses the discovered neuron pairs with frozen models and requires no additional training. It demonstrates how correspondence between individual neurons can connect a linguistic description to a visual response in a separately trained model.

## 2 Related Work

**Rosetta neurons and neuron universality.** Dravid et al. (2023) mine units whose activation maps correlate across a zoo of independently trained *vision* models, and argue that some visual concepts are intrinsic enough that any competent model finds them. Dravid et al. (2026) scales this approach and reports a scaling trend of $|\mathcal{R}| \propto N^{\alpha}$ with $\alpha \approx 0.5$–$0.7$ for language models to 30B and vision models to 5B (the latter on images generated under ImageNet-1k conditioning Deng et al. (2009)), plus a Neuron Polarization Effect in which matched units grow more selective while unmatched ones do not. Both works match LM to LM or vision to vision and neither crosses modalities, which is the gap we address. Related unit-level work includes universal neurons across GPT-2 replicas Gurnee et al. (2024), convergent learning Li et al. (2016) and a longer interpretability line Zhou et al. (2015); Bau et al. (2017; 2020); Mu & Andreas (2020); Hernandez et al. (2022). Single units are a contested object of study, since superposition predicts features need not align with neurons at all Elhage et al. (2022); Olah et al. (2020) and sparse autoencoders recover concepts from directions rather than channels Bricken et al. (2023); Cunningham et al. (2024).

**Convergence measured on representations.** A larger literature argues for convergence at the level of *representations*. The platonic representation hypothesis Huh et al. (2024) holds that as models grow they converge toward one shared representation of the world, whatever their modality, and supports this with rising alignment between vision and language embeddings. Model stitching Lenc & Vedaldi (2015) tests whether the layers of two networks can be joined through a small learned map, CKA and CCA Kornblith et al. (2019); Raghu et al. (2017); Morcos et al. (2018) measure how similar two representation spaces are up to a linear transformation, and Dhimoïla et al. (2026) recover sparse atoms shared by both sides of a vision–language embedding. Those measures are confounded by scale, and Gröger et al. (2026) show that calibrating against a permutation null removes most of the global convergence signal, which is why we score every population against a null recomputed under the same conditions.

**Cross-modal alignment.** Huh et al. (2024) run a cross-modal measurement using mutual $k$NN between whole representations on Wikipedia image–text pairs, and find only weak alignment. Koepke et al. (2026) then show that this alignment degrades sharply as the gallery (the pool of image–text pairs within which matches are searched) grows from 1K to 15M pairs, and again when the one-to-one image–caption assumption is relaxed. As the gallery grows, each input has more candidates competing to be its nearest neighbour. In contrast, we match neurons, and the candidate pool is the fixed set of units in the two models. Adding image–caption pairs supplies more observations for estimating each correlation.

**Bridging two unimodal models.** Paired data can couple separately trained encoders without joint training at the level of representations. For example, through shared anchors Moschella et al. (2023), two frozen encoders and a set of image–text pairs Norelli et al. (2023), or a single linear map Merullo et al. (2023). Closer to the unit level, Goh et al. (2021) describe multimodal neurons *inside a single* CLIP model, Schwettmann et al. (2023) find them in a text-only transformer fed *projected* visual input, and Cui et al. (2026) identify neurons shared between an LLM and a VLM *built on the same backbone*. In every case a single model, a learned projection or a shared ancestry supplies the bridge. To examine correspondence beyond caption supervision, we match language neurons with neurons in two vision encoders trained without text and compare independently trained models with a jointly trained reference (Sec. 5.2).

## 3 Method

**Setup and notation.** Following Dravid et al. (2026); Gurnee et al. (2024), we analyze neurons in the MLP blocks of each model. We define a neuron as one intermediate activation channel and include neurons from every transformer block, excluding attention and embedding components from matching. For GELU-based MLPs, we record post-GELU activations. For DINOv2 ViT-g/14's gated MLPs, we record the input to the output projection. We use each image–caption pair as a shared ob-

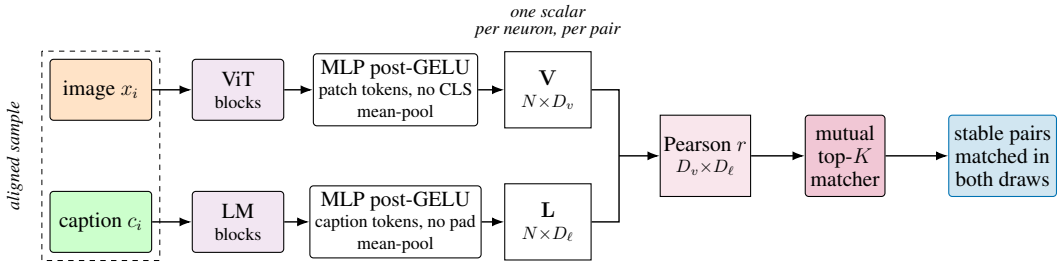


Figure 2: **Cross-modal matching.** Vision and language models share no input space, so the *(image, caption) pair* is the aligned sample. Each neuron is mean-pooled to one scalar per pair, giving two activation matrices over the same $N$ samples, and a mutual top-$K$ matcher is applied over a single correlation table spanning all layers of both models. A pair is kept only if it is matched in two independent draws of the data.

servation. For the image, we average each vision neuron's activation over the patch tokens of the center-cropped $224 \times 224$ image, excluding the CLS token. For the caption, we average each language neuron's activation over the caption's tokens, with no start token added and padding excluded. Each neuron therefore contributes one scalar response per image–caption pair. Given $N$ pairs, we collect these responses into activation matrices $\mathbf{V} \in \mathbb{R}^{N \times D_v}$ and $\mathbf{L} \in \mathbb{R}^{N \times D_\ell}$, where $D_v$ and $D_\ell$ are the total numbers of analyzed vision and language neurons. Corresponding rows describe the same image–caption pair, providing the shared sample axis used for matching (Fig. 2).

**Matching.** We standardize each column of the activation matrices by subtracting its mean and dividing by its standard deviation across the $N$ examples. We then compute the Pearson correlation between every vision neuron and every language neuron, giving a $D_v \times D_\ell$ correlation matrix spanning all layers of both models. We denote each pairwise correlation by $r$. Positive values mean that the neurons tend to activate together on corresponding images and captions. We identify matches using the mutual top-$K$ rule of Dravid et al. (2023): each neuron must be among its partner's $K$ most strongly correlated neurons. We additionally require $r \geq \tau$, so matches satisfy both a relative ranking criterion and a minimum correlation strength. We use $K = 5$ and $\tau = 0.2$ throughout unless stated otherwise. We also evaluate $K \in \{1, 3, 5, 10\}$, including the $K = 1$ setting used by Dravid et al. (2026), and report the correlation-threshold sweep in App. G. We rank neurons by signed correlation, so the positive threshold retains pairs that activate together. For the largest model pair, the correlation matrix has 737k $\times$ 393k entries. We compute it in blocks to limit memory use. On the reference model pair, this produces the same matches as computing the full matrix at once.

**Permutation baselines.** We assess the match count against a caption-permutation baseline. We shuffle the correspondence between images and captions and repeat the matching procedure, preserving each neuron's activation distribution. Repeating this shuffle gives a null distribution of match counts under random image–caption pairing. We use 30 permutations for the reference model pair, 100 for the $K$ and $\tau$ sweeps, and 20 for the model-size comparisons (Sec. 4), where larger correlation matrices make each permutation more expensive. We express the observed count relative to this baseline as $z = (m_{\text{obs}} - \mu_{\text{perm}})/\sigma_{\text{perm}}$, where $m_{\text{obs}}$ is the observed count and $\mu_{\text{perm}}$ and $\sigma_{\text{perm}}$ are the mean and standard deviation of the permuted counts.

**Controls for shared input properties.** We further examine how matching depends on dominant activation patterns and caption length. To test the first, we remove the leading $q$ principal components from each activation matrix, with $q = 1$ or 5, and repeat the matching procedure with residuals restandardized. This measures how much the matched population depends on these dominant directions. To control for caption length, we sort captions by word count into 10 equally sized bins and shuffle captions within each bin. Each image is therefore paired with a caption of approximately the same length, so this baseline preserves caption length while breaking the specific image–caption correspondence.

**Stability across samples.** For the main results, we retain neuron pairs found in both of two separately sampled sets of image–caption pairs. We refer to these pairs as *seed-stable*. In the gallery-size

| setting | vision model | language model | data | seed-stable pairs |
|---|---|---|---|---|
| reference | CLIP ViT-B/16 | Pythia-160M | COCO | 1,103 |
| text-free | DINOv2 ViT-B/14 | Pythia-160M | COCO | 1,055 |
| text-free | MAE ViT-B/16 | Pythia-160M | COCO | 781 |
| generative | SD-1.4 UNet | Pythia-160M | COCO | 775 |
| language | CLIP ViT-B/16 | GPT-2 | COCO | 1,402 |
| corpus | CLIP ViT-B/16 | Pythia-160M | Flickr30k | 1,086 |
| joint | CLIP ViT-B/16 | CLIP text encoder | COCO | 2,335 |

Table 1: **One component changed at a time.** Seed-stable neuron pairs at $N{=}5000$ samples. Grey entries are unchanged from the reference setting, and bars are scaled to the largest count.

experiment (Sec. 5.1), we keep the two sets disjoint so that the comparison across gallery sizes uses separate observations for the stability check.

## 4 Experimental setup

**Models and data.** As a jointly trained reference, we first run our pipeline on the vision and text encoders of OpenAI CLIP ViT-B/16, trained on 400M web image–text pairs Radford et al. (2021): a 12-layer vision transformer with 16-by-16 patches (about 86M parameters) and a 12-layer text transformer (about 63M parameters). These encoders were trained together with an objective that explicitly aligns them, providing a reference for the correspondence we find between independently trained models. For our reference pair we match the CLIP vision encoder against Pythia-160M, a 160M-parameter language model Biderman et al. (2023), trained on the Pile, roughly 300B tokens of English text Gao et al. (2021). The two models come from different groups, were trained on different corpora, and share no weights and no tokenizer. For matching we draw $N{=}5000$ images from MS-COCO val2014 Lin et al. (2014), each with the first of its five captions. A second draw with a different seed selects a different set of images (the two overlap by about 12%) and serves the seed-stability check. We test generality by changing one component at a time (Tab. 1). We replace the vision encoder with DINOv2-B, trained on images alone from LVD-142M Oquab et al. (2024), and the language model with GPT-2, trained on WebText Radford et al. (2019). We also repeat the analysis on Flickr30k Young et al. (2014). Finally, we test a generative model by matching FFN neurons in a Stable-Diffusion UNet Rombach et al. (2022) to language-model neurons.

**Scaling with model size.** We vary model size in five settings, and call each such series of models a *ladder*. First, we increase the Pythia language model from 70M to 12B parameters while keeping CLIP ViT-B/16 fixed. Second, we increase the OpenCLIP vision model from ViT-B/16 to ViT-bigG/14 while keeping Pythia-160M fixed. All OpenCLIP models in this comparison use LAION-2B weights, keeping the training corpus constant Cherti et al. (2023); Schuhmann et al. (2022). Third, we increase both models together. Finally, we repeat vision scaling with DINOv2 S/B/L/g Oquab et al. (2024) and MAE ViT-B/L/H He et al. (2022), each paired with Pythia-160M to test whether scaling also occurs when the vision model was trained without captions. On every ladder we count seed-stable neuron pairs $|\mathcal{R}|$, and each count must exceed its seed-stable caption-permutation baseline (App. H). We measure model size $x_D$ as the number of MLP neurons of the growing model, or the arithmetic mean of both models' counts when both grow, and fit $\log|\mathcal{R}| = \alpha \log x_D + c$ by least squares. We consider the results consistent with sublinear scaling if $R^2 \geq 0.9$ and the 95% confidence interval for the exponent $\alpha$ lies in (0, 1).

**Concept agreement.** To examine which concepts the matched neurons represent, we use a blinded judge, Qwen2.5-VL in its 7B and 32B versions Bai et al. (2025). For each model pair, we evaluate the 100 seed-stable neuron pairs with the highest correlations. For each unit, we select its 9 top-activating images or captions from the same MS-COCO pairs and ask the judge to name the concept. We label each unit in a separate call, present items in shuffled order, and omit partner identities and condition labels. Agreement is the cosine similarity between sentence embeddings of the two labels Reimers & Gurevych (2019). We test matched pairs against each control with a one-sided Mann–Whitney $U$ and report the AUC $P(\text{matched} > \text{control})$. We score labels rather than yes-or-no verdicts because the verdicts collapse to nearly constant answers.

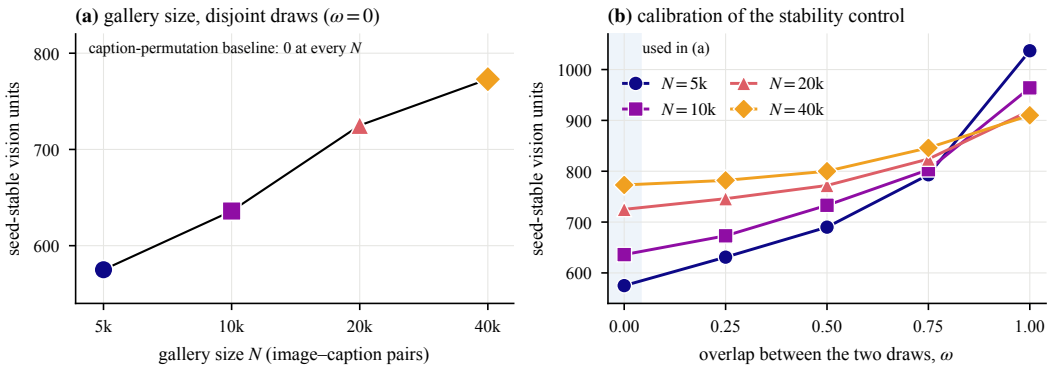


Figure 3: **(a)** Seed-stable matched vision units against gallery size, with disjoint draws. The population grows over an $8\times$ range while the caption-permuted null stays at zero. **(b)** Why the draws have to be disjoint. Two size-$N$ draws from a pool of size $M$ share $\omega = N/M$ of their data, and at $\omega{=}1$ the population seems to *fall* with gallery size. Colors and markers mark gallery size in both panels.

**Human ratings.** Twelve raters, not told about the different conditions, rated the same stimuli on a 1–5 scale, giving 204 study ratings. Exclusion rules and quality control are described in App. B.

## 5 Results

### 5.1 A stable population of cross-modal matches

We first test for cross-modal rosetta neurons in the reference pair, CLIP ViT-B/16 and Pythia-160M. Matching the two models yields about 2,000 matched pairs per draw, compared with about 20 under caption permutation, giving a $z$-score near 270 (App. F). Of these, 1,103 pairs are matched again in the second draw (Tab. 1, first row), whereas no pair survives both draws under caption permutation (App. H). In these seed-stable pairs, 61% of the vision units have exactly one language partner.

**Controls for shared input properties.** We examine how the matched population changes under controls for caption length and dominant activation directions. Permuting captions within length bins gives a similar baseline to unrestricted permutation, so caption length does not explain the matches. Removing the leading principal component from each side preserves a similar count, and removing five leaves more than 1,900 pairs per draw. Removing components does change which pairs are found: 78% of the seed-stable pairs keep their identity when one component is removed, but only 39% when five are, although the population stays large (App. J). The count also exceeds its baseline at every $K$, from 145 pairs at $K{=}1$ to 6,287 at $K{=}10$. Without threshold $\tau$, the seed-stable population is essentially unchanged and still far above its caption-permutation baseline (App. F). The threshold only removes the many weak mutual pairs from the single-draw counts, while seed stability, not $\tau$, separates matches from chance (App. G).

**Gallery size.** Is a gallery of 5,000 pairs large enough? Koepke et al. (2026) report that cross-modal alignment measured by mutual nearest neighbours *collapses* as the gallery grows from 1K to 15M pairs, so we swept the gallery from 5,000 to 40,000 pairs, drawing the two seeds from the pooled train2014 and val2014 sets so that they never overlap. With disjoint draws, the matched population increases from 575 to 773 vision units as the gallery grows from 5,000 to 40,000 pairs, while the caption-permutation baseline stays at zero (Fig. 3a). Both models are frozen, so the additional matches are recovered from an existing population as more observations become available. The candidate pool in neuron matching stays fixed as the gallery grows; in nearest-neighbour retrieval, it grows with the gallery. The overlap control shows why this design matters: reusing examples across draws changes the measured counts and their trend with gallery size (Fig. 3b). Two independently sampled draws of $N$ pairs from a pool of $M$ have an expected overlap fraction of $\omega = N/M$. Keeping them disjoint separates the effect of gallery size from that of shared observations. Pooling the two COCO splits supports disjoint draws of up to about 62,000 pairs each. Extending this comparison

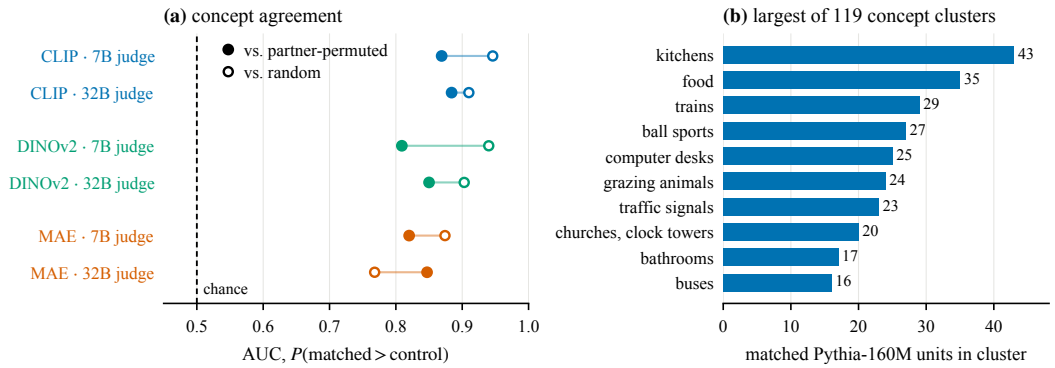


Figure 4: **What matched pairs encode. (a)** Concept agreement: the probability (AUC) that the labels of a matched pair agree more than those of a partner-permuted pair (filled) or a random pair (open), per vision encoder, each paired with Pythia-160M, for both judges. The judges evaluate the 100 seed-stable pairs with the highest correlations and 100 pairs per control condition. Human ratings are in Fig. 5. **(b)** The ten largest concept clusters among the 612 matched Pythia-160M units of the reference pair, named by reading their captions (App. E).

to the millions of pairs studied by Koepke et al. (2026) requires a larger paired corpus. We recover a large population of cross-modal matches across separately sampled datasets, larger galleries, and controls for caption length and dominant activation directions.

### 5.2 Generality across encoders, corpora, and training objectives

Does the result depend on CLIP, on COCO, or on discriminative training? We change one component at a time and rerun the pipeline (Tab. 1). We replace the vision encoder with DINOv2 or MAE to test text-free training, and with the FFN layers of a Stable-Diffusion UNet to test a generative model. We also replace the language model with GPT-2 or the paired corpus with Flickr30k. Across these settings, we find between 775 and 1,402 seed-stable pairs, with every count exceeding its caption-permutation baseline. DINOv2 and MAE let us examine correspondence learned without caption supervision. We therefore also test their concept agreement (Sec. 5.3) and their scaling (Sec. 5.4).

**Jointly trained reference.** Matching CLIP's vision encoder to its jointly trained text encoder yields 2,335 seed-stable pairs, $2.1\times$ as many as the 1,103 found with the independently trained Pythia-160M (last row of Tab. 1). This comparison places the independently trained pair alongside a jointly trained reference; the text encoders differ in architecture, size, training data, and objective. Both settings yield substantial populations of cross-modal matches.

### 5.3 Matched pairs encode shared concepts

We next examine whether matched neurons respond to the same concepts. We label each unit of the 100 highest-correlation pairs with a blinded judge (Sec. 4) and compare matched pairs with random pairs and with *partner-permuted* pairs, which keep exactly the same units and only shuffle the pairing. The partner-permuted control tests whether concept agreement depends on which neurons are paired. Among these high-correlation matches, both judges find greater concept agreement for matched pairs than for either control (Fig. 4a). Against partner-permuted pairs, the AUC is 0.87 for the 7B judge and 0.88 for the 32B judge. Appendix I reports the results for each model and judge.

**Human raters.** Human raters agree with the judges on the same stimuli (Fig. 5). Twelve raters, unaware that conditions existed, gave matched pairs a mean rating of 3.9 on a five-point scale, against 1.7 for partner-permuted and 1.3 for random pairs. Against partner-permuted pairs, the AUC is 0.87 at $p = 2\times10^{-7}$. Agreement between raters is high, with $\mathrm{ICC}(1,1) = 0.80$ (App. B).

**Vision encoders trained without text.** Concept agreement also holds when the vision encoder never saw text: against partner-permuted pairs, the AUC is 0.81 and 0.85 for DINOv2 and 0.82 and 0.85

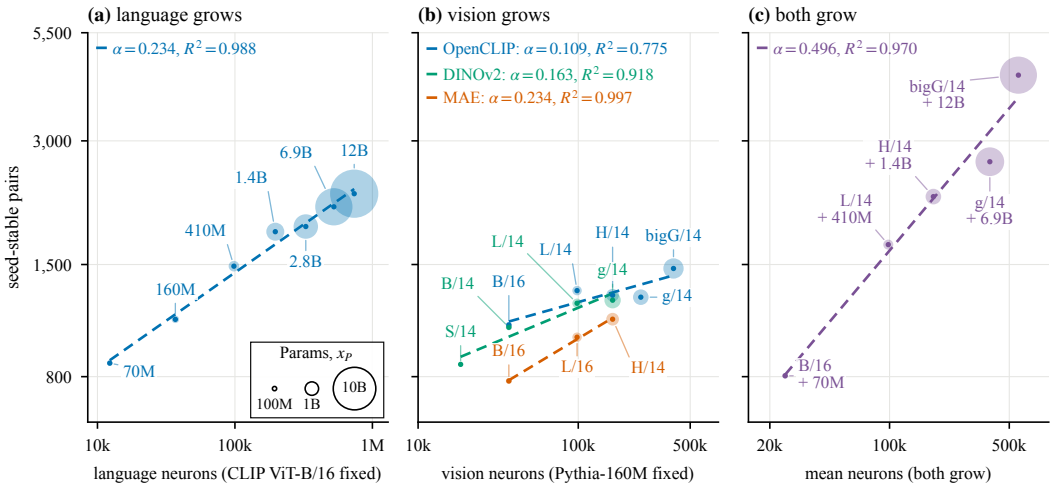


Figure 6: Seed-stable pairs against model size, measured in MLP neurons of the growing model (the mean over both models in (c)). **(a)** Pythia grows with CLIP ViT-B/16 fixed. **(b)** The vision encoder grows with Pythia-160M fixed. **(c)** Both models grow. Bubble area is proportional to parameter count, using one common area scale across panels. Dashed lines are power-law fits in log–log space. Legends give the exponent $\alpha$ and $R^2$ in log space.

for MAE, with the 7B and 32B judges respectively (Fig. 4a, App. I). Together, the judge and human evaluations support shared concepts among the selected high-correlation pairs, including pairs with vision encoders trained without text.

**Caption wording.** We next test whether matched neurons retain their correspondence across different descriptions of the same image. The COCO dataset lets us test this because each image has five captions written independently, while our pipeline used only the first. We keep the matched pairs fixed and score them again using captions two and three, which were not used for matching. The correlation barely changes (median $r'$ of 0.40 against 0.40 originally), 98% of pairs still pass the threshold, and partner-permuted correlations stay near zero. The captions can emphasize different details: for example, "a woman in the kitchen cooking some food" can become a lengthy description of the counters and pots. This result suggests that the models share the concept rather than the wording of a particular caption.

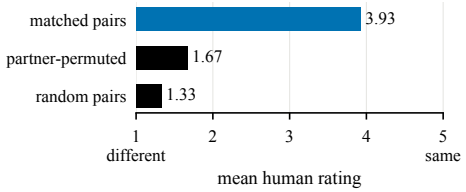


Figure 5: Human ratings of concept agreement, from 1 (different concepts) to 5 (same concept), averaged over the 204 study ratings of twelve raters, 30 items per condition.

**Concept coverage.** To measure concept coverage across the matched population, we group neurons by their top-activating captions. We cluster the 612 Pythia-160M units that appear in the 1,103 seed-stable pairs, representing each unit by the average sentence embedding of its nine top-activating captions. The units form 119 clusters, with 7% in the largest cluster. The 10 largest clusters correspond to scene and object categories: kitchens, food, trains, ball sports, computer desks, grazing animals, traffic signals, churches and clock towers, bathrooms, and buses (Fig. 4b, App. E). Unmatched units form about twice as many clusters, while their top captions have lower mean pairwise similarity (0.20, compared with 0.33 for matched units). The population therefore covers many concepts, and the matched units are the more coherent ones.

### 5.4 SUBLINEAR SCALING WITH MODEL SIZE

We now ask how the population grows with model size, on the five ladders of Sec. 4. The number of seed-stable matched pairs increases overall with model size, with sublinear fitted trends across all five series (Fig. 6). Scaling Pythia from 70M to 12B parameters with CLIP ViT-B/16 fixed gives $\alpha = 0.234$ (95% CI $[0.204, 0.265]$). With Pythia-160M fixed, the corresponding exponents are $0.109$ for OpenCLIP ViT, $0.163$ for DINOv2 ViT, and $0.234$ for MAE ViT. Growing both models together gives $\alpha = 0.496$ (95% CI $[0.336, 0.656]$), consistent with approximately square-root growth in the average neuron count. The strength of the power-law description varies across families. Four of the five ladders meet the criterion of Sec. 4, with $R^2 = 0.988$ for Pythia, $0.918$ for DINOv2, $0.997$ for MAE, and $0.970$ when both models grow. The OpenCLIP vision ladder does not ($R^2 = 0.775$), because its count stays nearly flat from ViT-L/14 to ViT-g/14. The fit quality for MAE should be interpreted cautiously because only three model sizes are available.

**Selectivity.** Matched units are also more selective than unmatched units from the same layers. From Pythia-410M upward, the mean vocabulary selectivity of matched language units exceeds that of layer-matched controls, and about 95% of matched vision units are judged monosemantic at every OpenCLIP size, against 71–84% of controls (App. C).

## 6 APPLICATION: TEXT-INDEXED SINGLE-NEURON RETRIEVAL

To retrieve images from a concept named in text, we select the language neuron it activates most among the stable pairs. We follow its match to a vision neuron and rank held-out images by that neuron's activation, without additional training, concept labels for the neurons, or a joint embedding space. For the six categories with the highest average precision (*cat*, *giraffe*, *teddy bear*, *cake*, *horse* and *clock*), all six top-ranked images are correct, with average precision between 0.73 and 0.92 (App. D, which shows the retrieved images). Across all 80 COCO categories, median average precision is 0.045, compared with 0.024 for partner-permuted units and 0.030 for random units. Performance varies across categories, with the strongest retrieval concentrated in the concepts above. In contrast to CLIP-Dissect Oikarinen & Weng (2023), our method uses text to reach a specific channel of a vision encoder without any joint embedding space to perform ranking.

## 7 CONCLUSION

We show that representational universality between independently trained networks extends beyond subspaces to individual units. Stable neuron pairs pass several robustness checks, and concept-agreement evaluations support shared concepts among selected high-correlation matches. The matched population grows with model size, following sublinear power-law scaling across language models and text-free vision encoders. When both models grow, the matched-pair count follows approximately square-root scaling in their average neuron count. A text-selected language unit can identify a useful vision channel without training, labels, or a joint embedding space. Our analysis characterizes correspondence through activations pooled over tokens and patches, with concept evaluation and scaling experiments focused on COCO. The human study evaluates a subset of the highest-correlation pairs. Future work can extend the scaling experiments to larger models and broader paired corpora and evaluate concept agreement across the full matched population. Examining correspondence at individual tokens and patches could connect these neuron-level findings to feature-based theories of superposition.

## REFERENCES

Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Int. Conf. Mach. Learn.*, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, et al. Reproducible scaling laws for contrastive language-image learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

Chenhang Cui, An Zhang, Yuxin Chen, Gelei Deng, Jingnan Zheng, Zhenkai Liang, Xiang Wang, and Tat-Seng Chua. Do llms and vlms share neurons for inference? evidence and mechanisms of cross-modal transfer. *arXiv preprint arXiv:2602.19058*, 2026.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *Int. Conf. Learn. Represent.*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009.

Grégoire Dhimoïla, Thomas Fel, Victor Boutin, and Agustin Picard. Cross-modal redundancy and the geometry of vision-language embeddings. In *Int. Conf. Learn. Represent.*, 2026.

Amil Dravid, Yossi Gandelsman, Alexei A. Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *Int. Conf. Comput. Vis.*, 2023.

Amil Dravid, Yasaman Bahri, Alexei A. Efros, and Yossi Gandelsman. Neuron populations exhibit divergent selectivity with scale. *arXiv preprint arXiv:2606.03990*, 2026.

Nelson Elhage, Tristan Hume, Catherine Olsson, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2021.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. In *Int. Conf. Mach. Learn.*, 2026. arXiv:2602.14486.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in GPT2 language models. *Trans. Mach. Learn. Res.*, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *Int. Conf. Learn. Represent.*, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *Int. Conf. Mach. Learn.*, 2024.

A. Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A. Efros. Back into plato's cave: Examining cross-modal representational convergence at scale. *arXiv preprint arXiv:2604.18572*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Int. Conf. Mach. Learn.*, 2019.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *Int. Conf. Learn. Represent.*, 2016.

Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft COCO: Common objects in context. In *Eur. Conf. Comput. Vis.*, 2014.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *Int. Conf. Learn. Represent.*, 2023.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Adv. Neural Inform. Process. Syst.*, 2018.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *Int. Conf. Learn. Represent.*, 2023.

Jesse Mu and Jacob Andreas. Compositional explanations of neurons. In *Adv. Neural Inform. Process. Syst.*, 2020.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. ASIF: Coupled data turns unimodal models to multimodal without training. In *Adv. Neural Inform. Process. Syst.*, 2023.

Tuomas Oikarinen and Tsui-Wei Weng. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *Int. Conf. Learn. Represent.*, 2023.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, 2021.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Adv. Neural Inform. Process. Syst.*, 2017.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP*, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Adv. Neural Inform. Process. Syst.*, 2022.

Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. Multimodal neurons in pretrained text-only transformers. In *ICCV Workshops*, 2023.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2014.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene CNNs. In *Int. Conf. Learn. Represent.*, 2015.

## REPRODUCIBILITY STATEMENT

We have uploaded the source code at https://anonymous.4open.science/r/crossmodal-rosetta-code to ensure reproducibility. We are planning to open-source it after the conference.

## USE OF LARGE LANGUAGE MODELS

In this work we used AI tools like Anthropic's Claude Code and OpenAI's Codex to write code as well as run experiments and gather results. We also used these tools to draft and edit the text. We have *not* used it to develop the main research ideas or interpret the results. We have reviewed AI-assisted suggestions and take full responsibility for the final contents of this work.

## APPENDIX CONTENTS

**A Further examples of matched pairs** **14**

**B Human concept-agreement evaluation details** **14**

**C Selectivity of matched units** **15**

**D Text-indexed retrieval, per concept** **16**

**E Concept coverage of the matched population** **17**

**F Existence and controls on the reference pair** **18**

**G Correlation threshold sweep results** **18**

**H Seed-stable baselines at every ladder rung** **19**

**I Concept agreement across vision models and judges** **20**

**J Matched pairs survival after removing principal components** **20**




Figure 7: **Illustrative matched pairs (1 of 2).** Each row shows a seed-stable pair between CLIP ViT-B/16 and Pythia-160M through 5 top-activating images and 3 top-activating captions. From the 400 strongest seed-stable pairs, we keep one pair per language neuron and sample at intervals from the resulting 174 pairs. Captions are shortened where needed. No concept labels enter the matching, and the two models share no weights and no input space.

## A Further examples of matched pairs

Figs. 7 and 8 show 12 matched pairs between CLIP ViT-B/16 and Pythia-160M, the reference pair. We rank the 1,103 seed-stable pairs by the smaller of their two correlations, keep the 400 strongest, and keep one pair per language unit, which leaves 174 pairs because the strongest language units each have several vision partners. We then take every fourteenth pair down that list to show a diverse set of pairs. Captions are ranked by the mean-pooled activation, the statistic used for matching and for selecting the judge's stimuli. Each row displays 5 images and 3 captions, with captions shortened where needed.

## B Human concept-agreement evaluation details

Each item showed two panels: the nine top-activating images of one CLIP ViT-B/16 neuron and the nine top-activating captions of one Pythia-160M neuron, the same stimuli the judges saw. Raters scored each item from 1 (different concepts) to 5 (the same concept). The 90 items were 30 matched pairs, 30 partner-permuted pairs and 30 random pairs, and raters were not told that conditions existed. The 90 items are a random subset, drawn with a fixed seed, of the 300 items shown to the judges: 30 of the 100 matched pairs with the highest correlations, 30 of the 100 partner-permuted pairs and 30 of the 100 random pairs. Twelve unpaid colleagues, uninvolved in the project and blind to its hypotheses and conditions, rated conceptual similarity from 1 (different concepts) to 5 (same concept) for paired grids of nine top-activating CLIP ViT-B/16 images and nine Pythia-160M captions. Each rater completed 20 screens: 17 study items, two catch items with an obvious answer, and one repeat of an earlier item. The last three only check attention and enter no statistic. A rater is excluded on two or more of a catch item that should score high rated $\leq 2$, a catch item that should

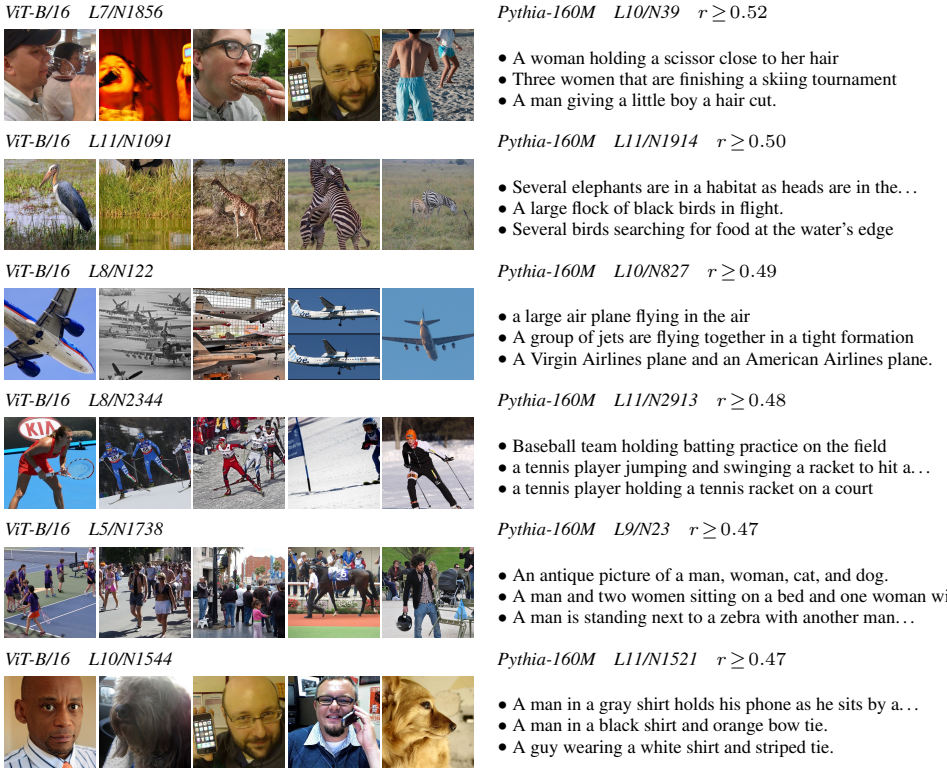


Figure 8: **Illustrative matched pairs (2 of 2).** Continued from Fig. 7, showing lower-ranked pairs within the same selected subset of 174 pairs. The displayed correlations are the minimum across the 2 data samples and range from 0.52 to 0.47. Each row shows 5 top-activating images and 3 top-activating captions, shortened where needed.

score low rated $\geq 4$, a median response time under 2 s, a repeat differing by 3 or more points from its first rating, or zero rating variance, the last of which excludes on its own. No rater met the exclusion rule. Over the 204 study ratings, matched pairs received a mean rating of 3.93, partner-permuted pairs 1.67 and random pairs 1.33. Inter-rater reliability is $\mathrm{ICC}(1,1) = 0.795$ (95% CI 0.713–0.857). The reliability of the item means themselves is $\mathrm{ICC}(1,k_0) = 0.90$.

## C SELECTIVITY OF MATCHED UNITS

To test whether cross-modally matched units are unusually selective, we follow Dravid et al. (2026): excess kurtosis of the vocabulary-projection cosines for language units, and the fraction a blinded judge calls monosemantic for vision units. Because cross-modal matches are spread unevenly across layers and selectivity itself varies by layer, our control units are sampled layer by layer, as many unmatched units per layer as matched ones, rather than from the whole model.

On the language side, excess kurtosis measures how strongly particular token associations stand out from the rest (Fig. 9a). In Pythia-70M and Pythia-160M, matched and control neurons both have relatively high mean scores, with little separation between the groups. Both scores drop sharply at 410M, so we compare the two groups within each size rather than across sizes. Beyond 410M, the matched group's mean generally increases, while controls reach lower values with an increasing gap. On the vision side, approximately 95% of matched neurons are judged monosemantic at every size, while control units are clearly below. These results resemble the selectivity advantage reported for within-modality rosetta neurons by Dravid et al. (2026), although our trends with model size are less consistent. Since we do not measure the overlap between the two selections, whether cross-modal and within-modality matching recover the same neurons remains open.

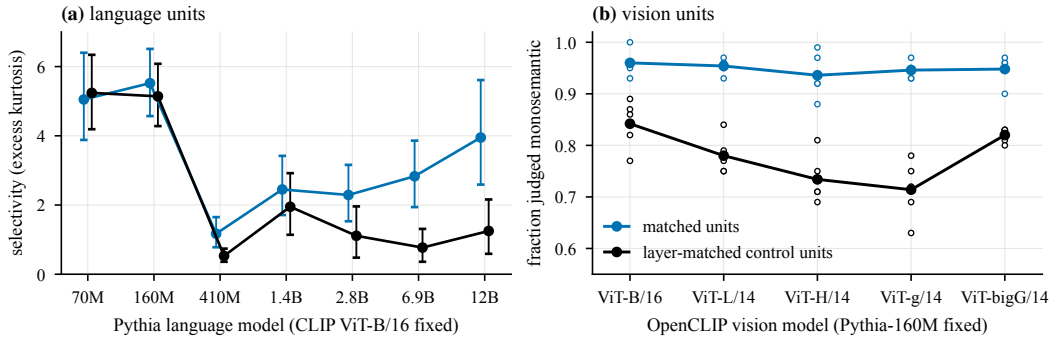


Figure 9: Vocabulary selectivity and judged monosemanticity of matched units (blue) and layer-matched control units (black). **(a)** Language side: selectivity of the matched Pythia units and of control units drawn from the same layers, with bootstrap 95% intervals, as the language model grows. **(b)** Vision side: the fraction of units a blinded judge calls monosemantic, for five disjoint subsets of 100 units per population (hollow markers) and their mean (line), as the OpenCLIP vision model grows.

**Language.** For a neuron with MLP output-projection column $w$, let $s_t = \cos(w, u_t)$, where $u_t$ is the unembedding vector for vocabulary token $t$. Its selectivity is the excess kurtosis $\mathbb{E}_t[(s_t - \bar{s})^4]/\mathbb{E}_t[(s_t - \bar{s})^2]^2 - 3$, computed over the full vocabulary. We include every distinct language neuron in a seed-stable pair and draw an equal number of controls within each layer, without replacement and excluding all neurons matched in either draw. The plotted 95% intervals are percentile intervals from 2,000 bootstrap resamples of individual neurons within each population and model size. Therefore, they quantify uncertainty under neuron resampling and do not model dependence among neurons from one network.

**Vision.** At each size we shuffle the stable vision-unit pool and select 500 units, split into five disjoint subsets of 100. Controls match their aggregate per-layer counts and exclude units matched in either draw. Each judge input is the unit's 20 highest mean-activating images with patch-activation overlays, independently normalized within each image. We ask Qwen2.5-VL-32B whether the highlighted responses represent a single coherent visual feature, requesting a YES/NO answer and a short feature label. We measure the YES fraction in each 100-unit subset. We compare the 5 fractions per population using an exact one-sided Mann–Whitney test.

## D TEXT-INDEXED RETRIEVAL, PER CONCEPT

Our method retrieves images by naming a concept in text and using a matched cross-modal neuron pair to connect the query to the images. For each of the 80 COCO categories, we give Pythia-160M two queries: the category name alone and "a photo of a [name]". We score each of the seed-stable language neurons. For each query, we average the neuron's post-activation responses across tokens, then standardize this value using its mean and population standard deviation from the first discovery draw. We average the two query scores and select the highest-scoring language neuron. We then use its matched CLIP ViT-B/16 neuron to rank images. If the language neuron has several vision partners, we choose the partner with the highest minimum correlation across both discovery draws. Each image's retrieval score is this vision neuron's average activation across image patches.

We evaluate the ranking on 4k held-out images from COCO val2014 that were not used to discover the neuron matches. COCO instance annotations tell us which images contain each category. We compare our method with two single-neuron controls and standard CLIP retrieval. The partner-permuted control uses the vision neuron selected for a different category. The random control uses a randomly selected neuron from the same layer as our selected vision neuron. Standard CLIP instead ranks images by the cosine similarity between its full image embeddings and the text embedding of "a photo of a [name]". All four methods rank the same 4k held-out images. We measure average precision by computing the proportion of correct results up to each relevant image in the ranking

and averaging these values over all relevant images. Fig. 10 shows the top-ranked images for eight concepts. Tab. 2 reports average precision for the six best-performing categories, alongside both controls and standard CLIP. Our results show that a single neuron retrieves some concepts accurately, but performance varies substantially across categories. The median average precision over all 80 categories is 0.045, exceeding the partner-permuted and random-neuron controls (0.024 and 0.030) while remaining far below full CLIP retrieval (0.533).

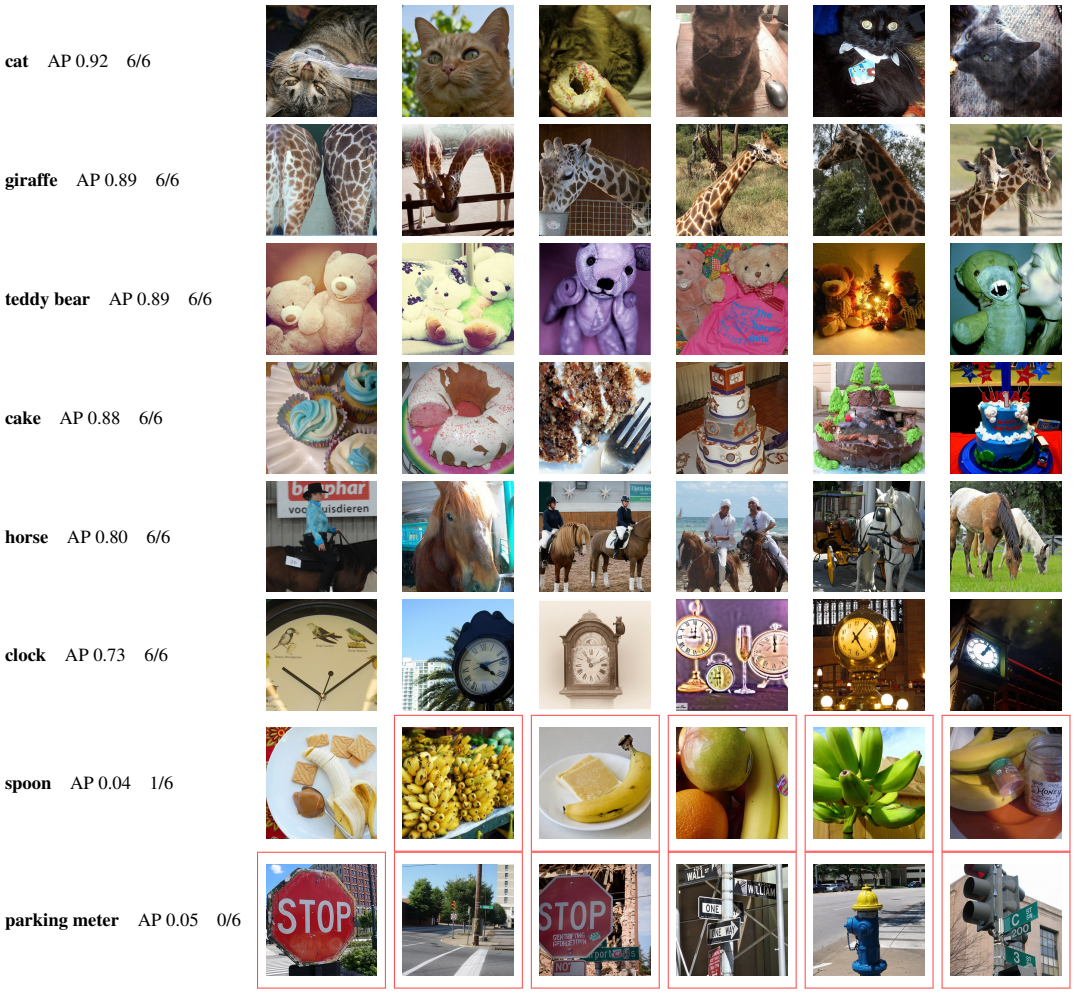


Figure 10: Images retrieved by a single vision neuron chosen from text alone. Each row names a concept. The Pythia-160M neuron it activates most is followed to its matched CLIP ViT-B/16 neuron, and that neuron alone ranks 4k held-out COCO images. Shown are the six highest-ranked images, the average precision over the whole ranking, and how many of the six contain the concept. A red frame marks a miss. The first six rows are the concepts with the highest average precision. The last two are the two concepts closest to the median over all 80 COCO categories.

## E Concept coverage of the matched population

We measure concept coverage by clustering the matched language neurons using their top-activating captions. To do so, each of the Pythia-160M units in the seed-stable pairs is represented by the normalized average sentence embedding of its nine top-activating captions (all-MiniLM-L6-v2, the encoder used for concept agreement). We compute cosine distances, apply agglomerative clustering with average linkage, and cut the dendrogram at distance $t \in \{0.20, 0.30, 0.40, 0.50\}$. Larger distance cutoffs merge more units and therefore yield fewer clusters. As a control we cluster the same

| concept | **matched unit** | partner-permuted | random unit | CLIP embeddings |
|---|---|---|---|---|
| cat | **0.924** | 0.035 | 0.037 | 0.875 |
| giraffe | 0.893 | 0.015 | 0.061 | **0.974** |
| teddy bear | **0.893** | 0.015 | 0.085 | 0.657 |
| cake | **0.881** | 0.028 | 0.018 | 0.753 |
| horse | **0.802** | 0.031 | 0.026 | 0.652 |
| clock | **0.731** | 0.037 | 0.033 | 0.533 |
| median over all 80 categories | 0.045 | 0.024 | 0.030 | **0.533** |

Table 2: Average precision of text-indexed single-neuron retrieval over 4k held-out COCO images sorted by highest retrieval average precision among all 80 COCO categories.

number of unmatched Pythia-160M units, sampled from the same layers. Next to the raw cluster count we give an *effective* count, which is computed as the exponential of the entropy of the cluster sizes (i.e., the number of equally sized clusters that would spread the units as evenly). A few large clusters and many singletons give a small effective count.

| | matched | | control | | |
|---|---|---|---|---|---|
| distance | clusters | effective | clusters | effective | ratio |
| 0.20 | 359 | 234 | 531 | 474 | 0.49 |
| 0.30 | 233 | 126 | 398 | 287 | 0.44 |
| 0.40 | 119 | 62 | 219 | 125 | 0.50 |
| 0.50 | 56 | 31 | 78 | 37 | 0.85 |

Table 3: Clusters of matched and control Pythia-160M units at four cut-off distances.

At threshold 0.40 the matched units form 119 clusters, effectively 62, and the largest holds 43 units, 7% of the population. The ten largest are kitchens, food, trains, ball sports, computer desks, grazing animals, traffic signals, churches and clock towers, bathrooms, and buses (Fig. 4b). The control units form about twice as many clusters, but their own top captions agree with each other much less (mean pairwise similarity 0.20 against 0.33 for matched units). The matched population combines coherent responses within units with coverage of many everyday objects and scenes.

## F Existence and controls on the reference pair

To measure how many matches exceed chance, we compare every count with the same procedure after shuffling the image–caption correspondence, which keeps each neuron's activation distribution but removes the shared content. Two further concerns are that matches could reflect shared input properties instead of content, such as caption length or a single dominant activation direction, and that they could depend on the choice of $K$. Fig. 11a addresses all three on the reference pair, which Sec. 5.1 summarizes. Removing one or five principal components leaves about two thousand pairs per draw, and permuting captions only within caption-length bins keeps the baseline at the level of a full permutation, so neither a dominant direction nor caption length explains the matches. The count grows with $K$, from 145 pairs at $K{=}1$ to 6,287 at $K{=}10$, and exceeds its baseline at every $K$. Finally, 1,103 pairs are matched in both draws, against none under caption permutation. The correlation threshold $\tau$ is examined separately in App. G.

## G Correlation threshold sweep results

Our matching rule adds a minimum correlation $\tau$ to the mutual top-$K$ rule of Dravid et al. (2023). Because $\tau$ is an additional free parameter, we test whether it creates the matched population or only removes weak pairs. To do so, we repeat the matching on the reference pair for $\tau$ from 0 to 0.3, in single draws and with the seed-stability requirement (Fig. 11b and Tab. 4).

Without a threshold ($\tau = 0$) the single-draw count lies far below its caption-permutation baseline. Every unit has nearest neighbours however weak its correlations, so permuted data still produce tens

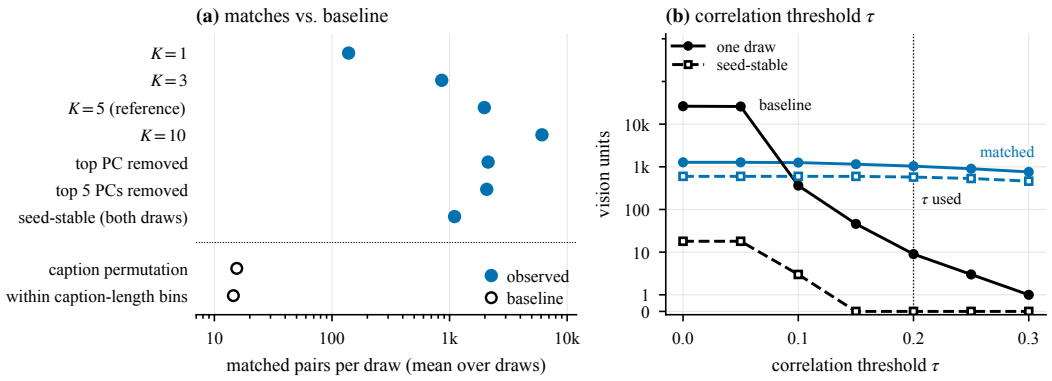


Figure 11: **Controls on the reference pair**, CLIP ViT-B/16 and Pythia-160M, mutual top-5, $r \geq 0.2$. **(a)** Matched pairs under every control (blue), averaged over the two draws of $N$=5000 MS-COCO pairs, against the mean caption-permutation baselines (black, with a maximum of 35 over 30 permutations). The seed-stable baseline is 0 over 20 permutations and is not shown. The sweep over $K$ uses 100 permutations. **(b)** Vision units against the correlation threshold $\tau$, in one draw and seed-stable, each with its baseline. This sweep uses disjoint draws from the pooled COCO splits, so its counts differ slightly from (a) (App. G).

of thousands of mutual pairs. Raising $\tau$ removes these, and from $\tau = 0.1$ the single-draw count exceeds its baseline. The seed-stable count never has this problem: at $\tau = 0$ it is 598 units against a baseline of at most 18, and it barely changes up to $\tau = 0.2$, because a spurious pair rarely survives two independent draws of the data. The threshold therefore does not create the population but stabilizes the single-draw count. Therefore, we keep it at 0.2. Dravid et al. (2023) need no threshold because within a modality both networks see the same input and the correlations are high enough for mutual top-$K$ alone to separate matches from non-matches. Across modalities the correlations are weaker, and seed stability does the separating.

| | single draw | | seed-stable | | seed-stable baseline | |
|---|---|---|---|---|---|---|
| $\tau$ | units | baseline | units | pairs | mean | max |
| 0.00 | 1,276 | 26,323 | 598 | 1,033 | 5.78 | 18 |
| 0.05 | 1,276 | 25,953 | 598 | 1,033 | 5.76 | 18 |
| 0.10 | 1,256 | 364 | 598 | 1,033 | 0.16 | 3 |
| 0.15 | 1,152 | 46 | 597 | 1,032 | 0.00 | 0 |
| **0.20** | 1,037 | 9 | 575 | 1,005 | 0.00 | 0 |
| 0.25 | 902 | 3 | 533 | 945 | 0.00 | 0 |
| 0.30 | 759 | 1 | 461 | 850 | 0.00 | 0 |

Table 4: Sweep of the correlation threshold $\tau$ on the reference pair, CLIP ViT-B/16 and Pythia-160M, $N$=5000 pairs per draw, mutual top-5, 100 caption permutations. "Single draw" gives the strong vision units in one draw and the mean of its caption-permutation baseline, and "seed-stable" requires the pair in both draws. The bold row is the value used for the main results. The counts differ from Sec. 5.1 because this sweep uses disjoint draws from the pooled train and val splits (Sec. 5.1), whereas the reference results use the original val-only draws.

## H SEED-STABLE BASELINES AT EVERY LADDER RUNG

Each rung of each ladder is compared with its own caption-permutation baseline, computed on the same seed-stable statistic that the fits use. Each draw is permuted with its own random permutation, the pairs found under the two permutations are intersected, and this is repeated 20 times. The two draws are different samples of images, so a shared permutation would not correspond to a shared corruption. Permuting them independently is also the stricter choice, since a corruption common to

both draws would let a spurious pair repeat. Table 5 lists all 24 rungs. At all 24 model settings, the seed-stable match count exceeds the largest count observed under caption permutation.

| ladder | rung | units | pairs | baseline max, units | baseline max, pairs |
|---|---|---|---|---|---|
| language (ViT-B/16 fixed) | Pythia-70M | 459 | 863 | 0 | 0 |
| | Pythia-160M | 612 | 1,103 | 0 | 0 |
| | Pythia-410M | 855 | 1,486 | 0 | 0 |
| | Pythia-1.4B | 1,084 | 1,803 | 0 | 0 |
| | Pythia-2.8B | 1,143 | 1,856 | 0 | 0 |
| | Pythia-6.9B | 1,303 | 2,073 | 0 | 0 |
| | Pythia-12B | 1,368 | 2,231 | 0 | 0 |
| vision (Pythia-160M fixed) | ViT-B/16 | 596 | 1,069 | 0 | 0 |
| | ViT-L/14 | 731 | 1,296 | 0 | 0 |
| | ViT-H/14 | 724 | 1,263 | 0 | 0 |
| | ViT-g/14 | 786 | 1,249 | 0 | 0 |
| | ViT-bigG/14 | 876 | 1,467 | 0 | 0 |
| both models grow | B/16 $\times$ 70M | 430 | 804 | 0 | 0 |
| | L/14 $\times$ 410M | 960 | 1,677 | 0 | 0 |
| | H/14 $\times$ 1.4B | 1,260 | 2,195 | 0 | 0 |
| | g/14 $\times$ 6.9B | 1,579 | 2,669 | 1 | 1 |
| | bigG $\times$ 12B | 2,495 | 4,336 | 5 | 6 |
| DINOv2 (text-free) | ViT-S/14 | 505 | 857 | 0 | 0 |
| | ViT-B/14 | 614 | 1,055 | 0 | 0 |
| | ViT-L/14 | 733 | 1,207 | 0 | 0 |
| | ViT-g/14 | 740 | 1,228 | 0 | 0 |
| MAE (text-free) | ViT-B/16 | 429 | 781 | 0 | 0 |
| | ViT-L/16 | 633 | 998 | 1 | 1 |
| | ViT-H/14 | 712 | 1,104 | 0 | 0 |

Table 5: Seed-stable units and pairs at every rung, against the maximum of a seed-stable caption-permutation baseline over 20 permutations, $\tau$=0.2, $K$=5.

## I Concept agreement across vision models and judges

To evaluate whether the matched units share concepts, we ask 2 VL judges to name the concepts represented by each unit's top-activating images or captions. Each unit is labelled in its own call, so the judge never sees a partner. We calculate label similarity as the cosine similarity between the all-MiniLM-L6-v2 sentence embeddings of the two independently generated neuron labels. We compare the pairwise label similarities with those of two controls: a partner-permuted control that pairs each unit with the other unit's partner, and a random control that pairs each unit with a random unit from the same layer. Table 6 reports the median similarity across pairs in each condition and the AUC from one-sided Mann–Whitney $U$ comparisons for the 3 vision models and both judges. For both judges and all 3 vision encoders, matched pairs have greater concept agreement than either control. For the 32B judge, random pairs have a higher median label similarity than partner-permuted pairs for all three encoders. For MAE, the AUC against random pairs is also lower than against partner-permuted pairs.

## J Matched pairs survival after removing principal components

Removing the leading principal components tests how matching depends on dominant activation directions. Section 5.1 shows that a large matched population remains after this removal. Here we ask whether the *same* pairs survive. For the reference pair, CLIP ViT-B/16 and Pythia-160M, we match both draws with $q = 0$, 1 and 5 components removed and compare the seed-stable pairs (Tab. 7).

Removing one component leaves the population intact and keeps most of the same pairs. After removing five principal components, the first draw retains 2,155 strong pairs, compared with a caption-

| | | median label similarity | | | AUC | |
|---|---|---|---|---|---|---|
| vision encoder | judge | matched | random | permuted | vs. rand. | vs. perm. |
| CLIP ViT-B/16 | Qwen2.5-VL 7B | 0.822 | 0.241 | 0.305 | 0.946 | 0.869 |
| | Qwen2.5-VL 32B | 0.596 | 0.210 | 0.139 | 0.910 | 0.884 |
| DINOv2 ViT-B/14 | Qwen2.5-VL 7B | 0.755 | 0.240 | 0.279 | 0.940 | 0.809 |
| | Qwen2.5-VL 32B | 0.569 | 0.223 | 0.141 | 0.903 | 0.850 |
| MAE ViT-B/16 | Qwen2.5-VL 7B | 0.581 | 0.235 | 0.271 | 0.874 | 0.820 |
| | Qwen2.5-VL 32B | 0.473 | 0.232 | 0.173 | 0.768 | 0.847 |

Table 6: Concept agreement for the 100 seed-stable pairs with the highest correlations and 100 pairs per control condition, tested with a one-sided Mann–Whitney $U$.

| $q$ | strong pairs, draws 1 / 2 | stable pairs | vision units | language units | shared with $q$=0 | kept |
|---|---|---|---|---|---|---|
| 0 | 2,071 / 1,879 | 1,103 | 645 | 612 | – | – |
| 1 | 2,223 / 2,034 | 1,124 | 671 | 626 | 861 | 78% |
| 5 | 2,155 / 1,988 | 955 | 507 | 466 | 434 | 39% |

Table 7: Seed-stable pairs after removing the top $q$ principal components from both sides, and how many of the original ($q$=0) pairs remain.

permutation baseline of 28.5 pairs on average. Across both draws, 955 pairs remain seed-stable, including 434 of the original 1,103 pairs (39%).