Scaling Laws for Task-Optimized Models of the Primate Visual Ventral Stream

Abdulkadir Gokce [1] **Martin Schrimpf** [1]

Abstract

When trained on large-scale object classification datasets, certain artificial neural network models begin to approximate core object recognition behaviors and neural response patterns in the primate brain. While recent machine learning advances suggest that scaling compute, model size, and dataset size improves task performance, the impact of scaling on brain alignment remains unclear. In this study, we explore scaling laws for modeling the primate visual ventral stream by systematically evaluating over 600 models trained under controlled conditions on benchmarks spanning V1, V2, V4, IT and behavior. We find that while behavioral alignment continues to scale with larger models, neural alignment saturates. This observation remains true across model architectures and training datasets, even though models with stronger inductive biases and datasets with higher-quality images are more compute-efficient. Increased scaling is especially beneficial for higher-level visual areas, where small models trained on few samples exhibit only poor alignment. Our results suggest that while scaling current architectures and datasets might suffice for alignment with human core object recognition behavior, it will not yield improved models of the brain's visual ventral stream, highlighting the need for novel strategies in building brain models.

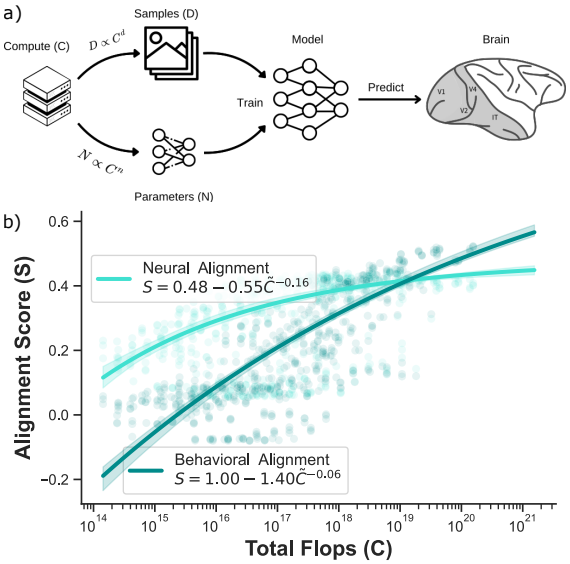

Figure 1: **a)** For a given compute budget (C), we determine the scaling laws for maximal neural and behavioral alignment to the primate visual ventral stream. **b)** We find consistent scaling laws for brain and behavioral alignment across over 600 models. While we predict models to approach perfect behavioral alignment at large scales, the effect of scaling on brain alignment is already saturating.

The advent of neural networks has revolutionized our understanding and modeling of complex neural processes. A particularly active area of study is the ventral visual stream in primates, a key pathway in the brain responsible for processing visual information (Goodale & Milner, 1992; Grill-Spector et al., 2001; Malach et al., 2002; Kriegeskorte et al., 2008). Neural networks, when trained on extensive datasets, have emerged as the most accurate quantitative tools for simulating the response patterns of neurons within this stream (Yamins et al., 2014; Schrimpf et al., 2018). These advanced models offer a precise computational account of how neural mechanisms in the brain give rise to visual perception.

Recent developments in machine learning have emphasized the significance of both the volume of training data and the complexity of model architectures (Kaplan et al., 2020; Hoffmann et al., 2022; Zhai et al., 2022; Bahri et al., 2022; Antonello et al., 2023; Muennighoff et al., 2023; Aghajanyan et al., 2023; Isik et al., 2024). These findings raise the question: Can we build better models of the brain by scaling up model architectures and dataset sizes? Recent studies have found that in pre-trained models, the number of parameters and dataset samples respectively seem to improve predictions of fMRI and behavioral measurements (Antonello et al., 2023; Muttenthaler et al., 2023). With the

[1]EPFL. Correspondence to: Abdulkadir Gokce <abdulkadir.gokce@epfl.ch>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

numerous differences between pre-trained models however, the relative contributions of model parameters and dataset size to brain and behavioral alignment are not clear.

Despite recent successes in using neural networks as models of the brain, a comprehensive understanding of how model scale—separately and jointly across parameters, dataset size, and compute—affects functional alignment with different cortical areas remains elusive. Previous studies have often relied on heterogeneously trained models using off-the-shelf checkpoints (Conwell et al., 2024), or focused narrowly on specific brain areas (e.g., IT only (Linsley et al., 2023)), frequently using proxy quantities such as task performance. Our work addresses these limitations through a systematic, from-scratch training protocol spanning over 600 models, enabling controlled comparisons and robust parametric estimation of scaling laws for both behavioral and neural alignment across the entire ventral visual hierarchy. This approach offers a clearer disentanglement of the respective contributions of architecture, data, and optimization objective to brain modeling.

In this paper, we examine how scaling – of model parameters and training dataset size – impacts the alignment of artificial neural networks with the primate ventral visual stream. We systematically train models from a variety of architectural families on image classification datasets which allows us to independently control and observe the effects of model complexity and data volume. To capture the observed trends, we introduce parametric power-law trends that describe the impact of scale on alignment with behavior and brain regions along the visual ventral stream. We summarize the contributions of this work as follows:

- While scale initially improves alignment, brain alignment saturates. Behavioral alignment on the other hand continues to improve.
- Increasing both parameter count and training dataset size improves alignment, with data providing more gains over model scaling.
- Architectures with stronger inductive bias (e.g., convolutions and recurrence) and datasets with higher-quality images are more sample- and compute-efficient.
- Fitting parametric power-law curves, we find that model alignment with higher-level brain regions and especially behavior benefits the most from scaling.
- We publicly release our training code, evaluation pipeline, and over 600 checkpoints for models trained in a controlled manner to enable future research.

1. Related Work

Primate Visual Ventral Stream. The ventral visual stream, a critical pathway in the primate brain, including humans, plays a key role in visual perception, extending from the occipital to the temporal lobes and serving as the "what pathway" for object recognition and form representation (Goodale & Milner, 1992; Grill-Spector et al., 2001; Malach et al., 2002; Kriegeskorte et al., 2008). Beginning in the primary visual cortex (V1), where basic visual information from retinal ganglion cells is processed, the ventral stream proceeds through areas such as V2, V3, V4, and the inferotemporal cortex (IT), each responsible for increasingly complex features of visual perception (Kandel et al., 2000). Despite decades of research and a wealth of brain data, the precise neural mechanisms underlying visual perception are not well understood.

Modeling the Primate Visual Ventral Stream. Particular artificial neural networks (ANNs) are the most accurate models of brain responses in the visual ventral stream and associated core object recognition behaviors (Schrimpf et al., 2018; 2020). Models optimized for ecologically viable tasks (Yamins & DiCarlo, 2016) in particular have demonstrated strong brain and behavioral alignment (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Cadena et al., 2019; Schrimpf et al., 2018; Nayebi et al., 2018; Kietzmann et al., 2019; Rajalingham et al., 2018; Zhuang et al., 2021; Geiger et al., 2022) – notably these models are trained purely on image classification datasets, without fitting to brain data.

Scaling Laws. Recent advancements in artificial intelligence are driven by scaling the model size and training data. Empirical evidence suggests a power-law relationship between model performance and both model parameters and dataset size, indicating that continued scaling will further improve performance (Kaplan et al., 2020; Cherti et al., 2023; Zhai et al., 2022; Hoffmann et al., 2022; Dehghani et al., 2023; Henighan et al., 2020; Brown et al., 2020; Bahri et al., 2022; Hestness et al., 2017). The power-law exponents enable the optimal allocation of compute between model parameters and dataset samples, such that performance is maximized (Kaplan et al., 2020; Hoffmann et al., 2022).

While scaling laws for machine learning *performance* has been extensively studied, the scaling laws for *brain alignment* remain unclear. Recent studies suggest an involvement of both model size and data volume in the functional alignment with brain data (Azabou et al., 2023; Benchetrit et al., 2023; Caro et al., 2024; Antonello et al., 2023). Conversely, Muttenthaler et al. (2023) indicate that sample size is critical for behavioral alignment. We here unify these results, in the realm of the primate visual ventral stream, into quantitative scaling laws for how model and dataset sizes relate to alignment with the brain and behavior.

2. Methods

Neural & Behavioral Alignment. To evaluate the alignment of our model with brain function, we utilize a range of benchmarks from Brain-Score (Schrimpf et al., 2018; 2020). These benchmarks assess model performance by comparing model activations or behavior with primate neural data using the same images. Specifically, the V1 and V2 benchmarks compare model outputs to primate single-unit recordings from (Freeman et al., 2013), using 315 texture images and data from 102 V1 and 103 V2 neurons. For the V4 and IT benchmarks, 2,560 images are used to match model activations to primate Utah array recordings from (Majaj et al., 2015), based on data from 88 V4 and 168 IT electrodes. A linear regression is trained on 90% of the images to correlate model and neural data, with prediction accuracy for the remaining 10% evaluated using Pearson correlation, repeated ten times for cross-validation. The behavioral benchmark assesses model predictions for 240 images against primate behavioral data from (Rajalingham et al., 2018) using a logistic classifier trained on 2,160 labeled images. Pearson correlation is used to measure the similarity in confusion patterns between model predictions and primate responses. All benchmark scores are normalized to their respective maximum possible values.

We define the model's alignment score S (and an inverse *Misalignment Score* $L = 1 - S$) as the average across the V1, V2, V4, IT, and behavioral benchmark scores. Layers are committed to brain regions based on models trained on a full dataset, and applied to all variants trained with subsampled datasets. As we reused the same neural and behavioral data both to select the optimal model layer for readout and to assess the model's alignment, we validated the benchmark results on a private split of each dataset on Brain-Score. We observed an almost perfect correlation between the results on the private and public splits (Appendix C).

Scaling Models and Data. We trained an array of standard models from several architecture families. Specifically, we used ResNet$18, 34, 50, 101, 152$ from (He et al., 2016); EfficientNet-B$0, 1, 2$ from (Tan & Le, 2019); Vision Transformer ViTT, S, B, L from (Dosovitskiy et al., 2021); ConvNeXtT, S, B, L from (Liu et al., 2022b); CORnet-S from (Kubilius et al., 2019); and AlexNet from (Krizhevsky et al., 2012). We also trained 33 modified versions of ResNet18: 22 models obtained by scaling the network width from $1/16$ to 4 times the original size, and 11 models derived by adjusting the depth. Similarly, we trained four additional ConvNeXt and ViT models by scaling the width of the ConvNeXt-T and ViT-S architectures.

For our experiments, we selected two image classification datasets: ImageNet (Deng et al., 2009) and EcoSet (Mehrer et al., 2021). ImageNet, with millions of labeled images across 1,000 categories, has long been a benchmark in computer vision, designed to challenge and evaluate automated visual object recognition systems. On the other hand, EcoSet is a more recent dataset, designed to provide an ecologically valid representation of human-relevant objects. It contains over 1.5 million images spanning 565 basic-level categories, curated to better reflect the natural distribution of objects in the real world, aligning with human perceptual and cognitive experiences.

To create subsets of ImageNet and EcoSet, we sampled $d \in 1, 3, 10, 30, 100, 300$ images per category. For $d \in 1, 10, 100$, we repeated the runs with three random seeds to ensure robustness. For ConvNeXts (Liu et al., 2022b) and ViTs (Touvron et al., 2022), we used the training recipes developed by the original model authors. The remaining models were trained for 100 epochs using a minibatch size of 512. We employed a stochastic gradient descent (SGD) optimizer with a cosine decaying learning rate schedule, starting with a peak learning rate of 0.1 and incorporating a linear warm-up phase spanning five epochs. We maintained the momentum at 0.9 and applied a weight decay of 10^{-4}. Cross-entropy loss was used as the minimization objective. We utilized standard ImageNet data augmentations, specifically random resized cropping and horizontal flipping.

Scaling Power-Law Curves. Following previous work on scaling laws (Zhai et al., 2022; Hoffmann et al., 2022; Besiroglu et al., 2024), we fit power law functions in the form

$$L = E + AX^{-\alpha} \tag{1}$$

on the data where L is the misalignment score, and X is an independent variable, such as the number of samples seen (D), number of parameters (N), and the total training floating point operations (FLOPs) (C). Coefficients E, A, and α are found by minimizing

$$\min_{a,e,\alpha} \sum_{i \in [\#\text{Runs}]} \text{Huber}_\delta \left(\text{LSE}(a - \alpha \log X_i, e) - \log L_i \right) \tag{2}$$

where $E = \exp(e)$, $A = \exp(a)$ and LSE is the log-sum-exp operator. We solve Eq. 1 using BFGS minimizer with $\delta = 1e-3$, and use a grid of initialiations as follows: $e \in \{-1, -0.5, \ldots, 1\}$, $a \in \{0, 5, \ldots, 25\}$, $\alpha \in \{0, 0.5, \ldots, 2\}$.

To capture the slow initial increase in benchmark scores of modern architectures like ConvNeXt and ViT models in the low-data regime, we introduce an additional parameter λ to Eq. 2. This parameter allows the fitted curve to saturate at lower scales, better reflecting the observed performance of

these models under limited data conditions:

$$L = E + A\left(X + 10^{\lambda}\right)^{-\alpha} \tag{3}$$

We minimize the modified equation as before, using $\lambda \in 0, 0.5, 1.0, 1.5, 2.0$. To fit the curve described by Eq. 3, we utilize all data points from the ConvNeXt and ViT models. For fitting the remaining curves, we select ConvNeXt and ViT runs that were trained on datasets with either 300 samples per class or the full dataset. This approach ensures that the fitted curves accurately represent the scaling behavior of these architectures across different data regimes.

Furthermore, we would like to describe the misalignment (L) as a function of both the model and data size (N, D) and predict optimal allocations N^* and D^* by solving

$$(N^*, D^*) = \underset{N,D}{\arg\min}\, L(N, D),\ \text{FLOPs}(ND) = C \tag{4}$$

In that regard, following (Hoffmann et al., 2022; Besiroglu et al., 2024) we fit a parametric function of the form

$$\hat{L}(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} \tag{5}$$

where the loss ($\hat{L}$) is a function of parameter count (N) and number of samples seen (D). In Eq. 5, the first term represents the loss in an ideal data generation scenario (entropy), the second and the third terms reflect the under-performance of a model due to limitations in parameter and data size (Hoffmann et al., 2022; Muennighoff et al., 2023). Following the example of Hoffmann et al. (2022), we learn variables $\{E,\ A,\ \alpha,\ B,\ \beta\}$ that characterizes misalignment by solving

$$\underset{e,\ a,\ \alpha,\ b,\ \beta}{\arg\min} \sum_{i \in [\#\text{Runs}]} \text{Huber}_{\delta}\Big(\log L_i - \text{LSE}\big(a - \alpha \log N_i, b - \beta \log D_i, e\big)\Big) \tag{6}$$

with $\delta = 10^{-3}$ and $E = \exp(e)$, $A = \exp(a)$ $B = \exp(b)$. Initialiations of b and β follow a and α, respectively.

Both Kaplan et al. (2020); Hoffmann et al. (2022) assume that compute follows the relationship $C(N, D) \approx 6ND$ to predict the optimal allocation of compute (C) to N and D using a set of equations with the learned variables mentioned above:

$$N^*(C) = G(C/6)^{a}, \quad D^*(C) = G^{-1}(C/6)^{b}$$
$$\text{where } a' = \frac{\beta}{\alpha+\beta},\ b' = \frac{\alpha}{\alpha+\beta},\ G = \left(\frac{\alpha A}{\beta B}\right)^{\frac{1}{\alpha+\beta}} \tag{7}$$

However, we observe that $C(N, D) \approx 6ND$ does not hold with different architectures, and various CNN families have a slightly different relationship of C, N, and D. As such, we assume a power-law relationship of the form

$$C(N,\ D) = m(ND)^{n} \tag{8}$$

where we fit m and n via linear regression of C and ND in log-log scale. Then, the updated equations governing the optimal allocation becomes

$$N^*(C) = G(C/m)^{a'/n},\ D^*(C) = G^{-1}(C/m)^{b'/n} \tag{9}$$

where a', b', and G are calculated as before.

To evaluate the uncertainty of our model fits, we performed bootstrapping with 1,000 resamples. We compute 95% confidence intervals for each point along the fitted curves based on the variability observed across the bootstrapped estimates.

Finally, to avoid large constants during curve fitting, we rescale the variables C, N, and D by setting $\tilde{C} = C/10^{13}$, $\tilde{N} = N/10^{5}$, and $\tilde{D} = D/10^{4}$.

3. Results

3.1. Scaling drives behavioral alignment, but saturates for neural alignment

Our experiments show a clear and consistent improvement in behavioral alignment as both model size and training dataset size increase. Fig 1.b illustrates this trend across different architectures and scaling axes. The curve $S = 1 - 1.4\tilde{C}^{-0.06}$ converges to perfect alignment score of 1 in the limit of C.

In contrast to behavioral alignment, neural alignment with specific brain regions demonstrated saturation as training compute scaled up in size. The curve represented by the formula $S = 0.48 - 0.55\tilde{C}^{-0.16}$ represents a saturation at 0.48. The diminishing returns in neural alignment imply that merely scaling up models and data is insufficient to achieve better alignment with higher-level neural representations.

3.2. Architectural Inductive Bias Influences Alignment and Scaling Dynamics

Experimental results indicate that modern architectures, such as ConvNeXt and Vision Transformers (ViTs), exhibit poorer neural alignment compared to models like ResNets and EfficientNets in low data regime. ResNets and EfficientNets, which have stronger inductive biases due to their fully convolutional structures, demonstrate high neural alignment even at initialization. In Fig. 2, alignment score of ResNets and EfficientNets increase steadily with additional compute in the form of training samples, however ConvNeXt and ViT requires more compute in order to start rising.

This difference in initial alignment also affects how the scaling laws evolve for each architecture. Models with weaker inductive biases require more extensive scaling—specifcally in terms of training data—to achieve levels of neural alignment comparable to those with stronger inductive biases. Consequently, the scaling curves for ConvNeXt and ViT

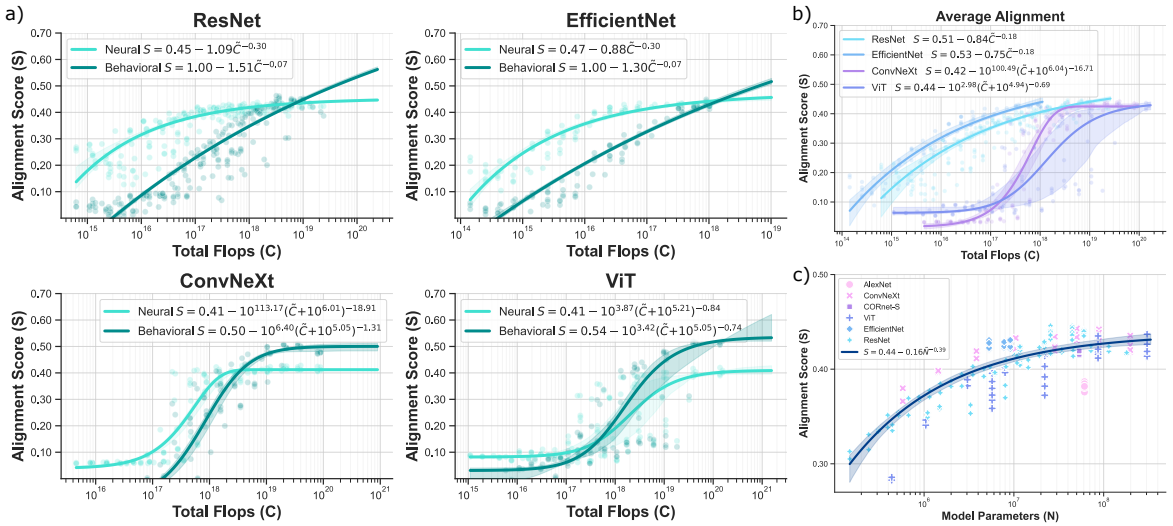

Figure 2: **Scaling Model Size. a)** Neural and behavioral alignments of four architecture families. Models with inductive biases (ResNet, EfficientNet) are more compute-efficient than less constrained models (ConvNeXt, ViT). **b)** Average alignment per model architecture. All models converge to similar alignments. **c)** Increasing parameters improves alignment (models trained on full datasets), but the effects saturate.

models develop differently, highlighting that architectural choices not only impact baseline alignment but also influence the efficiency of scaling strategies.

Fig. 3b highlights that architectural priors critically shape alignment dynamics, particularly in low-data settings. CORnet models, which incorporate recurrence, achieve relatively high alignment early in training—outperforming both convolutional and transformer-based models under limited supervision. Yet as training data increases, this initial advantage wanes, and alignment scores across architectures begin to converge. This suggests that while certain inductive biases offer sample efficiency, their long-term benefits may be outpaced by deeper or more flexible architectures given sufficient data. Overall, these findings emphasize that strong inductive biases—such as convolution and recurrence—facilitate better alignment when data is limited, whereas extensive task-driven optimization on larger datasets eventually mitigates differences across architectures.

3.3. More Data Is Better Than More Parameters

Our analysis reveals that increasing the size of the training dataset has a more significant impact on improving brain alignment than simply enlarging the number of model parameters. While both strategies lead to performance enhancements, the benefits from data scaling exhibit less severe diminishing returns compared to model scaling. Specifically, models trained on larger datasets consistently demonstrate superior neural and behavioral alignment with the primate ventral visual stream, following a predictable power-law relationship.

In contrast, expanding the model size without proportionally increasing the training data results in steeper diminishing returns in alignment performance. Larger models rapidly reach a point where additional parameters do not translate into meaningful improvements. Fig. 2c estimates a saturation level of 0.44 by scaling model sizes with all samples of training data whereas Fig. 3a predicts maximum alignment of 0.46 and 0.48 for ImageNet and Ecoset respectively. This indicates that scaling training datasets overall improves brain alignment better than models scaling. Furthermore, Fig. 4b demonstrates that larger models of the same architecture family require much more samples to achieve the same level of alignment.

To quantitatively capture the joint interaction between data and model scaling, we fitted a parametric curve based on Eq.5, as shown in Fig.4a. This curve effectively models how compute (C), dataset size (D), and model size (N) collectively influence brain alignment. Utilizing the parametric relationships described in Eq. 9, we estimate that additional compute should be allocated following the scaling laws $D \approx C^{0.7}$ and $N \approx C^{0.3}$. These exponents indicate that, for optimal brain alignment, computational resources should be predominantly invested in increasing the dataset size rather than the model size.

3.4. Ordered effect of scale on alignment

Our study reveals a graded effect of scaling on alignment across the cortical hierarchy of the primate visual system.

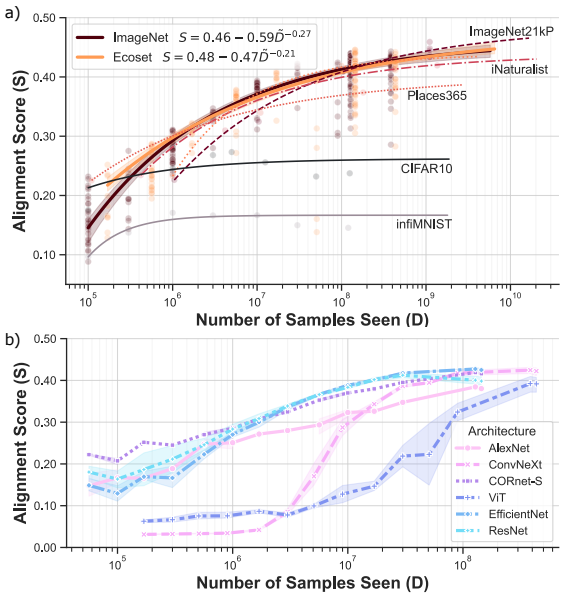

Figure 3: **Scaling Dataset Size. a)** Training on larger datasets enhances brain alignment. The alignment scaling curves derived from ImageNet and EcoSet closely estimate the alignment achieved when using ImageNet21k. In contrast, datasets with specialized image distributions—such as Places365—fall below the alignment scaling laws established by these generalist datasets. In the extreme case of handwritten digits (infiMNIST), the impact of training is minuscule on the alignment. **b)** Model families with weaker inductive bias start at a lower alignment and require more data to improve.

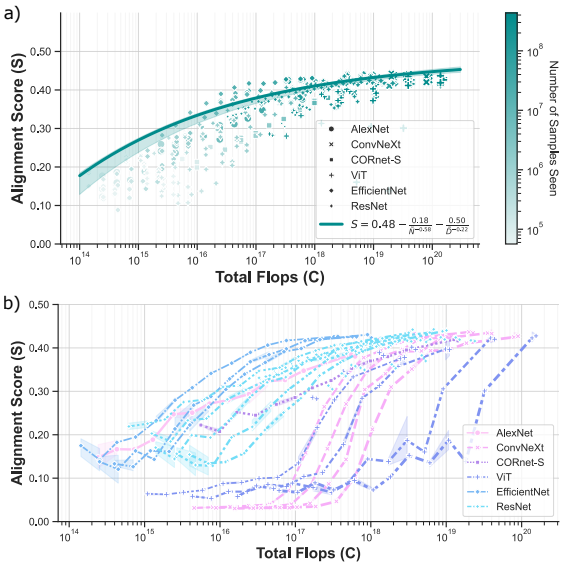

Figure 4: **Optimal Compute Allocation. a)** Alignment as a function of both model and training dataset sizes. Marker size is log-proportional to model size. Compute should be spent 0.3/0.7 on model/dataset size respectively. **b)** Models start out at different alignments but converge to the same saturating point.

Specifically, we observe that the benefits of increased training compute—achieved through larger datasets and more complex models—vary systematically among different brain regions, reflecting their position in the visual processing pathway. Fig. 5.a illustrates the alignment as a function of training compute across various brain regions. We categorized the models into two groups based on their architectural inductive biases. Group 1 includes most models with strong inductive biases, such as ResNets and EfficientNets. These models start with higher neural alignment scores even at initialization due to their fully convolutional architectures. Group 2 consists of models with weaker inductive biases, specifically ConvNeXt and Vision Transformers (ViTs). These models exhibit lower neural alignment in the low-data regime and require more compute to achieve similar alignment levels.

To quantify the impact of scaling on each brain region, we define the alignment gain per region as $A10^{\alpha}$ where A and α are parameters of Eq. 2. Our findings indicate that higher regions in the cortical hierarchy show greater benefits from increased compute. Fig. 5b illustrates the alignment gain per region, highlighting how higher cortical areas benefit more from scaling efforts. This ordered effect suggests that regions higher up in the visual hierarchy, such as the Inferior Temporal (IT) cortex and behavioral outputs, gain more substantially from additional data and increased model complexity. In contrast, early visual areas like V1 and V2 exhibit smaller alignment gains with increased compute, indicating a potential saturation effect.

4. Discussion

We establish scaling laws governing the effect of model and dataset scale on behavioral and brain alignment with the primate visual ventral stream. While scale is a necessary component for all brain-like models, model architectures with priors such as convolutions, and datasets with high-quality images are more sample efficient, leading to alignment with smaller compute requirements. Scale especially improves alignment with higher-level visual regions, but brain alignment saturates across all conditions tested here whereas behavioral alignment continuously improves with increased scale.

We find a saturation of neural alignment under current modeling approaches, consistent with trends reported in prior work (Linsley et al., 2023; Conwell et al., 2024; Muttenthaler et al., 2023). Critically, our results reveal a disconnect between neural and behavioral alignment: while behavioral

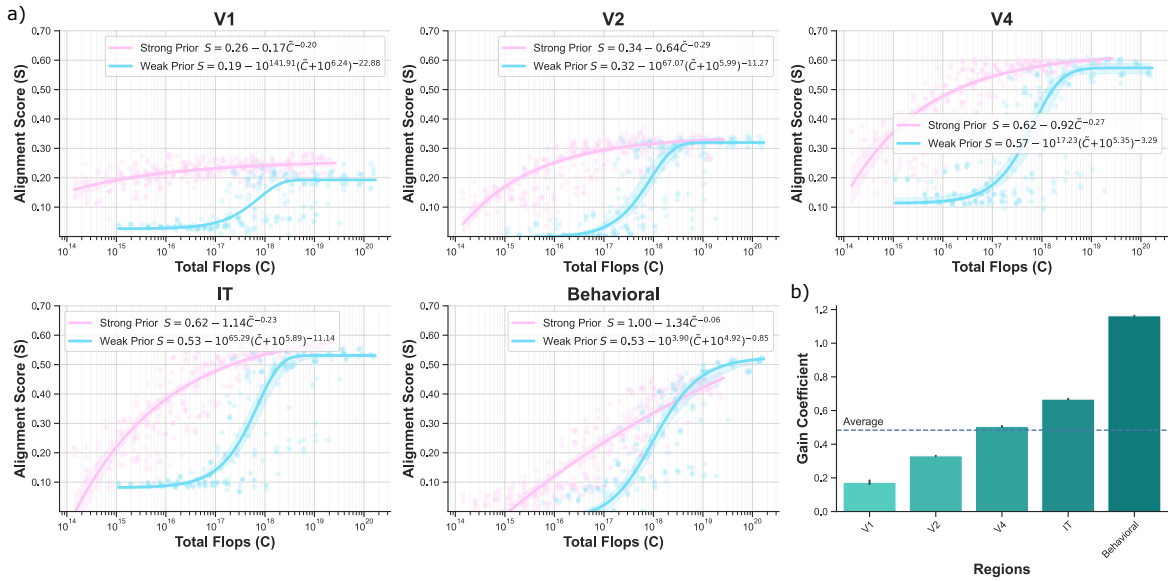

Figure 5: **Graded Effect of Scale across Cortical Hierarchy. a)** Alignment as a function of training compute across different brain regions. Group 1 contains most models except those with low inductive bias (Group 2; ConvNeXt, ViT). **b)** Alignment gain per region, defined as $A10^{\alpha}$. Regions higher in the cortical hierarchy show greater benefits from increased compute (Behavior > IT > V4 > V2 > V1).

.

alignment continues to improve with increased scale, neural alignment plateaus. By quantifying scaling laws across model families and data regimes, we show that improvements in brain alignment are more efficiently achieved by increasing dataset size rather than model parameters. These findings offer concrete guidance for developing brain-like models more effectively, emphasizing the importance of dataset diversity and biologically inspired architectural priors over brute-force model scaling.

Dissociation of behavioral and neural alignment. Our findings reveal a dissociation between behavioral and neural alignment as models are scaled with more parameters and larger datasets. While behavioral alignment continues to improve consistently with increased model parameters and training data – exhibiting a strong power-law relationship – neural alignment reaches a saturation point beyond which additional scaling yields minimal gains. This divergence suggests that behavioral alignment benefits more substantially from scaling efforts, whereas neural alignment may require alternative approaches beyond merely increasing model size and data volume to achieve further improvements.

This disparity is further highlighted by the correlation between task performance and alignment depicted in Figure 6. Behavioral alignment closely tracks validation accuracy, improving hand-in-hand as models become more accurate. Consistent with prior work(Schrimpf et al., 2018; Linsley et al., 2023), neural alignment eventually saturates, indicating that factors other than task performance influence neural alignment.

Generalization Beyond Supervised Training. We assessed whether alternative training paradigms can overcome the limitations observed in neural alignment under supervised learning. Figure 7a illustrates the scaling of alignment as a function of compute spent during self-supervised training of ResNet models using SimCLR (Chen et al., 2020) on ImageNet. The results confirm the trends observed in supervised training: behavioral alignment continues to improve with increased compute, following a strong power-law relationship, while neural alignment approaches a saturation point. This consistency suggests that the saturation in neural alignment is not exclusive to supervised learning but may be inherent to the models or datasets employed.

The region-specific breakdown (as illustrated in Supp. Fig. S6) further reinforces this observation. Even in a self-supervised learning context, higher-level visual areas like IT and behavioral outputs demonstrate more pronounced improvements with increased compute, while early visual areas like V1 and V2 show minimal gains. This suggests that the hierarchical nature of neural alignment is a fundamental characteristic that transcends specific training methods.

Additionally, we explored the impact of adversarial fine-tuning on alignment performance. In Figure 7b, ResNet models trained on subsets of ImageNet were fine-tuned

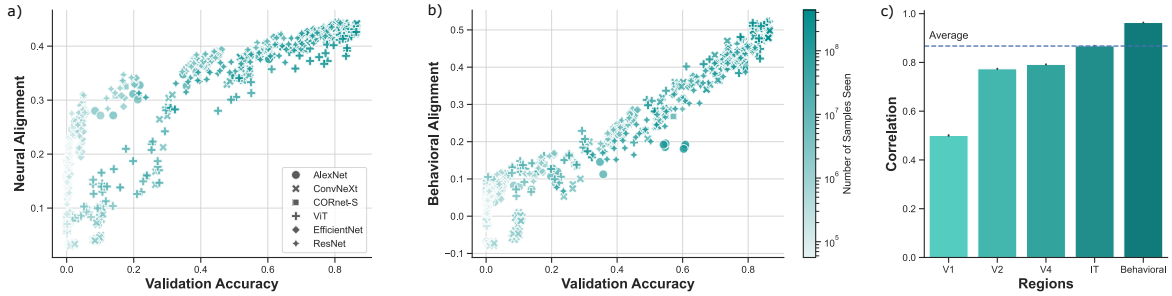

Figure 6: **Correlation between Task Performance and Alignment. a,b)** Correlation between validation accuracy (ImageNet & EcoSet) and brain (a) and behavioral (b) alignment. Behavioral alignment strongly correlates with task performance, whereas neural alignment shows a non-linear trend, reaching saturation. **c)** Pearson's correlation coefficient per region, with all p-values less than 10^{-40}.

adversarially for 10 epochs using the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015; Wong et al., 2020). Importantly, the scaling curves were estimated solely from the non-adversarial runs, yet the adversarially fine-tuned models exhibited improvements along these existing scaling curves. This indicates that adversarial training can enhance alignment without deviating from the established scaling behavior.

Impact of Architectural Inductive Biases on Alignment Dynamics. Our evaluation of alignment during training reveals that the alignment behavior varies significantly across different model architectures. Figure 7.c shows that while various models eventually converge to similar alignment levels with sufficient training, fully convolutional architectures—such as ResNets and EfficientNets—exhibit substantially higher alignment scores at the very beginning of training. This early advantage suggests that these architectures possess inherent features that align closely with neural data from the primate ventral visual stream even before learning from data occurs.

Further analysis in Figure S3 confirms that this initial high alignment is due to the strong inductive biases present in fully convolutional networks. These biases enable the models to start with representations already well-suited for neural alignment. Figure S4 reinforces this finding by demonstrating that models with strong inductive biases achieve higher initial alignment compared to architectures like ConvNeXt and ViT, which have weaker inductive biases.

Influence of Learning Signals on Alignment Dynamics. Our investigation reveals that the type of learning signal plays a crucial role in the dynamics of alignment during training. Figure 7d illustrates the alignment trajectories of ResNet50 and ViT-S models trained on ImageNet using supervised learning, SimCLR, and DINO (Caron et al., 2021) methods. Notably, the ViT-S model requires significantly more training steps to achieve the same level of alignment under supervised learning compared to when trained with self-supervised objectives like DINO and SimCLR. In contrast, the ResNet50 model, which possesses strong inductive biases due to its convolutional architecture, exhibits relatively consistent alignment dynamics across different learning signals. This robustness implies that models with strong inductive biases are less affected by the choice of training objective, whereas architectures like ViT-S benefit more substantially from rich, self-supervised feedback to achieve optimal alignment.

Limitations and Future Directions. Our study has several limitations. First, the extrapolation of our scaling functions is constrained by the specific range of model sizes and dataset volumes we examined. While we observed power-law relationships between scaling factors and brain alignment, these functions may not generalize beyond the scales tested.

Second, we evaluated a subset of models focusing primarily on standard and modern convolutional neural networks (e.g., ResNets and ConvNeXts), transformer-based architectures (e.g. ViTs) and recurrent networks (CORnet-S). While these architectures cover a range of inductive biases and complexities, they do not encompass the full spectrum of possible neural network designs, such as more biologically plausible models. We see scaling laws as an opportunity to extrapolate the alignment of models at scale, even if their current training is compute-constrained.

Third, our experiments utilized a subset of training datasets primarily from ImageNet and EcoSet. Although these datasets are extensive and widely used, they may not capture all the nuances of visual stimuli relevant to the primate ventral visual stream. Therefore, models trained on other datasets might exhibit improved scaling properties.

While transformer-based architectures like ViTs dominate in many ML tasks due to their flexibility and strong task performance, our findings highlight a clear dissociation between

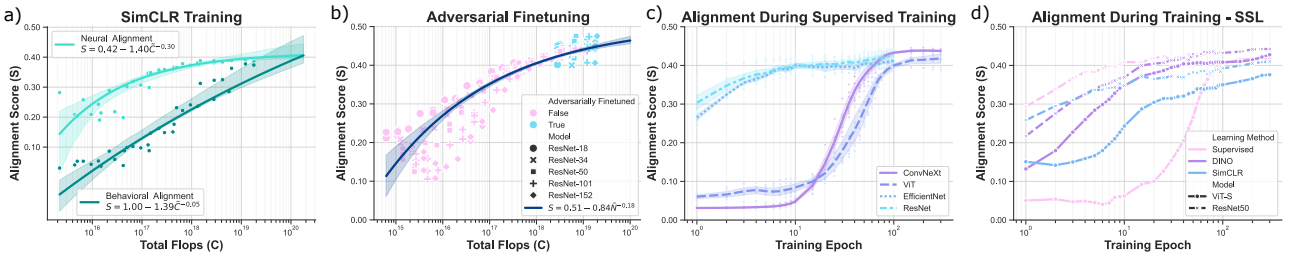

Figure 7: **Alignment Scaling of Alternative Training Strategies.** **a)** SimCLR Training: ResNet models trained with SimCLR improve behavioral alignment with compute, while neural alignment saturates. **b)** Adversarial Robustness: Finetuning ResNets with adversarial training (FGSM) enhances alignment along the scaling curve. **c)** Supervised Training Dynamics: ResNets and EfficientNets exhibit higher initial alignment in early phases of training, due to strong inductive biases, unlike ConvNeXt and ViT. **d)** Self-supervised Training (variants):Alignment dynamics vary with training objectives, but converge to the same alignment saturation.

behavioral and neural alignment. Convolutional models exhibit significantly better neural alignment, especially in early training and low-data regimes—suggesting that biologically inspired inductive biases play a unique role in approximating cortical representations. Interestingly, we also show that recurrence (e.g., CORnet) offers benefits in sample efficiency, but its advantage diminishes with more data. These observations motivate hybrid architectures that blend convolutional, recurrent, and transformer components.

Taken together, our results demonstrate that while scaling both model parameters and training data size enhances behavioral alignment with human visual perception, it leads to saturation in neural alignment with the primate ventral visual stream. Data scaling proves more effective than model scaling in improving alignment, emphasizing the critical role of extensive and diverse training datasets. We also find that architectural choices significantly influence alignment efficiency, with models possessing strong inductive biases—such as fully convolutional networks—achieving higher neural alignment even at initialization. Additionally, the impact of scaling varies across different brain regions, benefiting higher cortical areas more than early visual areas. These findings suggest that merely increasing scale is insufficient for modeling the intricate neural representations of the brain's visual system. Future work should investigate new approaches, including alternative architectures and training strategies, to develop models that more accurately reflect the complexities of neural processing in the primate visual cortex.

To push neural alignment beyond current saturation levels, future research should explore adversarial training methods that encourage models to learn more robust, brain-like representations. Leveraging biologically inspired architectures such as VOneNets (Dapello et al., 2020) may lead to more compute-efficient models achieving higher neural alignment without extensive scaling. Additionally, investigating co-training with brain data—integrating neural recordings directly into the training process—could enhance both neural and behavioral alignment, paving the way for more accurate and efficient brain-like models.

Acknowledgements

This work was partly supported by the Swiss National Science Foundation (SNSF) through the Spark project CRSK-3_228579. We thank the members of the EPFL NeuroAI Lab for their valuable discussions and feedback throughout the project.

Software and Data

We open-source our training and analysis code, along with benchmark results and model checkpoints from our model zoo. All resources are available at: https://github.com/epflneuroailab/scaling-primate-vvs.

Impact Statement

This work seeks to advance our understanding of how scaling model size and dataset size influences alignment with primate visual processing. While the primary goal is scientific—improving models of the visual cortex—it also intersects with broader concerns about resources and impact. Training large-scale models has nontrivial computational costs and environmental implications. At the same time, more accurate brain-like models could yield downstream benefits for neuroscience and machine learning applications (for instance, improved medical imaging or more robust computer vision systems). We do not foresee direct ethical risks specific to this research beyond the usual resource considerations in large-scale model training, and we have used existing public neurophysiological and behavioral datasets.

References

Aghajanyan, A., Yu, L., Conneau, A., Hsu, W.-N., Hambardzumyan, K., Zhang, S., Roller, S., Goyal, N., Levy, O., and Zettlemoyer, L. Scaling laws for generative mixed-modal language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 265–279. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/aghajanyan23a.html.

Antonello, R. J., Vaidya, A. R., and Huth, A. G. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36, 2023.

Azabou, M., Arora, V., Ganesh, V., Mao, X., Nachimuthu, S. B., Mendelson, M. J., Richards, B. A., Perich, M. G., Lajoie, G., and Dyer, E. L. A unified, scalable framework for neural population decoding. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=sw2Y0sirtM.

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. Explaining scaling laws of neural network generalization, 2022. URL https://openreview.net/forum?id=FvfV64rovnY.

Benchetrit, Y., Banville, H., and King, J.-R. Brain decoding: toward real-time reconstruction of visual perception. In *Submitted to The Twelfth International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3y1K6buO8c.

Besiroglu, T., Erdil, E., Barnett, M., and You, J. Chinchilla scaling: A replication attempt, 2024. URL https://arxiv.org/abs/2404.10102.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. Albumentations: Fast and flexible image augmentations. *Information*, 11 (2), 2020. ISSN 2078-2489. doi: 10.3390/info11020125. URL https://www.mdpi.com/2078-2489/11/2/125.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., and Ecker, A. S. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLOS Computational Biology*, 15(4):e1006897, April 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006897. URL http://dx.doi.org/10.1371/journal.pcbi.1006897.

Caro, J. O., de Oliveira Fonseca, A. H., Rizvi, S. A., Rosati, M., Averill, C., Cross, J. L., Mittal, P., Zappala, E., Dhodapkar, R. M., Abdallah, C., and van Dijk, D. BrainLM: A foundation model for brain activity recordings. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RwI7ZEfR27.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.00276. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00276.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., and Konkle, T. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15 (1), October 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-53147-y. URL http://dx.doi.org/10.1038/s41467-024-53147-y.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., and DiCarlo, J. J. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13073–13087. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/98b17f068d5d9b7668e19fb8ae470841-Paper.pdf.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., Jenatton, R., Beyer, L., Tschannen, M., Arnab, A., Wang, X., Riquelme Ruiz, C., Minderer, M., Puigcerver, J., Evci, U., Kumar, M., Steenkiste, S. V., Elsayed, G. F., Mahendran, A., Yu, F., Oliver, A., Huot, F., Bastings, J., Collier, M., Gritsenko, A. A., Birodkar, V., Vasconcelos, C. N., Tay, Y., Mensink, T., Kolesnikov, A., Pavetic, F., Tran, D., Kipf, T., Lucic, M., Zhai, X., Keysers, D., Harmsen, J. J., and Houlsby, N. Scaling vision transformers to 22 billion parameters. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7480–7512. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/dehghani23a.html.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ding, M., Xiao, B., Codella, N., Luo, P., Wang, J., and Yuan, L. Davit: Dual attention vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pp. 74–92. Springer, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., and Movshon, J. A. A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, 16(7):974–981, 2013. doi: 10.1038/nn.3402.

Geiger, F., Schrimpf, M., Marques, T., and DiCarlo, J. J. Wiring up vision: Minimizing supervised synaptic updates needed to produce a primate ventral stream. In *International Conference on Learning Representations (ICLR) Spotlight*, 2022. URL https://openreview.net/forum?id=g1SzIRLQXMM.

Goodale, M. A. and Milner, A. D. Separate visual pathways for perception and action. *Trends in Neurosciences*, 1992.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572.

Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jegou, H., and Douze, M. Levit: A vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12259–12269, October 2021.

Grill-Spector, K., Kourtzi, Z., and Kanwisher, N. The lateral occipital complex and its role in object recognition. *Vision Research*, 41(10-11):1409–1422, 2001.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pp. 770–778. IEEE, June 2016. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., and McCandlish, S. Scaling laws for autoregressive generative modeling, 2020.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically, 2017.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Vinyals, O., Rae, J. W., and Sifre, L. An empirical analysis of compute-optimal large language model training. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=iBBcRUlOAPR.

Isik, B., Ponomareva, N., Hazimeh, H., Paparas, D., Vassilvitskii, S., and Koyejo, S. Scaling laws for downstream task performance of large language models, 2024. URL https://arxiv.org/abs/2402.04177.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. *Principles of Neural Science*. McGraw-Hill, 2000.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. 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.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., and Kriegeskorte, N. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, October 2019. ISSN 1091-6490. doi: 10.1073/pnas.1905544116. URL http://dx.doi.org/10.1073/pnas.1905544116.

Kim, H. Torchattacks: A pytorch repository for adversarial attacks, 2020.

Kriegeskorte, N. et al. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Kubilius, J., Schrimpf, M., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L. K., and DiCarlo, J. J. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. In Wallach, H., Larochelle, H., Beygelzimer, A., D'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Neural Information Processing Systems (NeurIPS)*, pp. 12785—-12796. Curran Associates, Inc., 2019.

LeCun, Y., Cortes, C., and Burges, C. J. The mnist database. https://yann.lecun.com/exdb/mnist/, 1998. [Accessed 16-11-2024].

Li, W., Wang, L., Li, W., Agustsson, E., and Gool, L. V. Webvision database: Visual learning and understanding from web data. *CoRR*, abs/1708.02862, 2017. URL http://arxiv.org/abs/1708.02862.

Linsley, D., Rodriguez Rodriguez, I. F., FEL, T., Arcaro, M., Sharma, S., Livingstone, M., and Serre, T. Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 28873–28891. Curran Associates, Inc., 2023.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Loosli, G., Bottou, L., and Simard, P. projects:infimnist [leon.bottou.org] — leon.bottou.org. https://leon.bottou.org/projects/infimnist, 2007. [Accessed 16-11-2024].

maintainers, T. and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016.

Majaj, N. J., Hong, H., Solomon, E. A., and DiCarlo, J. J. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *The Journal of Neuroscience*, 35(39):13402–13418, 2015. doi: 10.1523/jneurosci.5181-14.2015.

Malach, R., Levy, I., and Hasson, U. The topography of high-order human object areas. *Trends in Cognitive Sciences*, 6(4):176–184, 2002.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., and Kietzmann, T. C. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8), 2021. doi: 10.1073/pnas.2011417118.

Mehta, S. and Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vh-0sUt8HlG.

Muennighoff, N., Rush, A. M., Barak, B., Scao, T. L., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. Scaling data-constrained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=j5BuTrEj35.

Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., and Kornblith, S. Human alignment of neural network representations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ReDQ1OUQR0X.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., and Yamins, D. L. Task-driven convolutional recurrent models of the visual system. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/6be93f7a96fed60c477d30ae1de032fd-Paper.pdf.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2023.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J. J. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *The Journal of Neuroscience*, 38(33): 7255–7269, 2018. doi: 10.1523/jneurosci.0388-18.2018.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. Imagenet-21k pretraining for the masses, 2021.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., and DiCarlo, J. J. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018. URL https://www.biorxiv.org/content/10.1101/407007v2.

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., and DiCarlo, J. J. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 2020. URL https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S. R., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

Susmelj, I., Heller, M., Wirth, P., Prescott, J., and Ebner, M. Lightly, 2020.

Tan, M. and Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Team, T. M. M. composer. https://github.com/mosaicml/composer/, 2021.

Touvron, H., Cord, M., and Jégou, H. Deit iii: Revenge of the vit, 2022. URL https://arxiv.org/abs/2204.07118.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., and Li, Y. Maxvit: Multi-axis vision transformer. *ECCV*, 2022.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Vasu, P. K. A., Gabriel, J., Zhu, J., Tuzel, O., and Ranjan, A. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJx040EFvH.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

Yamins, D. L. K. and DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, February 2016. ISSN 1546-1726. doi: 10.1038/nn.4244. URL http://dx.doi.org/10.1038/nn.4244.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12104–12113, June 2022.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., and Yamins, D. L. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), 2021. doi: 10.1073/pnas.2014196118.

A. Implementation Details

Our experiments are conducted using the PyTorch framework (Paszke et al., 2019), with Composer (Team, 2021) employed as the GPU orchestration tool to efficiently manage computational resources.

For image augmentations, we leverage the Albumentations Buslaev et al., 2020 library due to its rich set of augmentation techniques, which are crucial for enhancing model robustness and preventing overfitting. In experiments involving self-supervised learning, we use the Lightly (Susmelj et al., 2020) library to facilitate the implementation of self-supervised losses, augmentations, and model heads. This library streamlines the process of setting up models for SimCLR and DINO training methods.

To generate adversarial examples for adversarial fine-tuning, we employ the Torchattacks library (Kim, 2020). Specifically, we use the Fast Gradient Sign Method (FGSM) to create perturbations that challenge the models, aiming to enhance their alignment with neural representations by exposing them to adversarial inputs.

B. Additional Image Datasets

To further validate our findings across diverse image distributions and to estimate scaling curves across different sample scales, we trained ResNet18 models on subsets of several large-scale image datasets: ImageNet-21k-P, WebVision-P, iNaturalist, and Places365. Below, we provide detailed descriptions of each dataset.

B.1. ImageNet21k-P

ImageNet-21k-P is a processed subset of the full ImageNet-21k dataset (Ridnik et al., 2021), which originally contains over 14 million images organized into more than 21,000 categories following the WordNet hierarchy. The ”P” denotes a pruned version where classes with insufficient images or noisy labels are filtered out to enhance dataset quality. This results in a refined dataset that maintains the richness of the original ImageNet-21k while improving label accuracy and image relevance. The resulting dataset contains approximately 11 million training images across 10,450 classes.

B.2. WebVision-P

The WebVision dataset (Li et al., 2017) is a large-scale web image dataset designed to provide a real-world, noisy alternative to ImageNet. It originally contains over 16 million images categorized into 5,000 classes. The images are collected from the internet using queries from search engines like Google and Flickr, leading to a dataset that includes label noise, varying image resolutions, and diverse visual contexts. Due to classes with very few available samples, we processed the WebVision dataset similarly to ImageNet-21k-P to remove classes with insufficient images. The resulting dataset, which we denote as WebVision-P, contains approximately 13.5 million training images across 4,189 categories.

B.3. iNaturalist

iNaturalist (Van Horn et al., 2018) contains 2.7 million photographs of organisms in their natural environments, representing 10000 species. The dataset features highly specialized fine-grained categories and natural backgrounds, offering insight into how domain-specific visual features influence alignment scaling.

B.4. Places365

Places365 (Zhou et al., 2017) is a large-scale scene-centric dataset containing approximately 1.8 million training images across 365 scene categories. Unlike object-centric datasets such as ImageNet, Places365 focuses on the recognition of environmental scenes, including natural landscapes, urban settings, and indoor environments. Each category includes a wide variety of images to capture the diversity within scene types.

B.5. infiMNIST

The MNIST dataset (LeCun et al., 1998) is a classic benchmark in machine learning, comprising 70,000 grayscale images of handwritten digits (0-9), each sized 28×28 pixels. To expand this dataset for more extensive experimentation, we utilize the Infinite MNIST (Infimnist) tool Loosli et al., 2007, which generates additional MNIST-like samples through data augmentation techniques. We create an extended dataset by modifying the original training dataset 19 additional times,

resulting in a total of 1.2 million images. This enlarged dataset allows for a more thorough evaluation of scaling effects on the alignment.

B.6. CIFAR10

CIFAR-10 (Krizhevsky et al.) is a widely used benchmark of 60,000 low-resolution (32×32) color images divided evenly into 10 object classes. The dataset comprises 50,000 training images and 10,000 test images, with 6,000 samples per class. To match our scaling protocol, we created class-balanced subsets by sampling $d \in \{10, 30, 100, 300, 1000, 3000\}$ images per class. Because CIFAR-10 is two orders of magnitude smaller than our other baselines, in addition to our standard 100-epoch runs we also trained models on the full dataset for extended durations of 250, 1000, and 2500 epochs to assess convergence and scaling effects.

C. Validation on private data

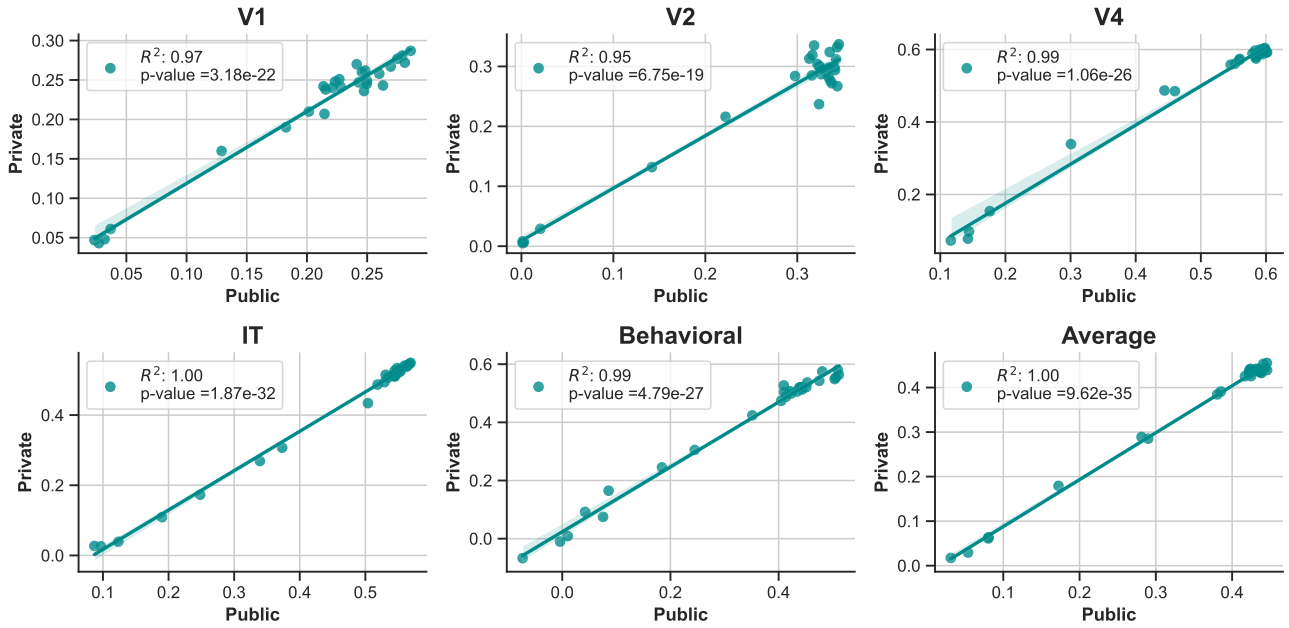

Figure S1: Public benchmarks used in this study correlates highly with private benchmarks on Brain-Score

As described in Section 2 we test a diverse set of models on private benchmarks on Brain-Score platform. All R^2 values are above 0.95 with p-values less than 10^{-18}.

D. Pretrained Models

As part of our comprehensive evaluation, we benchmarked a diverse set of pretrained models sourced from both `torchvision` (maintainers & contributors, 2016) and the `timm` (Wightman, 2019) libraries. We tested a total of 94, including ViT (Dosovitskiy et al., 2021), DaViT (Ding et al., 2022), LeViT (Graham et al., 2021), ConvNeXt (Liu et al., 2022a), MobileViT (Mehta & Rastegari, 2022), MaxVit (Tu et al., 2022), FastViT (Vasu et al., 2023). Each model varies in parameter count, training sample size, dataset source, and training objective, providing a broad spectrum for analysis.

To verify the generalizability of our findings, we conducted evaluations with these pretrained models, including larger networks like CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023), which are pretrained on richer and more diverse datasets such as LAION (Schuhmann et al., 2021; 2022). We also compared variations of these models by examining base pretrained models alongside their fine-tuned counterparts on ImageNet, aiming to investigate the impact of fine-tuning on scaling behavior.

Our results indicate that models with extensive pretraining achieve enhanced behavioral alignment, likely due to their exposure to richer and more varied data. However, similar to models trained solely on ImageNet or EcoSet, these pretrained models still exhibit a saturation effect in neural alignment with the primate visual ventral stream (VVS). This suggests that while larger and more diverse datasets improve behavioral predictability, they do not substantially extend the scaling of neural alignment beyond the observed plateau.

The curves in Figure S2 closely follow the scaling patterns estimated for our trained models shown in Figure 2.c, further validating that the observed saturation is consistent across different pretraining regimes and dataset scales. This reinforces our conclusion that scaling alone is insufficient to overcome the limitations in neural alignment and highlights the need for alternative approaches to improve alignment with neural representations.

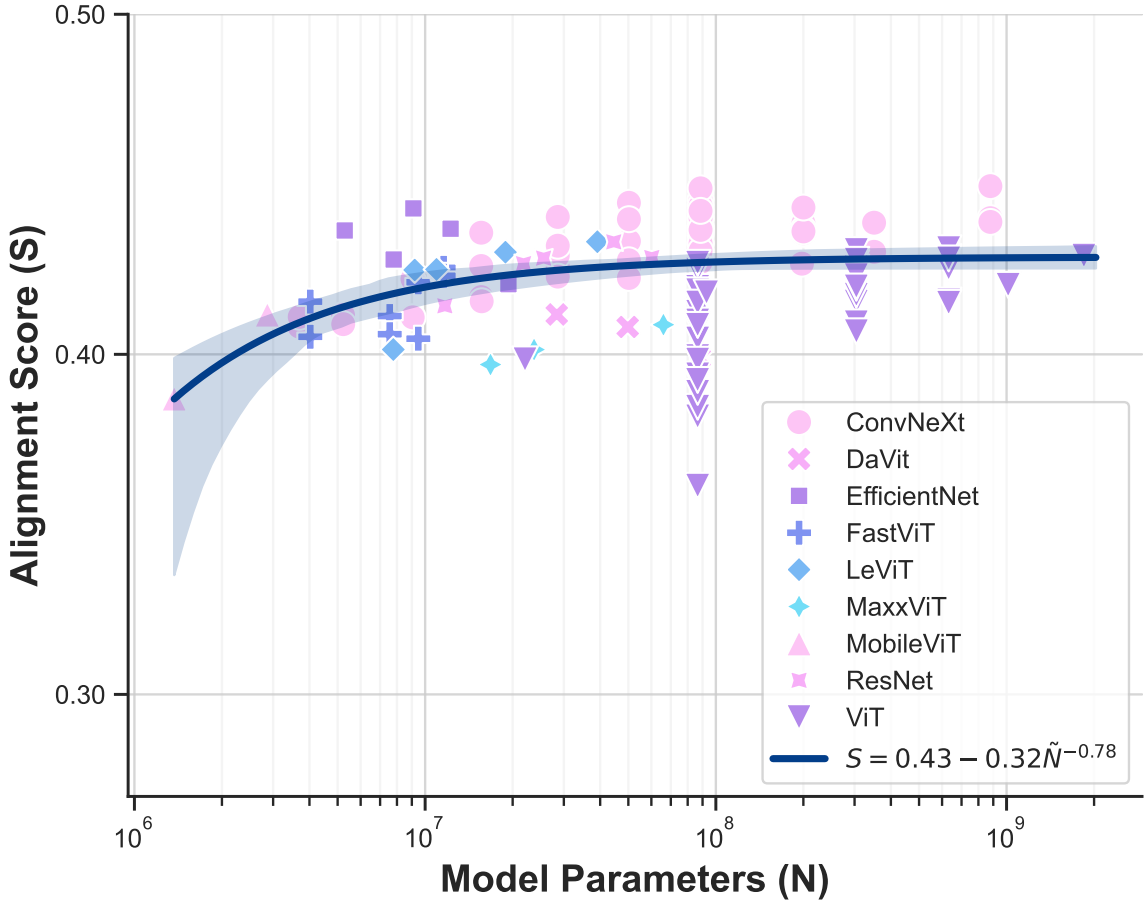

Figure S2: Alignment of pretrained vision models as a function of model parameters

E. Training Evolution

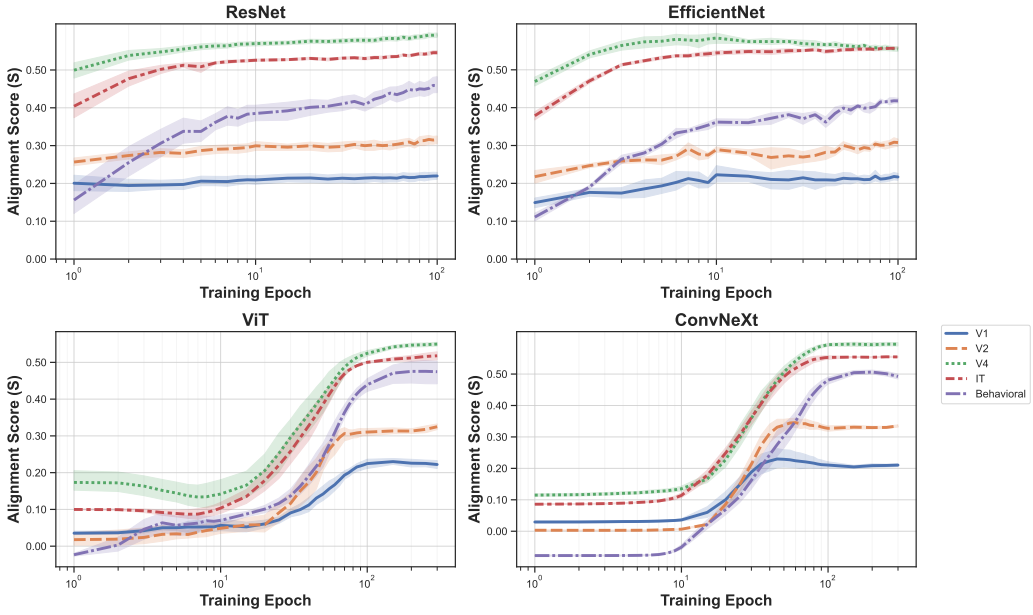

Figure S3: Evolution of per-region alignment throughout training. Models with stronger priors—such as ResNet and EfficientNet—exhibit higher neural alignment initially. However, the gap in representational power diminishes as more generalist models are trained on data.

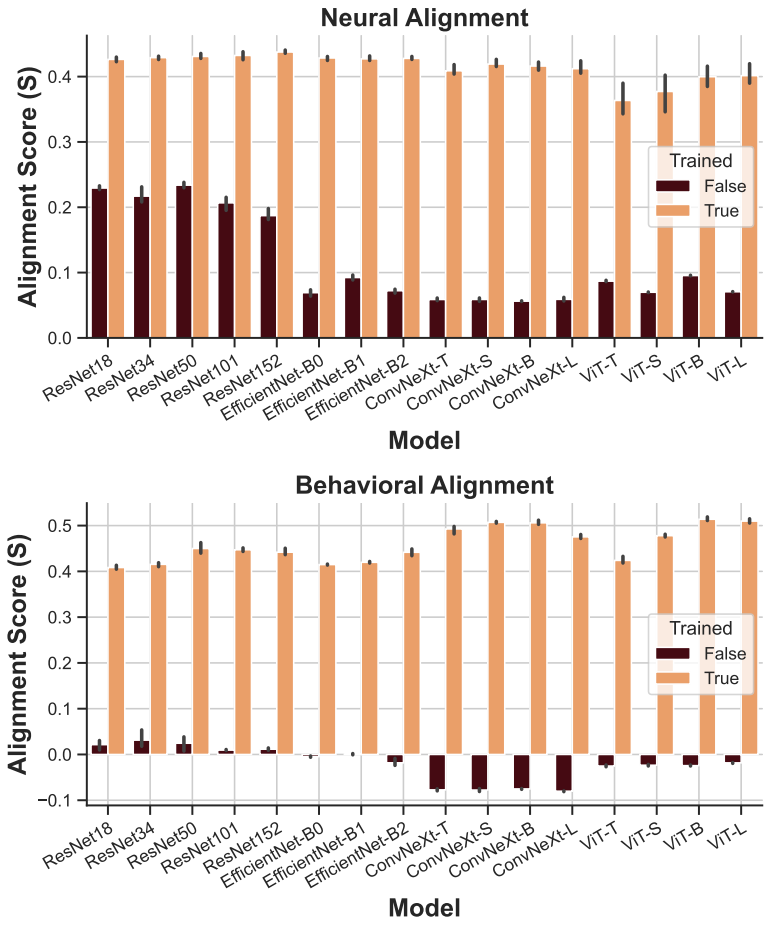

Figure S4: Some untrained models demonstrate non-zero neural alignment at initialization. Nevertheless, all models start with almost zero behavioral alignment, indicating that initial neural alignment arises from architectural biases rather than learned behavior.

F. Effect of Training Objective

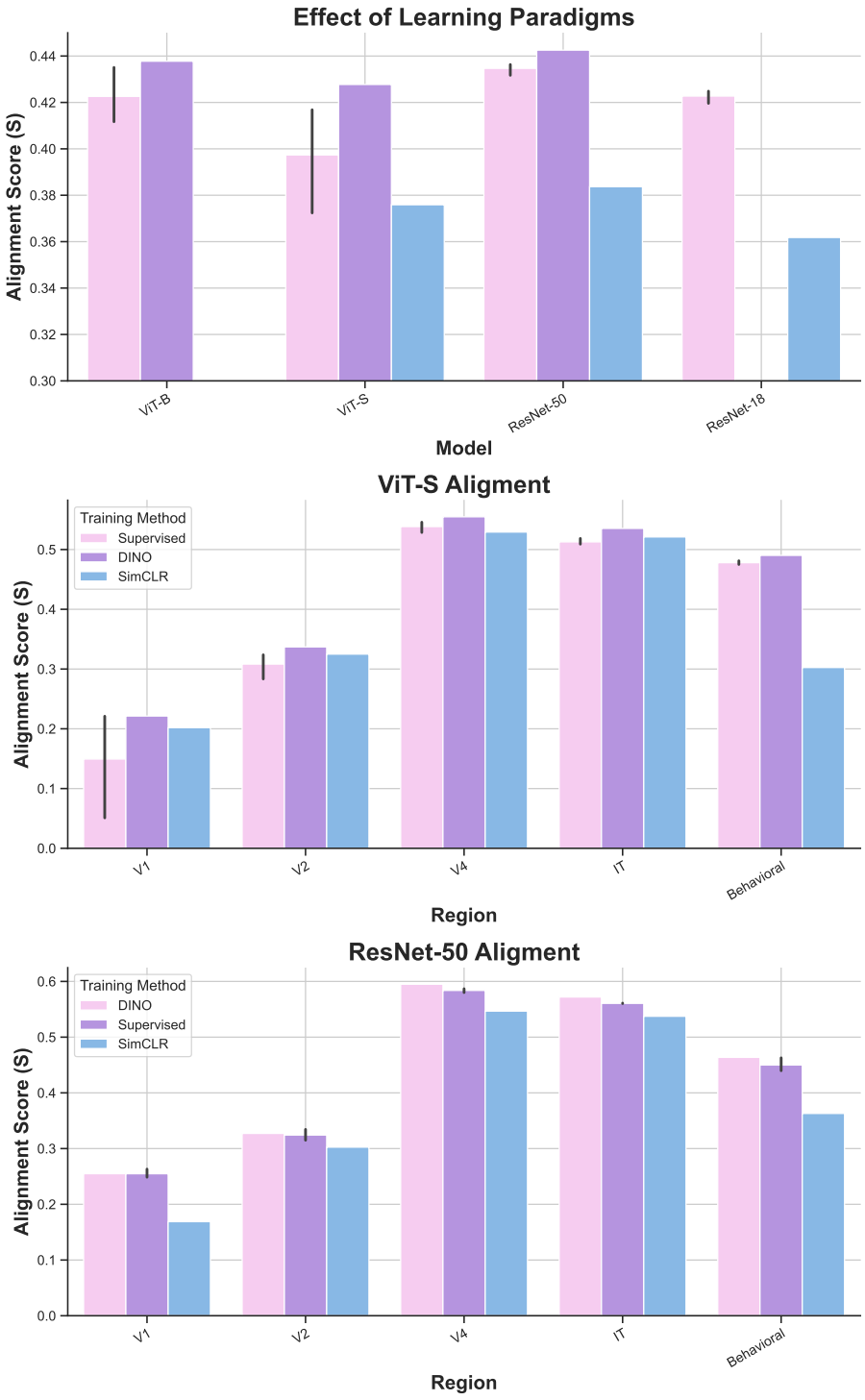

Figure S5: Effect of training objective on alignment

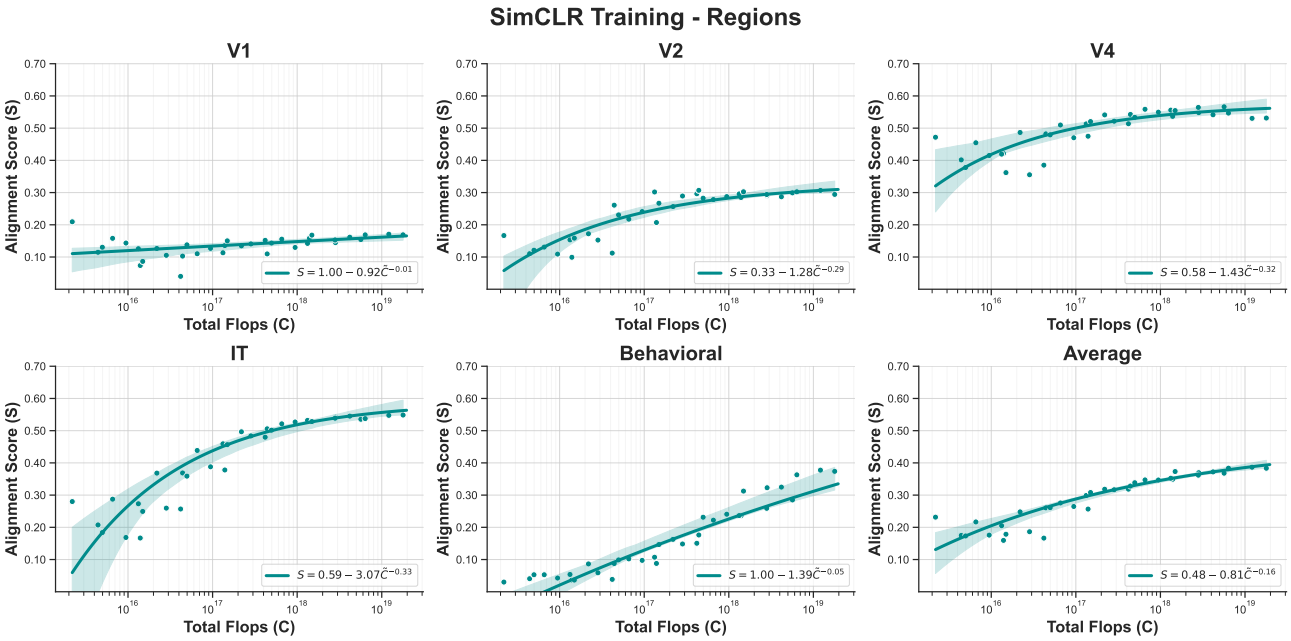

Figure S6: Scaling of SimCLR training across regions

G. Region-wise scatter plots

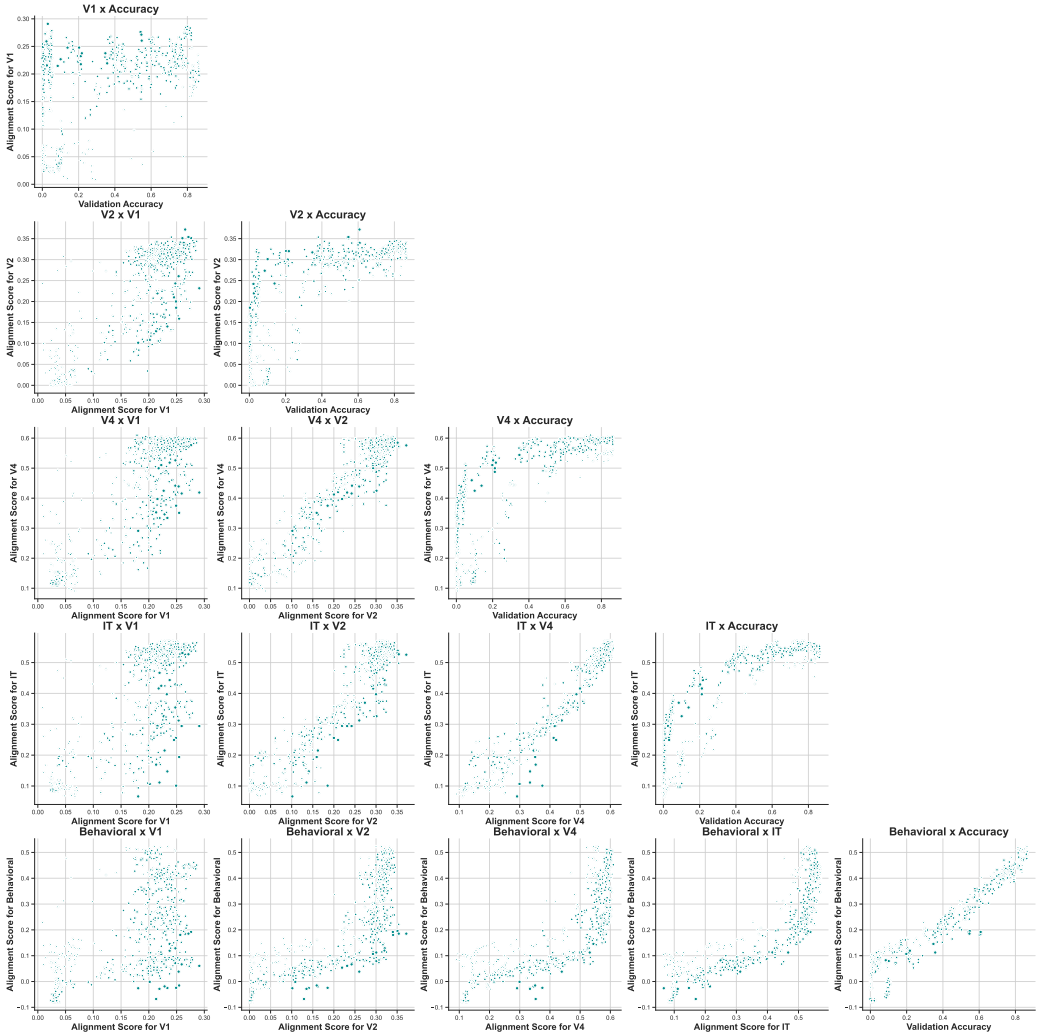

Figure S7: Region vs. Region Comparisons: This figure shows how the alignment scores for each brain region correlate with those of other regions. The diagonal plots illustrate the relationship between the alignment score of each region and the validation accuracy on ImageNet and EcoSet datasets.