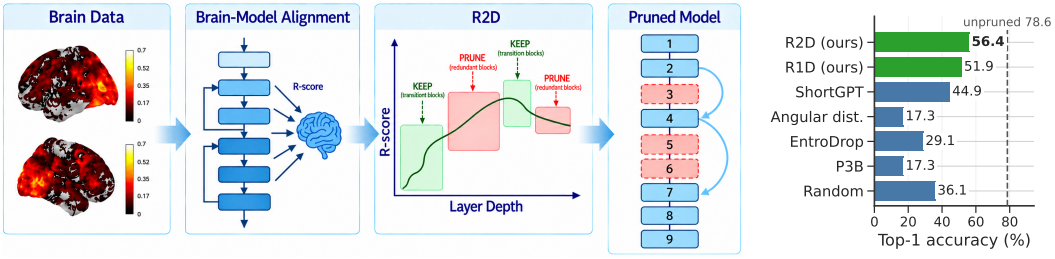
BRAIN DATA AS A BLOCK-IMPORTANCE METRIC FOR DEPTH PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain-model alignment quantifies how well a deep neural network's internal representations correlate with human brain activity. It remains unclear however whether this alignment actually reflects the functional importance of individual deep network components. In this work, we investigate whether changes in this correlation between brain data and the model's internal representations can be used to infer block importance. Using per-block brain-model alignment scores computed on the Natural Scenes Dataset, we derive two model-agnostic block-importance metrics that capture the first derivative and second derivative of the alignment score, denoted R1D and R2D, respectively. We evaluate R1D and R2D through structured depth pruning by ranking blocks according to these metrics, removing the lowest-scoring blocks, and comparing the resulting performance against conventional pruning baselines. Across multiple models and vision tasks, R2D matches or outperforms several other block-importance metrics for pruning, with the largest margins at moderate compression rates. Notably, after pruning half the blocks of the Swin Small model using our brain-inspired metrics, it retains 61.5% accuracy on ImageNet without any additional finetuning, against 26.9% for the strongest depth pruning baseline. Our work demonstrate that brain-model alignment provides a direct signal that allows for more efficient model design.

1 INTRODUCTION



(a) Brain-guided pruning

(b) Pruning results

Figure 1: **Brain alignment tells us which transformer blocks a model can prune to maintain the most accuracy. (a)** We compute the second derivative of the brain-model alignment score and use it to guide model pruning. **(b)** Our brain-inspired metrics outperform several baselines on COCO classification after pruning 6 of the 12 blocks in DeiT-Base (87M) with no additional finetuning.

A substantial body of work at the intersection of neuroscience and AI has shown that deep neural networks develop representations during training that correlate strongly with those of the human brain (DiCarlo et al., 2012; Yamins et al., 2014; Schrimpf et al., 2018; Caucheteux & King, 2022; Raugel et al., 2026). This convergence may arise from the high flexibility of these two learning systems, which, when exposed to sufficient data across evolution, learning or training, may converge toward similar computational solutions for processing the same input (Mehrer et al., 2020; Hasson et al., 2020; Huh et al., 2024; van Rossem & Saxe, 2024; Cagnetta et al., 2024; Mahner et al., 2025).

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.

Crucially, if these two kinds of learning systems converge towards shared computational mechanisms, then the more optimal one could serve as template to steer the other system toward more efficient representations and computations. The human brain, shaped by millions of years of evolution and a lifetime of experience and supporting the full range of human cognitive abilities on an energy budget of $\sim 20W$ (Balasubramanian, 2021), might therefore serve as a good template to steer our deep learning models towards better compression. In other words, representational convergence opens a prescriptive path where biological mechanisms can be directly exploited to improve the efficiency of artificial ones.

In this work, we investigate whether brain-model alignment captures the functional importance of individual blocks in vision models. We address this question by translating changes in block-wise brain-encoding scores into quantitative block-importance metrics, which we then evaluate through structured depth pruning, as Figure 1a illustrates. If our brain-inspired metrics accurately measure the importance of the blocks in a vision model, preserving the highest ranking blocks according to these metrics should lead to the least degradation in model capabilities after pruning. We demonstrate in this work that our brain-inspired metrics do indeed match or outperform several activation-based depth pruning baselines in multiple settings, including the one highlighted in Figure 1b.

We make the following contributions:

1. We analyze the brain–model alignment at the block level across network depth, revealing a consistent, multi-phase trajectory whose organization follows the cortical hierarchy.
2. We introduce two brain-derived, model-agnostic block-importance metrics: R1D, capturing brain-model alignment change across depth, and R2D, capturing alignment change rate.
3. We evaluate R1D and R2D for structured depth pruning across model families, scales, and datasets, and demonstrate that R2D is competitive with or outperforms established activation-based pruning baselines across all pruning strengths. R1D also shows considerable performance improvements, especially for DINOv3 at the small scale.
4. We isolate the role of neural data from the second-derivative, by applying the same second-derivative operator to activation-based metrics, showing that the brain-signal is necessary for the good performance achieved by R2D.

2 Related Work

Brain-AI Alignment. Deep neural networks were designed for industrial applications, not to mimic the brain. It is surprising then that they develop internal representations through their training that resemble those observed in the mammalian visual system (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Guclu & van Gerven, 2015; Cichy et al., 2014; Simkova et al., 2025). In particular, the bottom-up hierarchy of representations across layer depth in vision models has repeatedly been found to align with the cortical hierarchy of early-to-late visual processing (Raugel et al., 2026; Wang et al., 2023). We leverage this alignment to build block-importance metrics using neural data.

Brain-Guided Model Design. Anand et al. (2021) and Oota et al. (2026) showed that conventionally pruned or compressed networks can preserve or even improve alignment with cortical data. Beyond measuring brain-model alignment, recent works have investigated whether neural data can inform how artificial networks are modified across training and post-training. For example, Moussa et al. (2025); Xiao et al. (2026) have demonstrated the utility of brain signals for guiding model training. In addition, Tarigopula et al. (2023) used human behavioral and neural similarity spaces to prune individual representational dimensions of DNN representations, isolating features that better capture human representational geometry. While this finding is promising, it remains unclear how the capabilities of such brain-guided compressed models compare to conventional pruning baselines.

Block Importance and Pruning. Previous work has developed block-importance metrics for structured depth pruning. Men et al. (2025) measured block influence using cosine similarity between input and output representations, assigning low importance to blocks that apply near-identity transformations. Closely related, Gromov et al. (2025) used the angular distance between representations before and after a candidate contiguous span of blocks, with smaller distances indicating lower importance. Similarly, Yang et al. (2025) ranked blocks according to changes in representation entropy, while Glandorf & Rosenhahn (2025) used a task-dependent loss-based measure. All follow

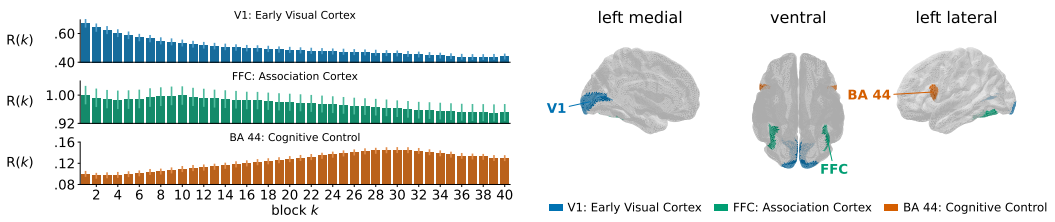


Figure 2: **Brain-Model Alignment peaks at different depths for different parts of the cortex. Left:** per-block scores $R_i(k)$ of DINOv3-7B for V1 (primary visual cortex), FFC (fusiform face complex), and BA 44 (association area). **Right:** relevant regions plotted on the cortical surface. As we move up the cortical hierarchy (from V1 to BA 44), each curve peaks later in block progression.

the same general procedure: compute block importance, rank blocks, and prune the least important ones. That such removal is possible at all reflects substantial redundancy in trained networks, and this compressibility is itself predictive of generalization in both vision and language models (Zhou et al., 2018; Lotfi et al., 2022; 2024; 2026).

Positioning and Impact of this Work. Prior work motivates our central question: can brain–model alignment guide identification of the most functionally important blocks in a vision model? To answer this question, we conduct a thorough analysis of the brain–model alignment signal in Section 3 and introduce two block-importance metrics based on variations in this alignment across network depth. We evaluate these metrics through structured depth pruning in Section 4, ranking blocks and removing the lowest-scoring ones, and show that our brain-inspired metrics due indeed outperform standard pruning baselines in several settings that cover multiple models, datasets, and pruning strengths. To the best of our knowledge, this is the first evaluation of brain-guided pruning that compares to conventional baselines on standard benchmarks. Our work demonstrates that brain-alignment surprisingly helps achieve more efficient model design.

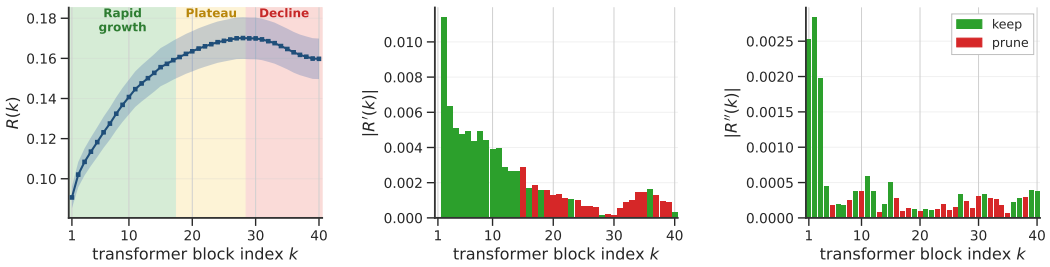
3 Methods: Translating Brain-Model Alignment into Block-Importance Metrics

3.1 Quantifying brain-model alignment for each model block

Following Raugel et al. (2026), for a model with K residual blocks, we take the output of each block $k \in \{1, 2, \ldots, K\}$ and treat it as one candidate representation, so a K-block model gives us K representations to score. We obtain brain responses to natural images from the Natural Scenes Dataset (NSD) (Allen et al., 2022), an fMRI dataset in which 8 participants each viewed about 10,000 natural scenes. Following Raugel et al. (2026) again, we use a standard mesh covering the cortex, the outer sheet of the brain where visual processing takes place. Each vertex of that mesh carries one voxel, which is the smallest volume of cortex that fMRI resolves.

For each block k, we fit one linear encoding model per voxel, mapping model activations $X \in \mathbb{R}^{n \times d}$ to the fMRI responses $Y \in \mathbb{R}^{n \times m}$ of m voxels across n images: $\arg\min_W \ \|Y - XW\|_2^2 + \lambda \|W\|_2^2$, with 10 logarithmically-spaced regularization values $\lambda \in [10^0, 10^8]$ and per-voxel regularization selection. We evaluate performance via 5-fold cross-validation, where each fold yields a Pearson correlation R between predicted and held-out voxel responses, averaged across folds and subjects, using packages from Gramfort et al. (2013); King et al. (2026). This procedure yields, for each block k, a brain-encoding score $R(k)$ that quantifies how well that block's representation linearly predicts the neural code.

Regions of interest. We select a representative set of 15 regions of interest (ROIs) spanning the anatomy of the cortex from visual system in posterior-occipital lobe to prefrontal cortex (Felleman & Van Essen, 1991). These ROIs are V1, V2, V3, V4, FFC, TPOJ2, TPOJ1, TE2p, STV, TE1p, IFJp, IFJa, IFSp, IFSa, 44 (Glasser et al., 2016). They span early retinotopic cortex, category-selective ventral temporal regions including the fusiform face complex (Kanwisher et al., 1997), temporo-parieto-occipital junction areas, and lateral prefrontal regions associated with cognitive control (Sundermann & Pfleiderer, 2012).



(a) Three-phase in R-score (b) R1D block-importance (c) R2D block-importance

Figure 3: **Brain-inspired metrics for block importance.** All panels use DINOv3-7B and are averaged over 8 subjects and 15 ROIs regions. **(a)** The brain-encoding curve $R(k)$ rises steeply over the early blocks, plateaus, then declines. **(b)** R1D, magnitude of the first derivative, is largest across the growth phase. Colours give the greedy pruning decision at 50% compression, green for keep and red for prune. R1D therefore marks every early block as important and removes a contiguous run from the second half. **(c)** R2D, the magnitude of the second derivative, yields a more sparse block selection for pruning, particularly preserving blocks that mark phase transitions.

3.2 Alignment across depth follows the cortical hierarchy

Writing $R_i(k)$ for the score of block k averaged solely over the voxels comprising region i gives us a mapping between model depth and the cortical hierarchy, as per prior work (Raugel et al., 2026; Wang et al., 2023; Conwell et al., 2022). Early blocks best predict early visual areas such as V1, V2 and V3, middle blocks best predict ventral temporal regions such as V4 and FFC, and late blocks best predict prefrontal regions, such as IFJa, IFJp, and BA 44. Figure 2 shows this for three regions of DINOv3-7B (Siméoni et al., 2025). The peak alignment shifts across the network rather than maintaining a single depth for all regions; the depth of the peak block corresponds to the cortical order from early-visual processing to later associative areas.

3.3 Defining first and second derivative metrics for block importance

Averaging $R_i(k)$ across the 15 ROIs from Section 3.2 yields a single curve summarizing block-level alignment. Figure 3a shows this curve for DINOv3-ViT-7B (Siméoni et al., 2025), and it has three phases. *(1) Early growth*: $R(k)$ rises rapidly across the first blocks as low-level visual features build up. *(2) Middle plateau*: $R(k)$ barely changes from one block to the next, so consecutive blocks predict brain responses about equally well. *(3) Mild decline*: $R(k)$ decreases slightly in the final blocks, consistent with late blocks specializing to the training objective. We find the same three phases across different DINOv3 (Siméoni et al., 2025) sizes and for the DeiT (Touvron et al., 2021) and Swin (Liu et al., 2021) models as well, with shorter plateaus and decline phases for shallower models, as shown in Section B.1.

Our first hypothesis is that a block may be redundant when it leaves the brain-alignment unchanged. We measure this notion using the first derivative of the R-score, R1D, and compute the finite-difference approximation of this first derivative as follows:

$$\mathrm{R1D}(k) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \big| R_i(k{+}1) - R_i(k) \big|, \tag{1}$$

where $\mathcal{I}$ is the set of ROIs, $R_i(k)$ is the mean Ridge R-score for ROI i at block k, and $k{+}1$ is the next remaining block (the last block is scored by its mean R-score).

Our second hypothesis is that blocks in low-curvature regions of the R-score across depth, where neighboring R-scores can be approximated from one another with a linear transformation, may belong to the same representational stage. Therefore these blocks may be good candidates for pruning. High curvature, in contrast, indicates a departure from this local linear trend in alignment. This notion aligns with the distinct representational phases across depth in vision models discussed in Neyshabur et al. (2020), where they argue that early layers are responsible for learning general features and later layers are specialized to the target task. If representational phases are separated in

Table 1: **Block importance baselines.** We use x_k to denote the input to block k and x_{k+1} its output. Baselines include: block influence from ShortGPT (Men et al., 2025); angular distance on the `[class]` token (Gromov et al., 2025); entropy increase from EntroDrop (Yang et al., 2025), where $\bar{x}_k$ is x_k averaged over all tokens and H is a nearest-neighbour estimate of its entropy across the calibration images; and the block performance indicator from P3B (Glandorf & Rosenhahn, 2025), with $\mathcal{L}$ the cross-entropy loss and h_k a linear probe on block k. *Selection*: *Greedy* recomputes the score on the remaining blocks after each removal and drops the lowest; *Static* scores all blocks on the unpruned model and drops the n lowest; *Span* removes the n consecutive blocks $k, \ldots, k{+}n{-}1$ with the smallest distance. Random is averaged over 10 seeds. Full definitions are in Section A.1.

Method	Block Importance Score	Selection	Brain data
R2D (ours)	$\lvert R_i(k{-}1) - 2R_i(k) + R_i(k{+}1)\rvert$	Greedy	✓
R1D (ours)	$\lvert R_i(k{+}1) - R_i(k)\rvert$	Greedy	✓
ShortGPT	$1 - \mathbb{E}[\cos(x_{k,t}, x_{k+1,t})]$	Static	–
Angular dist.	$\frac{1}{\pi}\,\mathbb{E}[\arccos\cos(x_k, x_{k+n})]$	Span	–
EntroDrop	$H(\bar{x}_{k+1}) - H(\bar{x}_k)$	Static	–
P3B	$\mathbb{E}[\mathcal{L}(h_k(x_{k-1})) - \mathcal{L}(h_k(x_k))]$	Static	–
Random	random permutation	Static	–

such a manner, it is intuitive to want to preserve the blocks that connect different phases. The second derivative of the R-score may capture these transitions, allowing us to identify such blocks. We compute the finite-difference approximation of the second derivative of the R-score as follows:

$$\mathrm{R2D}(k) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\big|R_i(k+1) - 2R_i(k) + R_i(k-1)\big|, \tag{2}$$

where again $\mathcal{I}$ is the set of ROIs and $R_i(k)$ is the mean Ridge R-score for ROI i at block k. For both R1D and R2D, we average across the 8 NSD subjects for each block and ROI, take the first and second derivatives, then the absolute value, then the mean across ROIs. Both metrics are recomputed on the remaining blocks after each removal, and neither one uses labels, gradients, or task data.

Figure 3b shows R1D on DINOv3-7B (Siméoni et al., 2025). R1D is highest across the early phase, where the score increases quickly, so it marks almost every early block as important. It is smaller across the plateau, but critically also small at the block where the curve transitions from plateau into decline, so R1D treats a transition block the same as the flat blocks beside it. In contrast, Figure 3c shows that the R2D peaks where the curve bends and drops to near zero across the middle of the plateau. Ranking by R2D and removing the lowest scores therefore keeps the transition blocks. We evaluate both metrics in what follows since they capture different hypotheses.

4 Results: Brain Data Effectively Guides Depth Pruning

4.1 Experimental setup

We use depth pruning to evaluate the effectiveness of our block-importance metrics. We prune the lowest-ranked blocks under each metric and its intended block selection rule. We compare our brain-derived metrics to established depth pruning baselines. Whenever we remove a block, we remove both the attention and feed-forward sublayer together.

Greedy pruning for brain-inspired metrics. We fix a compression computational budget and prune a fixed number of blocks n. For R1D and R2D, rather than ranking all blocks once and dropping the bottom n, we apply the criterion iteratively: at each step, we recompute the metric on the remaining blocks, prune the least important one, and repeat. Once we prune a block, we measure the new R1D and R2D at its neighbors under the new architecture rather than the original geometry. This greedy computation takes a few seconds to run since the R-score remain unchanged. The baselines follow their original, intended selection rules.

Baselines. Table 1 lists all the block-importance metrics, including the baselines. We treat random selection as a strong baseline rather than a worst-case scenario, since previous work on weight and channel pruning finds random masks to be far stronger than their simplicity suggests (Su et al.,

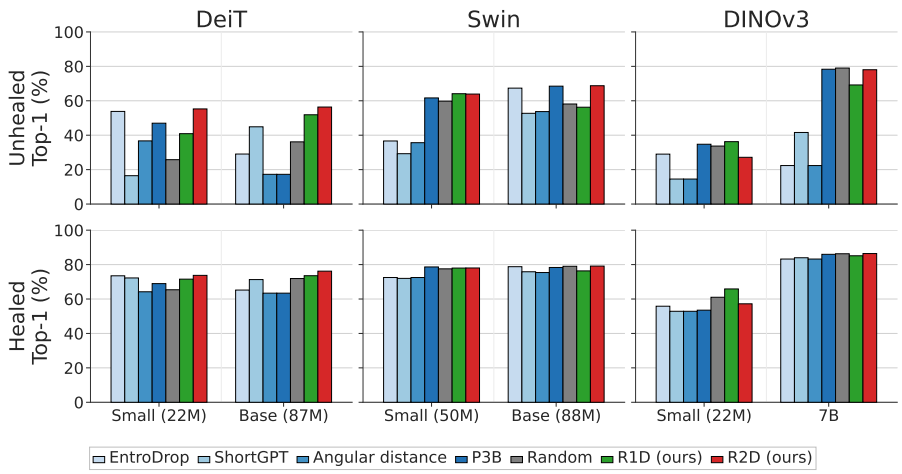


Figure 4: **Brain-inspired metrics perform competitively on data in-distribution to the images the participants saw, across multiple models and scales.** Top-1 accuracy (%) on COCO at 50% pruning with and without healing. Our brain-inspired metrics outperform the baselines in 10 our of these 12 settings, with R1D leading on DINOv3 Small and R2D on DeiT and on Swin Base.

2020; Frankle et al., 2020; Li et al., 2022; Liu et al., 2022; Gadhikar et al., 2023). We provide a pseudo-code for the computation and evaluation of these block-importance metrics in Algorithm 1.

Healing. In practice, one may be willing to incur additional computational costs after pruning to finetune the model in order to improve its performance while saving on inference costs. We refer to this post-pruning finetuning as *healing*, similarly to related work (Gromov et al., 2025). Therefore, we evaluate every configuration twice. First, the unhealed version: we prune the blocks and measure the accuracy directly after pruning. Second, the healed version: we prune the blocks, finetune the resulting pruned network with LoRA (Hu et al., 2022), and then measure the accuracy after finetuning. This way, we can evaluate both the performance of the block selection itself and whether the quality of block selection still matters once the network can adapt.

Models and vision tasks. We evaluate three model families that differ in architecture and training objective: DINOv3 (Siméoni et al., 2025), a self-supervised vision transformer, DeiT (Touvron et al., 2021), a supervised vision transformer, and Swin (Liu et al., 2021), a supervised hierarchical transformer. We cover 3 sizes for each model family. We evaluate classification performance on COCO (Lin et al., 2014), which is in-distribution for the images seen by the participants for fMRI data collection, as well as ImageNet-1k (Deng et al., 2009; Russakovsky et al., 2015), CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and DTD (Cimpoi et al., 2013).

4.2 Brain-derived metrics are competitive in-distribution

We first evaluate our brain-inspired metrics on the same distribution as the one the participants saw, by building a superset of the COCO subset used in NSD with no overlap between the images seen by the participants and the data used for finetuning and evaluation. Our rationale is that if the brain data carries a prescriptive signal for pruning, it would most likely do so for in-distribution data to what the participants saw. For all block-importance metrics we follow Algorithm 1: we finetune a head on the training data, make forward passes with a class-balanced 5% of the training data to compute the metrics and the rankings they induce, and evaluate the pruned model on the validation data. R1D and R2D are computed from the R-scores on COCO alone.

Figure 4 shows that at 50% pruning, our metrics are competitive across all three model families and different scales. Without healing, R2D achieves 55.3% accuracy on DeiT small and 56.4% on DeiT base, outperforming all baselines, the best of which achieves 53.8% and 44.9% respectively. R1D also outperforms every baseline on DeiT base, at 51.9%. With healing, R2D still outperforms all baselines on both DeiT sizes. Similarly on Swin small and base, R2D is on par with or outperforms all baselines, and R1D is the strongest of the seven metrics on Swin small without healing at 64.1%.

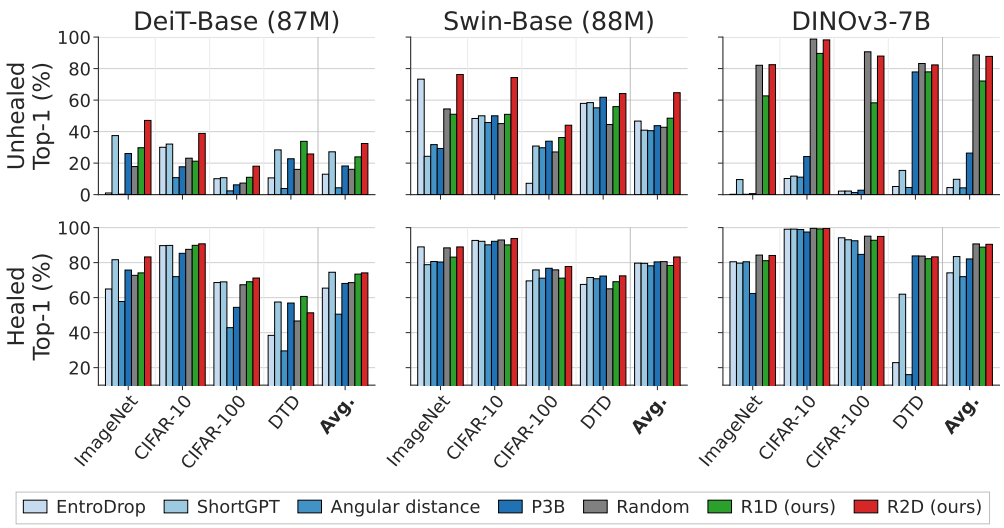


Figure 5: **Brain-metrics computed only from the images the participants saw successfully transfer to datasets out-of-distribution for the brain data.** Top-1 accuracy (%) at 50% pruning on the four held-out datasets and their mean, for DeiT base, Swin base and DINOv3 7B. Top row: no healing. Bottom row: LoRA healing on each dataset. The activation baselines are computed on a class-balanced 5% of each target dataset's own training split, whereas R1D and R2D are computed once from the COCO images in NSD and never see data from these datasets.

Overall, one of our two metrics is the best in 10 of the 12 settings. R1D performs best for DINOv3 small while R2D performs better for DINOv3 7B. We hypothesize that this is because DINOv3 small is the only model in Figure 4 that does not have a decline phase as shown in Figures 7 to 9. In fact, its plateau runs to the final block, whereas R2D has two transitions to preserve for all other models. DeiT and Swin are moreover trained with label supervision, which drives late blocks to specialize to the class posterior, whereas all of DINOv3's sizes below 7B are distilled from the 7B teacher, which may be why smaller DINOv3 models have shorter or inexistent decline phases.

4.3 Brain-derived block-importance transfers out of distribution

While activation-based baselines are computed on exactly the distribution they are evaluated on as in Algorithm 1, which is a faithful implementation of what their authors intended, our brain-inspired metrics can only be computed from the R-score averaged over the COCO images each participant saw. It would therefore be impressive if an R-score computed on COCO could tell us how to prune a model for other datasets that are potentially out-of-distribution.

Figure 5 shows that indeed, without healing, R2D outperforms all baselines on average over four out-of-distribution datasets, achieving 32.5% accuracy on DeiT base and 64.7% on Swin base, where the best baselines achieve 27.2% and 46.7%. With healing, R2D still outperforms all baselines on Swin base at 83.2% over 80.6%, and remains competitive on DeiT base at 74.1%. On DINOv3 7B, random selection is a competitive baseline with and without healing, which was observed in prior works as well (Su et al., 2020; Frankle et al., 2020; Li et al., 2022; Liu et al., 2022; Gadhikar et al., 2023). We look more in-depth into why random selection is competitive in Section 5.

The fact that brain-derived metrics generalize out-of-distribution for classification and outperform activation-based baselines in several settings implies that the brain signal captures something essential about vision tasks that goes beyond the particular images the participants were viewing. This is not only a significant result but one that is also practical since one can collect fMRI responses once, as in the publicly available data we use here, and reuse them to prune any model on any dataset at no additional data collection cost.

Table 2: **Brain-inspired pruning metrics perform on-par or better on average for top-1 accuracy (%) averaged over COCO, ImageNet, CIFAR-10, CIFAR-100 and DTD, after healing.** We vary model size and pruning strength for DeiT, Swin and DINOv3 at different sizes. – : Angular distance has no valid pruning for Swin at 75%; it removes one contiguous run of blocks, which must lie within a single Swin stage, and no stage can hold a run of 18 blocks (see Section A.4). Within each row, cells within 1% of the best are shaded green and within 1% of the worst are shaded red.

Family	Size	Pruning strength (%)	EntroDrop	ShortGPT	Angular Distance	P3B	*Random*	**R1D (ours)**	**R2D (ours)**
DeiT	Tiny 5.7M	25%	**71.4**	70.6	66.5	**72.1**	68.1	71.0	**71.1**
		50%	57.5	54.8	44.1	57.8	51.2	**58.5**	**59.3**
		75%	37.1	35.4	26.7	35.1	31.1	**39.5**	35.5
	Small 22M	25%	**80.5**	**79.8**	73.4	**80.3**	75.8	**79.9**	**79.9**
		50%	**70.7**	66.1	51.5	66.4	59.3	67.7	**69.8**
		75%	44.6	41.6	36.1	41.7	36.2	**46.7**	40.6
	Base 87M	25%	82.4	**84.7**	81.9	82.8	82.8	79.6	83.4
		50%	65.4	**73.9**	53.1	67.2	69.2	**73.5**	**74.5**
		75%	43.0	43.6	42.0	43.5	41.9	40.5	**46.7**
Swin	Small 50M	25%	80.7	**83.5**	83.0	**84.2**	**84.0**	**84.2**	**84.5**
		50%	66.8	71.7	68.6	71.6	77.1	**78.2**	**77.5**
		75%	41.0	**50.5**	–	48.7	47.0	45.3	47.4
	Base 88M	25%	84.2	**85.9**	**85.0**	**85.4**	**84.9**	**85.9**	**85.7**
		50%	79.5	78.8	77.6	80.0	80.3	78.0	**82.4**
		75%	54.0	55.6	–	**57.9**	51.6	48.5	**58.0**
	Large 197M	25%	**87.9**	87.1	86.8	87.8	**88.0**	87.6	**88.9**
		50%	78.1	75.0	70.6	76.1	**82.4**	80.0	**82.9**
		75%	48.8	**53.3**	–	50.9	50.5	51.5	49.0
DINOv3	Small 22M	25%	54.8	62.9	56.3	64.7	70.7	**72.3**	67.2
		50%	38.5	32.5	32.5	35.2	43.5	**55.4**	38.3
		75%	**32.9**	28.5	28.3	27.7	**32.9**	31.2	31.2
	Large 303M	25%	76.8	**87.3**	82.5	84.0	86.0	85.3	**86.3**
		50%	46.6	56.6	48.4	61.3	68.2	**70.1**	**69.8**
		75%	38.4	40.2	34.6	40.4	40.1	**42.4**	**43.0**
	Largest 7B	25%	**90.0**	**89.9**	**90.2**	89.0	**90.1**	**90.0**	**90.1**
		50%	76.0	83.6	74.2	82.9	**89.8**	88.1	**89.7**
		75%	48.2	43.6	40.5	72.5	**84.4**	81.4	**84.0**

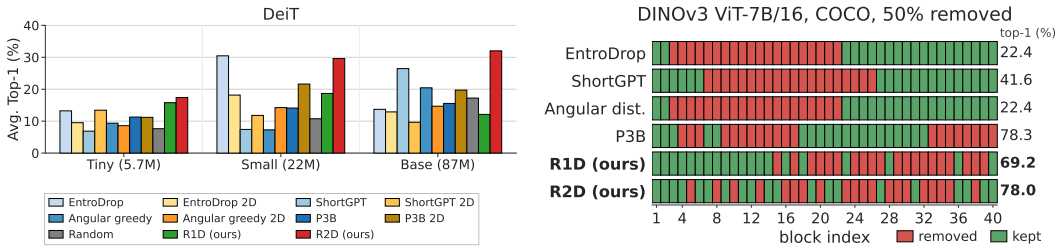
4.4 Brain-derived metrics remain competitive across model sizes and compression levels

In Table 2, we report accuracy averaged over COCO, ImageNet, CIFAR-10, CIFAR-100 and DTD after healing, for different models and sizes, each pruned at 25%, 50% and 75% of their initial number of blocks. Across these 27 settings, R1D or R2D outperform or perform on par with all five baselines in 85% of the cases. These wins are particularly significant at 50% compression, where our metrics are ahead in 7 out of 9 settings. The largest gap is on DINOv3 small at 50% compression, where R1D achieves 55.4% and the best baseline 43.5%.

DINOv3 7B illustrates how over-parameterization masks the significance of block selection. At 25% compression the 7 metrics fall within 1.2% of each other, so a model of that capacity has to be pruned aggressively before the block selection matters at all. Within DINOv3, R1D outperforms R2D on the small model at 25% and 50% compression and matches it at 75%, while R2D is ahead on the 7B model at all three levels. As discussed in Section 4.2, DINOv3 small is the one model whose plateau runs to the final block as in Figure 7.

5 Critically examining brain-inspired block-importance metrics

Is the brain signal necessary? R1D outperforms activation-based baselines in some settings, most clearly on DINOv3 small, and falls behind in others, while R2D is the more consistent of the two.



(a) Evaluating 2D variants of all baselines

(b) Blocks pruned by different methods

Figure 6: **Which blocks each metric removes, and what the second derivative buys.** Both panels are at 50% pruning and without healing. **(a)**The second-derivative (2D) variant of activation-based metrics does not consistely improve their accuracy on DeiT Tiny, Small and Base averaged over 5 datasets. **(b)** Pruning selection by different metrics for DINOv3-7B on COCO, red for a pruned block and green for a kept one, with top-1 accuracy for each drop set on the right. The metrics that take one contiguous collection of blocks lose the most accuracy.

This raises the question of whether the improvement comes from the second-derivative operator itself or from the brain signal being better extracted by that operator. The baselines can be seen as approximate first derivatives of other signals, such as entropy or angular distance, so we recompute each of them with the same second-derivative operator and refer to the result as its 2D variant. Figure 6a shows that no 2D variant outperforms R2D on DeiT at 50% pruning, and that the operator on its own is not consistently helpful, improving 5 of the 12 baseline-size pairs, degrading 6 and leaving 1 unchanged. In all of these settings, we conclude that the improvement comes from the brain data in combination with the operator rather than from the 2D operator alone.

Why is random pruning so competitive? Random masks are known to be strong baselines despite their simplicity (Su et al., 2020; Frankle et al., 2020; Li et al., 2022; Liu et al., 2022; Gadhikar et al., 2023), and we observe the same pattern in Section 4. We look closely into the pruned blocks for DINOv3 7B since random pruning performs most competitively for this model. Figure 6b shows that for such an overparameterized model, accuracy seems to correlate strongly with whether the block-importance metric selects contiguous blocks to be pruned or prunes the network sparsely. EntroDrop, Angular distance, and ShortGPT each delete a single contiguous run of 20 blocks and have a significant drop in accuracy. R1D breaks its selection into 6 runs and retains 69.2% accuracy, whereas R2D spreads its 20 deletions over 10 selections of at most 4 blocks and achieves 78.0% accuracy, performing on par with P3B. Random selection would in most cases lead to sparse pruning by construction, which is sufficient to place it among the strongest methods at this scale.

How robust are brain-inspired metrics across participants? Figure 10 shows the performance of the R-score from each participant separately as a function of the pruning strength. Up to 45% pruning, all 8 selections stay within 2% of one another, and the participants agree on which blocks to remove more than chance at every pruning strength. The 8 participants induce different pruning patterns, yet their accuracies stay close to one another, which tells us that each individual's data carries helpful signal on its own, and confirms the robustness of our metrics to subject selection.

6 Conclusion

Understanding why artificial networks develop representations that resemble those of the human brain remains a fundamental challenge for both neuroscience and AI. Here, rather than seeking to reverse-engineer the biological algorithms underlying these representations, we use neural data as proxy for algorithms allowing cortical computations, and treat them a direct signal of block importance in vision models. We show that, even without a mechanistic account of the underlying cortical computations, brain recordings provide a meaningful signal of block importance, enabling a biologically-grounded pruning method that reaches state-of-the-art performance in several settings. More broadly, our results suggest that biological representations can serve not only as objects of study, but also as a guiding signal for building more efficient artificial models.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, make plots for this paper, and to polish the writing. We have not used generative AI tools to develop the conceptual frameworks presented in this work, nor to propose or refine our hypotheses. The remaining categories requiring disclosure are not applicable to this work: we did not generate synthetic data sets, formulate mathematical claims, obtain critical ingredients for proving mathematical claims, write proofs, perform translation, or edit the paper to improve readability with generative AI tools.

We have reviewed all AI-assisted work. All AI-generated or AI-assisted code was read, tested, and verified for correctness by at least two authors, and all reported numbers were produced by code that the authors independently inspected and re-ran. Figures produced with AI assistance were checked against the data they show.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve the collection of new human data, and the authors did not interact with human subjects at any point. All brain recordings used here come from the publicly released Natural Scenes Dataset (Allen et al., 2022), which was collected under the ethical approvals and informed-consent procedures described by its authors, and which we use in accordance with its terms of release. We use the data in its released, de-identified form and we do not release any new neural data.

We also note that our use of these recordings is not aimed at decoding the content of brain activity. The brain-encoding scores R(k) are used only in aggregate, averaged across subjects and cortical regions, to obtain a per-block importance signal for a vision model. We do not reconstruct stimuli, infer individual mental states, or attempt to recover any subject-specific information, and the resulting metrics carry no information about individual participants. The purpose of the analysis is to use the neural signal as a reference for identifying redundant components of artificial networks and thereby to build more efficient models. All other data used in this work are standard public benchmarks.

REPRODUCIBILITY STATEMENT

Our aim is to make the results in this paper reproducible from public data and the details provided in the paper. All neural data come from the publicly available Natural Scenes Dataset (Allen et al., 2022), and the brain encoding procedure, including the ridge regression setup, the regularization grid, and the cross-validation and aggregation scheme, is described in Section 3. We also describe in Section 3 the two block-importance metrics, the pruning setup, the greedy pruning procedure, the models and vision tasks, and the LoRA healing process. All the block-importance metrics are summarized in Table 1. All additional experimental details to reproduce our work are described in Section A and additional ablations are provided in Section B.

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, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL http://dx.doi.org/10.1038/s41593-021-00962-x.

Aditi Anand, Sanchari Sen, and Kaushik Roy. Quantifying the brain predictivity of artificial neural networks with nonlinear response mapping. *Frontiers in Computational Neuroscience*, 15, August 2021. ISSN 1662-5188. doi: 10.3389/fncom.2021.609721. URL http://dx.doi.org/10.3389/fncom.2021.609721.

Vijay Balasubramanian. Brain power. *Proceedings of the National Academy of Sciences*, 118(32), August 2021. ISSN 1091-6490. doi: 10.1073/pnas.2107022118. URL http://dx.doi.org/10.1073/pnas.2107022118.

Francesco Cagnetta, Leonardo Petrini, Umberto M. Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *Physical Review X*, 14(3), July 2024. ISSN 2160-3308. doi: 10.1103/physrevx.14.031001. URL http://dx.doi.org/10.1103/PhysRevX.14.031001.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455–462, 2014.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild, 2013. URL https://arxiv.org/abs/1311.3618.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, pp. 2022–03, 2022.

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.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Pruning neural networks at initialization: Why are we missing the mark? *arXiv preprint arXiv:2009.08576*, 2020.

Advait Harshal Gadhikar, Sohom Mukherjee, and Rebekka Burkholz. Why random pruning is all we need to start sparse. In *International Conference on Machine Learning*, pp. 10542–10570. PMLR, 2023.

Patrick Glandorf and Bodo Rosenhahn. Pruning by block benefit: Exploring the properties of vision transformer blocks during domain adaptation. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3027–3037. IEEE, 2025.

Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, 7:267, 2013.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A Roberts. The unreasonable ineffectiveness of the deeper layers. In *International Conference on Learning Representations*, volume 2025, pp. 81906–81920, 2025.

U. Guclu and M. A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, July 2015. ISSN 1529-2401. doi: 10.1523/jneurosci.5023-14.2015. URL http://dx.doi.org/10.1523/JNEUROSCI.5023-14.2015.

Uri Hasson, Samuel A. Nastase, and Ariel Goldstein. Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434, February 2020. ISSN 0896-6273. doi: 10.1016/j.neuron.2019.12.002. URL http://dx.doi.org/10.1016/j.neuron.2019.12.002.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11):e1003915, November 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003915. URL http://dx.doi.org/10.1371/journal.pcbi.1003915.

J-R. King, C. Bel, L. Evanson, J. Gadonneix, S. Houhamdi, J. Lévy, J. Raugel, A. Santos Revilla, M. Zhang, J. Bonnaire, C. Caucheteux, A. Défossez, T. Desbordes, P. Diego-Simón, S. Khanna, J. Millet, P. Orhan, S. Panchavati, A. Ratouchniak, A. Thual, T. Brooks, K. Begany, Y. Benchetrit, M. Careil, H. Banville, S. d'Ascoli, S. Dahan, and J. Rapin. Neuralset: A high-performing python package for neuro-ai. 2026. URL https://kingjr.github.io/files/neuralset.pdf. Preprint; URL will be updated when the paper lands on arXiv.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yawei Li, Kamil Adamczewski, Wen Li, Shuhang Gu, Radu Timofte, and Luc Van Gool. Revisiting random channel pruning for neural network compression. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 191–201. IEEE, 2022.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

Shiwei Liu, Tianlong Chen, Xiaohan Chen, Li Shen, Decebal Constantin Mocanu, Zhangyang Wang, and Mykola Pechenizkiy. The unreasonable effectiveness of random pruning: Return of the most naive baseline for sparse training. *arXiv preprint arXiv:2202.02643*, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew G Wilson. Pac-bayes compression bounds so tight that they can explain generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

Sanae Lotfi, Yilun Kuang, Brandon Amos, Micah Goldblum, Marc Finzi, and Andrew G Wilson. Unlocking tokens as data points for generalization bounds on larger language models. *Advances in Neural Information Processing Systems*, 37:9229–9256, 2024.

Sanae Lotfi, Polina Kirichenko, Steven Li, and Zechun Liu. Quantized reasoning models think they need to think longer, but they do not. *arXiv preprint arXiv:2606.00206*, 2026.

Florian P. Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N. Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*, 7(6):848–859, June 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01041-7. URL http://dx.doi.org/10.1038/s42256-025-01041-7.

Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and Tim C. Kietzmann. Individual differences among deep neural network models. *Nature Communications*, 11(1), November 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19632-w. URL http://dx.doi.org/10.1038/s41467-020-19632-w.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20192–20204, 2025.

Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving semantic understanding in speech language models via brain-tuning, 2025. URL https://arxiv.org/abs/2410.09230.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

Subba Reddy Oota, Vijay Rowtula, Satya Sai Srinath Namburi, Khushbu Pahwa, Anant Khandelwal, Manish Gupta, Tanmoy Chakraborty, and Bapi S. Raju. Linguistic properties and model scale in brain encoding: from small to compressed language models, 2026. URL https://arxiv.org/abs/2602.07547.

Joséphine Raugel, Marc Szafraniec, Huy V. Vo, Camille Couprie, Patrick Labatut, Piotr Bojanowski, Valentin Wyart, and Jean-Rémi King. Disentangling the factors of convergence between brains and computer vision models, 2026. URL https://arxiv.org/abs/2508.18226.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

Katerina Marie Simkova, Adrien Doerig, Clayton Hickey, and Ian Charest. Representations in vision and language converge in a shared, multidimensional space of perceived similarities, 2025. URL https://arxiv.org/abs/2507.21871.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Jingtong Su, Yihang Chen, Tianle Cai, Tianhao Wu, Ruiqi Gao, Liwei Wang, and Jason D Lee. Sanity-checking pruning methods: Random tickets can win the jackpot. *Advances in neural information processing systems*, 33:20390–20401, 2020.

Benedikt Sundermann and Bettina Pfleiderer. Functional connectivity profile of the human inferior frontal junction: involvement in a cognitive control network. *BMC neuroscience*, 13(1):119, 2012.

Priya Tarigopula, Scott Laurence Fairhall, Anna Bavaresco, Nhut Truong, and Uri Hasson. Improved prediction of behavioral and neural similarity spaces using pruned dnns. *Neural Networks*, 168: 89–104, November 2023. ISSN 0893-6080. doi: 10.1016/j.neunet.2023.08.049. URL http://dx.doi.org/10.1016/j.neunet.2023.08.049.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Loek van Rossem and Andrew M. Saxe. When representations align: Universality in representation learning dynamics, 2024. URL https://arxiv.org/abs/2402.09142.

Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, 5(12):1415–1426, November 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00753-y. URL http://dx.doi.org/10.1038/s42256-023-00753-y.

Mingqing Xiao, Kai Du, and Zhouchen Lin. Beyond representational alignment with brain-guided language models for robust reasoning. *Nature Machine Intelligence*, 8(8):1275–1289, August 2026. ISSN 2522-5839. doi: 10.1038/s42256-026-01278-w. URL http://dx.doi.org/10.1038/s42256-026-01278-w.

Daniel L. K. 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, May 2014. ISSN 1091-6490. doi: 10.1073/pnas.1403112111. URL http://dx.doi.org/10.1073/pnas.1403112111.

Liangwei Yang, Yuhui Xu, Juntao Tan, Doyen Sahoo, Silvio Savarese, Caiming Xiong, Huan Wang, and Shelby Heinecke. Entropy-based block pruning for efficient large language models. *arXiv preprint arXiv:2504.03794*, 2025.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. *arXiv preprint arXiv:1804.05862*, 2018.

A Experimental Details

Throughout the appendix, we use x_k to denote the input to block k and x_{k+1} its output collected on the unpruned model. All activation baselines are computed on a calibration set disjoint from the evaluation data: for COCO, a class-balanced 5% of the NSD stimuli (3,528 images); for every other dataset, a class-balanced 5% of its training split.

A.1 Description of Baseline Metrics.

Block Influence (ShortGPT). Men et al. (2025) measures the importance of each block by how much it changes its input, averaging the cosine similarity over images and all tokens t; blocks are ranked once and those that act as near-identity maps are dropped first.

$$\mathrm{BI}(k) = 1 - \mathbb{E}\big[\cos(x_{k,t},\, x_{k+1,t})\big].$$

Angular Distance. Following Gromov et al. (2025), we first compute the angular distance between the input and output averaged over images for each candidate span of n contiguous blocks. To calculate this score, we use the `[class]` token (denoted as $x_{k,\mathrm{cls}}$) for DINOv3 and DeiT, and the spatially averaged token for Swin, whose candidate spans stay within one stage:

$$d(k,n) = \tfrac{1}{\pi}\,\mathbb{E}\Big[\arccos\big(\cos(x_{k,\mathrm{cls}},\, x_{k+n,\mathrm{cls}})\big)\Big], \qquad k^\star = \arg\min_k d(k,n),$$

and the block span with the lowest score $\{k^\star, \ldots, k^\star{+}n{-}1\}$ is removed.

Entropy Increase (EntroDrop). Yang et al. (2025) measure block importance by the change in entropy of the hidden states across the block. We average each image's hidden states over tokens, $\bar{x}_k$, and estimate the entropy over the calibration set. Blocks are ranked once by the signed increase and those with the smallest increase are dropped first; the first two blocks and the last block are never removed.

$$\Delta H(k) = H(\bar{x}_{k+1}) - H(\bar{x}_k).$$

Block Performance Indicator (P3B). Glandorf & Rosenhahn (2025) defines a block performance indicator (BPI) of a given block based on the increase in cross-entropy loss when that block is bypassed. Adapting their method, for each block k we fit a linear classifier h_k on the frozen representations x_k to predict the dataset's labels. Then $\mathrm{BPI}(k)$ is the difference in cross-entropy (CE) loss on the calibration set when h_k is evaluated on the previous block's features x_{k-1} instead of x_k:

$$\mathrm{BPI}(k) = \mathbb{E}\left[\mathrm{CE}\big(h_k(x_{k-1})\big) - \mathrm{CE}\big(h_k(x_k)\big)\right]. \tag{3}$$

Negative values are set to zero and the scores are normalized to sum to one over blocks. We compute this for two probes per block, on the `[class]` token and on the mean of the patch tokens (Swin: the spatially averaged token only), and average the two. Blocks are ranked once and the lowest-scoring ones are dropped first.

Random Selection. A chance-level control on ranking. Per seed we fix one random permutation of $3, \ldots, L-1$ and remove its first n blocks, so drop sets are nested across compression levels. Any useful importance score should beat it. We report the results on random selection as a mean over 10 seeds.

A.2 Details of R score calculation

To compute R scores, the model activation matrix X must have dimensions $n \times d$, where n is the number of images and d is the hidden state size. However, the natural output of a Vision Transformer includes an extra dimension for the token sequence. While X could theoretically incorporate every token, doing so makes the subsequent ridge regression computationally prohibitive. To enable less computation, Raugel et al. (2026) aggregates the tokens to eliminate this extra dimension.We adopt this practice to construct X by isolating the `[class]` token, aligning with the design of the original Vision Transformer (Dosovitskiy et al., 2020). The `[class]` token is a learned vector appended to the input sequence. Inherited from text-based models, it is engineered to aggregate global context and serve directly as the holistic image representation.

A.3 COMPUTING AND EVALUATING BLOCK-IMPORTANCE METRICS

Every metric follows the same four steps (Algorithm 1). Steps (1) and (4) are shared, while steps (2) and (3) differ across metrics.

(1) Linear probe. For each model and dataset, we train a linear probe on the features of the frozen, unpruned backbone using the training split $\mathcal{D}_t$. The probe consists of layer normalization followed by a linear classifier, trained with AdamW (learning rate 10^{-3}, weight decay 10^{-4}) for 20 epochs. This probe serves as the classification head to evaluate every metric.

(2) Computing the metric. Our R1D and R2D metrics are computed from per-block R-scores using only the NSD data (images seen by participants). We compute these R-scores once per model without using evaluation data (Section 3.1). In contrast, the activation baselines (ShortGPT, EntroDrop, P3B, and Angular distance) require evaluation data. They are computed from a single forward pass of the unpruned model on a class-balanced calibration set $\mathcal{C}$ (a 5% sample of $\mathcal{D}_t$ for ImageNet, CIFAR-10/100, and DTD, and NSD pool). Thus, the baselines use data from every evaluated dataset, while R1D and R2D solely rely on NSD data. Random selection uses no metric score.

(3) Selecting the blocks. We remove n out of K model blocks. R1D and R2D select blocks greedily: after each removal, derivatives are recomputed on the remaining blocks, and the lowest-scoring block is removed next (Sections 3.3 and 4.1). The baselines follow their original selection rules (Table 1). ShortGPT, EntroDrop, and P3B are static, removing the n lowest-scoring blocks based on a single, fixed ranking. Angular distance removes the n contiguous blocks with the smallest input-output angle, chosen independently for each n (within one stage for Swin). Random selection removes the first n blocks of a fixed random permutation (reported as the mean over 10 seeds). EntroDrop never removes blocks 1, 2, and K.

(4) Prune, heal and evaluate. We remove each selected block alongside its attention and MLP sublayers. Without healing, we evaluate the pruned model and frozen probe on the validation split $\mathcal{D}_v$. With healing, we train LoRA adapters ($r = 32$, $\alpha = 64$, dropout 0.05) on the attention projections alongside the probe. This training runs on $\mathcal{D}_t$ for 5 epochs using AdamW (learning rate 10^{-4}, weight decay 0.01). For both approaches, we use the epoch that achieved the highest accuracy on $\mathcal{D}_v$.

Algorithm 1 Evaluation protocol shared by all block-importance metrics.

Require: model M with L blocks, training and validation splits $\mathcal{D}_t$ and $\mathcal{D}_v$, metric s, number of blocks to remove n

(1) Linear probe

train a linear probe (the classification head) on the frozen, unpruned M with $\mathcal{D}_t$

(2) Compute the metric

if $s \in \{\mathrm{R1D}, \mathrm{R2D}\}$ **then**

  compute the R-scores $R_i(k)$ from the NSD data ▷ Section 3.1

else ▷ other baselines

  compute s as the baseline specifies, on a calibration set from the evaluation dataset ▷ Table 1

end if

(3) Select n blocks

if $s \in \{\mathrm{R1D}, \mathrm{R2D}\}$ **then**

  $P \leftarrow$ greedy removal, recomputing s after each step ▷ Section 4.1

else ▷ other baselines

  $P \leftarrow$ the n blocks chosen by the baseline's own selection rule ▷ Table 1

end if

(4) Prune, heal and evaluate

$M \leftarrow M$ without the blocks in P ▷ attention and MLP sublayers together

optionally heal M with LoRA on $\mathcal{D}_t$

report the top-1 accuracy of M with the probe on $\mathcal{D}_v$

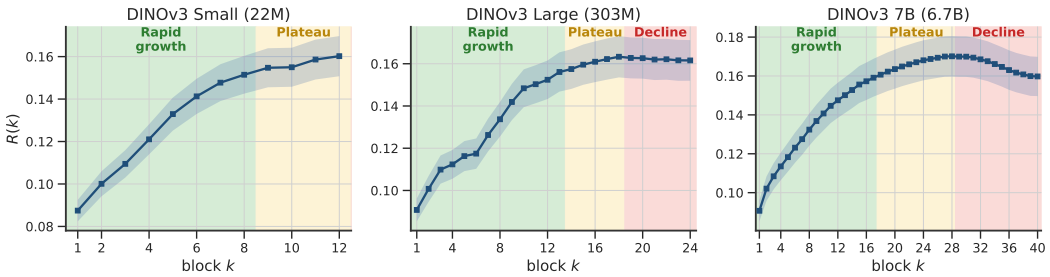


Figure 7: **Brain-encoding curves for DINOv3.** Panels and shading as in Figure 8. The plateau grows from 5 blocks at Small to 11 at 7B. Small is the one model in our sweep that peaks at its final block, so it has no decline phase.

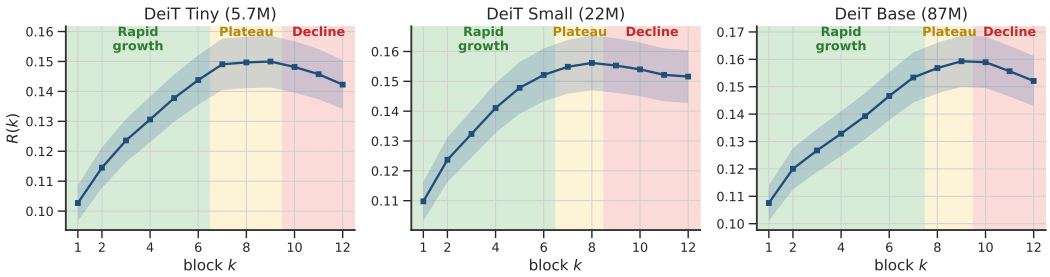


Figure 8: **Brain-encoding curves for DeiT.** Each panel is the mean over 8 participants and the 15 ROIs, with the band giving the standard error over participants. We recover the phases from each curve. Growth ends once R has completed 85% of its total rise, the plateau ends at the peak, and the decline continues to the last block. All three sizes have 12 blocks and plateaus of 2 to 3 blocks.

A.4 Angular distance on Swin at 75% pruning.

Angular distance (Gromov et al., 2025) removes the single contiguous run of n blocks whose input and output representations are closest in angle. In Swin, consecutive stages are separated by patch-merging layers that halve the spatial resolution and double the channel width, so representations on either side of a stage boundary cannot be compared and a run cannot cross one. We therefore restrict each run to a single stage and require it to leave at least one block of that stage, otherwise the method would have a single forced candidate and no longer select anything. Swin-S/B/L have 24 blocks in stages of 2, 2, 18 and 2, so the longest admissible run has 17 blocks since we require at least one block per stage to survive. At 75% pruning, $n = 18$, and no valid run exists. We report these cells as unavailable (–) rather than substitute a different selection rule. The other methods score blocks individually and may remove any 18 blocks across stages, so they are unaffected.

B Additional Results

B.1 Alignment curves for different models

Figure 3 reports three phases on DINOv3-ViT-7B. The same three phases hold for DeiT and Swin across sizes and for DINOv3 large as Figures 7 to 9 show. Alignment rises quickly over the first blocks, then barely moves for a stretch in the middle, then falls slightly over the last few. Shallower models have shorter plateaus and shorter declines. DINOv3 Small is the one exception to the three-phase shape, having its plateau phase extend until the end of the network. In all of these cases, we can make an argument that there is a phase transition at least once in these models and therefore blocks at the boundaries of these transitions should be preserved.

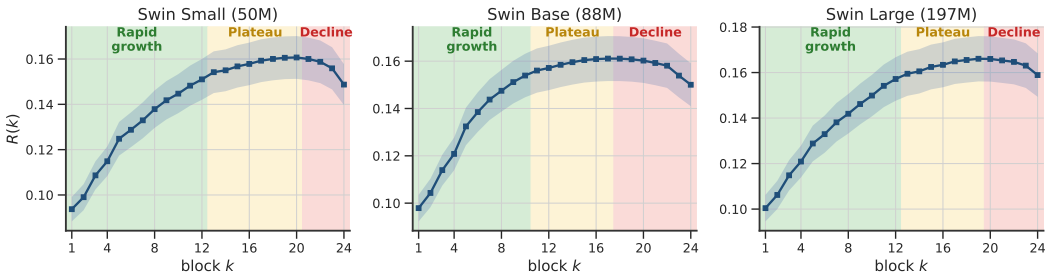


Figure 9: **Brain-encoding curves for Swin.** Panels and shading as in Figure 8. Tiny has 12 blocks and a plateau of 3 blocks, and Small, Base and Large have 24 blocks and plateaus of 7 to 8.

B.2 Ablation on participants

In the main results, R2D is computed from R-scores averaged over the eight NSD participants. Here we compute R2D from each participant's R-scores alone, with the same 15 ROIs and the same greedy removal, which gives one pruning order per participant. We prune pretrained DINOv3-7B with each order from 5% to 95% of its 40 blocks and report COCO top-1 accuracy without healing (Figure 10, left). To measure how much participants agree on which blocks to remove, we compute the intersection over union (IoU) of the pruned sets between pairs of participants and between each participant and the averaged R-scores, and compare it with the IoU of two random sets of the same size (right).

Up to 45% pruning, all selections stay within 2 points of one another. From 50% to 70%, the averaged R-scores give 1 to 3 points lower accuracy than the mean single participant. The participants agree on the pruned blocks more than chance at every level (IoU 0.45 against 0.33 at 50%).

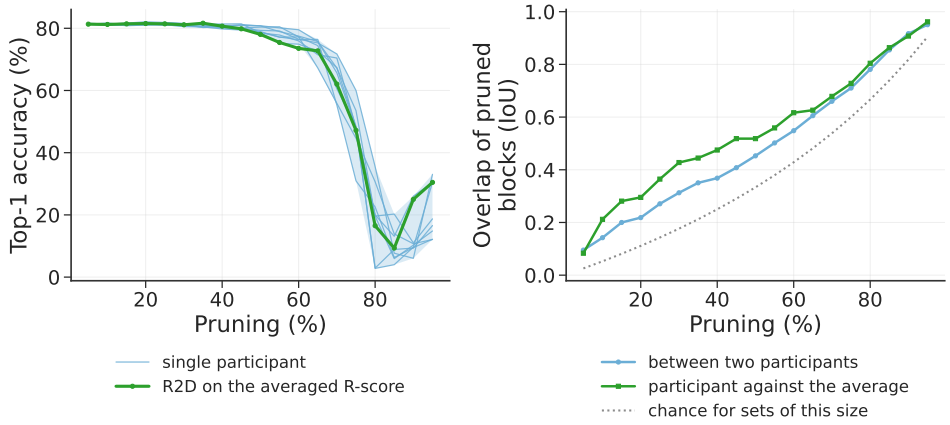


Figure 10: **R2D from each participant separately** (DINOv3-7B, COCO, no healing). **Left:** top-1 accuracy. **Right:** overlap (IoU) of the pruned blocks.