Beyond neural predictivity: Scale tolerance as a mechanistic test of ventral-stream models

Iván Felipe Rodríguez[1†], Xizheng Yu[1†], Nishka Pant[1], Jacob Mulliken[1], Thomas Serre[1*]

[1]Department of Cognitive & Psychological Sciences, Brown University, 190 Thayer Street, Providence, 02912, RI, USA.

*Corresponding author(s). E-mail(s): thomas_serre@brown.edu;
†These authors contributed equally to this work.

Abstract

How the primate ventral stream builds invariant object representations remains a central problem. Task-optimized deep networks can achieve strong recognition and high neural predictivity on standard benchmarks, yet predictive scores alone do not specify mechanism. We probe retinal-size tolerance as a mechanistic test. After large-scale categorization training, an HMAX-family model that assumes a scale-space organization enabling local Hubel–Wiesel pooling builds size tolerance gradually and transfers it to novel categories. On neuroscience probes, it matches human cross-size generalization for novel objects, preserves macaque IT predictivity under scale-split mapping, and captures V4-like scale-normalized curvature coding. Performance-matched CNNs, which must learn tolerance from data without an explicit scale-pooling circuit, fail these tests despite similar benchmark accuracy and standard predictivity. These results show that purely data-driven predictive benchmarks can mask mechanistic mismatches, motivating targeted perturbation tests when using models as explanations of cortical computation.

Introduction

A central goal of systems neuroscience is to explain how the primate ventral visual stream transforms retinal inputs into representations that support rapid, transformation-tolerant object recognition. Modern task-optimized deep networks can achieve strong benchmark accuracy and high neural predictivity on standard datasets,

but predictive scores alone do not determine mechanism: distinct computations can yield similar stimulus–response alignment. This motivates diagnostic tests that target specific computations thought to underlie cortical processing.

Here we use *retinal scale tolerance*—the ability to preserve object identity across changes in retinal size—as a mechanistic probe of ventral-stream models. Humans can generalize across large size changes after minimal exposure [1], and ventral-stream neurons exhibit tolerant codes that remain stable across changes in size [2, 3], including curvature tuning in area V4 [4]. These signatures provide constraints that go beyond single operating-point accuracy or predictivity.

Classic work by Hubel and Wiesel established a foundational computational motif for early vision: receptive fields increase in complexity and tolerance through alternating stages of selectivity and pooling [5]. In this view, tolerance can be constructed from *local* pooling operations when the relevant stimulus dimension is organized topographically. Retinotopy provides such an organization for position (and phase) and motivates repeated local circuits in V1 [6]. Multiple studies also report systematic organization of spatial-frequency preference in primate V1 and extrastriate cortex [7–11]. HMAX extends the Hubel–Wiesel pooling motif by positing an additional organization across scale—often conceptualized as a size or spatial-frequency axis—closely related to classical scale-space theory [12–15]. A complementary normative perspective derives families of early receptive fields from scale-space axioms and covariance requirements, providing theoretical support for representing images over an explicit scale parameter [16].

A strong empirical motivation for quantitative models of tolerance-building came from the paperclip experiments of Logothetis and colleagues [17]. Monkeys trained to recognize novel three-dimensional objects exhibited IT neurons that were selective for object identity and view, yet tolerant to changes in retinal position and scale. Substantial scale tolerance was observed despite fixed viewing distance during training, suggesting that tolerance to common two-dimensional transformations is not learned anew for each object but instead relies on largely generic computations [18]. Riesenhuber and Poggio's HMAX model was explicitly designed to account for these properties [19]. It instantiates a feedforward hierarchy of alternating simple (S) and complex (C) stages, where selectivity increases through template matching while tolerance emerges through pooling over afferent units with similar tuning. In this framework, generic invariance is built gradually through cascaded local pooling in space and (under an explicit scale-space organization) in scale [20]. A key prediction—that complex cells implement a max-like pooling operation—was later supported by intracellular recordings showing that complex-cell responses are well described by max-like nonlinearities over simple-cell inputs [21].

In parallel, deep learning for vision has advanced rapidly through large-scale supervised training and engineering-driven refinements—deeper hierarchies stabilized by skip connections and normalization, improved optimization and regularization, and feature aggregation across spatial scales—optimized end-to-end for benchmark recognition tasks (reviewed in [22]). This progress helped establish a goal-driven modeling strategy: optimize a model to solve an ethologically relevant recognition task and then test its internal representations against neural measurements. Seminal work showed

that task-optimized networks can predict responses in macaque IT and V4, often outperforming earlier biologically inspired models [23, 24], and this relationship was later formalized in standardized benchmarks such as Brain-Score [25].

Critically, however, most modern convolutional networks are developed with few explicit neurophysiological constraints beyond convolution and local spatial pooling; invariances such as scale tolerance are therefore induced by task optimization on the available training examples. Empirically, convolution and pooling alone do not guarantee broad translation or scale tolerance: invariance can increase with depth but remains incomplete and strongly dependent on architecture, training objectives, and augmentation [26–30]. Thus, a model can be accurate and neurally predictive at a reference operating point while still implementing a transformation code that differs from the ventral stream.

Here we test whether scale tolerance can serve as a targeted diagnostic for ventral-stream models at contemporary performance levels. We systematically modernize the HMAX framework while preserving its tolerance-building organization and compare it to standard convolutional networks matched for large-scale recognition performance. We then evaluate whether scale tolerance generalizes to novel categories, whether it matches human cross-size generalization under one-shot conditions, whether neural predictivity is preserved under a *scale-split* perturbation of the mapping regime, and whether intermediate representations reproduce a physiological signature of V4 scale-normalized curvature coding.

Results

We asked whether models that achieve comparable task performance and conventional neural predictivity nevertheless differ in how they construct invariant visual representations. We used large-scale object recognition both to train these models and to match their overall performance, then probed whether the resulting representations generalize scale tolerance to unfamiliar objects. We compared modernized HMAX-family models and standard convolutional networks across four analyses: (i) large-scale recognition performance, (ii) generalization across retinal size when test objects differ from those used to train the model, (iii) human cross-size discrimination after minimal exposure, and (iv) neural alignment and physiological signatures along the primate ventral stream.

Modernizing a ventral-stream architecture preserves object recognition performance

A common concern regarding biologically grounded models is that architectural commitments may limit performance on large-scale recognition tasks. To test this directly, we progressively modernized the HMAX architecture to operate in a contemporary performance regime while preserving its core tolerance-building organization. These updates included replacing fixed feature extractors with learnable convolutional filters, increasing depth via residual connections, and enabling end-to-end optimization.

Across this progression, each modernization step yielded systematic gains in ImageNet top-1 accuracy (Figure 1). The final model achieved performance comparable to

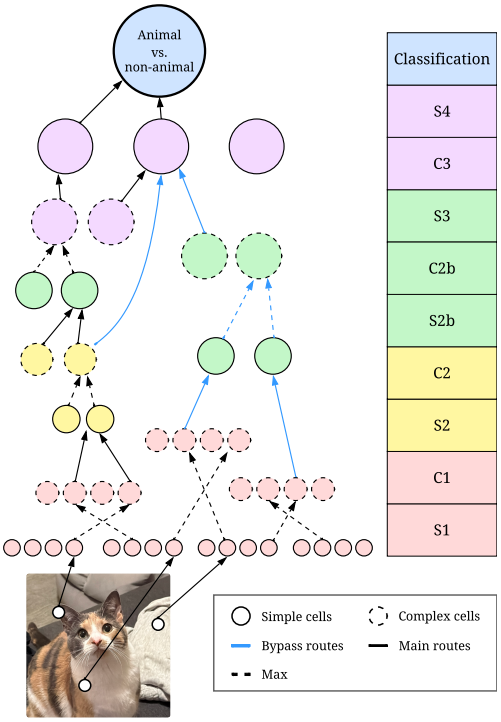


Fig. 1: HMAX Architecture

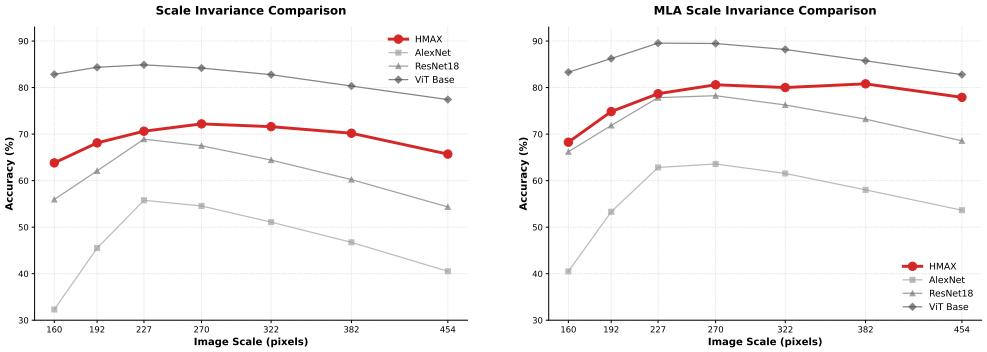
widely used convolutional architectures in this regime. Importantly, these gains were obtained without abandoning explicit multi-scale processing or hierarchical tolerance-building, establishing that ventral-stream–inspired model families are not intrinsically confined to low-performance regimes. This provides a controlled basis for comparing invariance at similar levels of task accuracy.

Models differ in how they generalize across retinal scale

We next examined whether performance-matched models differ in their ability to generalize across changes in retinal object size. To isolate intrinsic scale tolerance from effects of augmentation, models were trained at a single reference scale and evaluated across a range of test scales spanning approximately one octave.

Under this off-reference evaluation, standard convolutional networks showed pronounced degradation as test scale deviated from the training scale (Figure 2a). In contrast, modernized HMAX-family models maintained stable performance across the same range. This difference persisted despite comparable accuracy at the reference scale, indicating that it is not explained by overall task difficulty or underfitting.

Consistent with prior work, these results do not imply that convolutional networks lack scale tolerance altogether, but rather that such tolerance is typically restricted to the range of transformations represented in the training data and does not generalize reliably beyond that range [28, 29].



(a) ImageNet results.

(b) Multi-label ImageNet results.

Fig. 2: Scale generalization across seven test scales on ImageNet and Multi-Label ImageNet.

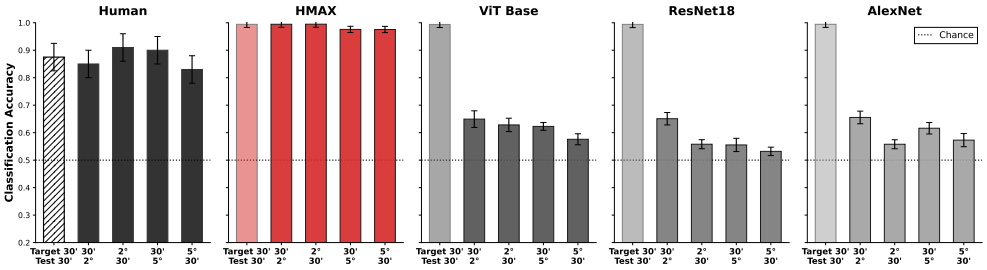


Fig. 3: One-shot Hangul discrimination across scale.

Human-like one-shot generalization distinguishes model families

To assess whether the differences observed above extend to regimes relevant for perception, we evaluated models on a one-shot discrimination task using unfamiliar Hangul characters. This paradigm probes generalization of identity across large size changes after minimal exposure, a hallmark of human visual recognition.

Modernized HMAX-family models matched human performance under out-of-distribution scale conditions, whereas standard convolutional networks showed marked declines despite strong in-distribution performance (Figure 3). This dissociation mirrors the ImageNet scale generalization results and demonstrates that differences in scale tolerance manifest not only in large-scale classification benchmarks but also in controlled behavioral paradigms that approximate human perceptual demands.

Neural predictivity diverges under scale perturbation

We next asked whether differences in scale tolerance are reflected in alignment with neural responses in inferotemporal cortex. Using the Brain-Score framework, we first

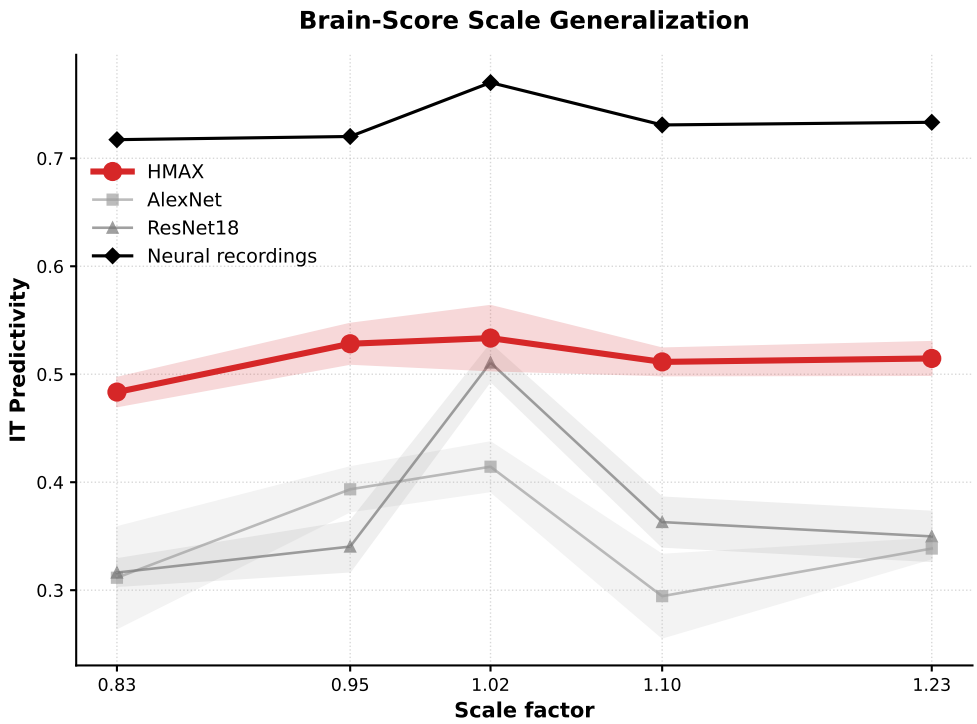


Fig. 4: Brain-Score performance under scale variation.

confirmed that modernized HMAX models and standard convolutional networks can achieve comparable neural predictivity under conventional evaluation protocols.

We then applied a *scale-split* analysis as a diagnostic probe of generalization beyond the regime used to fit neural mappings (rather than as a new benchmark): mappings were fit at a reference scale and evaluated on stimuli at unseen scales. Under this procedure, a clear divergence emerged (Figure 4). Neural predictivity of standard convolutional networks declined substantially under scale shifts, whereas modernized HMAX models preserved alignment across scale. Notably, this divergence was not apparent under standard evaluation, indicating that single-operating-point predictivity can obscure differences in how models represent identity-preserving transformations.

Intermediate representations capture a physiological signature of V4 invariance

Finally, we tested whether architectural differences in scale tolerance are reflected at intermediate stages of processing. We replicated a classic electrophysiological paradigm demonstrating that neurons in macaque area V4 encode contour curvature in a scale-normalized manner, maintaining tuning stability across changes in retinal size [4].

Model	Layer	Invariance Score	MC χ^2	p
HMAX	C2	0.080	8.58	0.21
AlexNet	Best	0.609	26.76	0.0004
ResNet-18	Best	0.754	31.82	< 0.001
ViT-Base/16	Best	0.156	51.37	< 0.001

Table 1: Replication of scale-normalized curvature tuning in macaque V4. Invariance score was defined as the median absolute value of the per-neuron scale-invariance slopes, details listed in S1.1.

Intermediate layers of the modernized HMAX hierarchy reproduced normalized curvature tuning consistent with macaque V4, whereas no layer in standard convolutional networks exhibited comparable stability, even when selecting the best-matching layer post hoc (Table 1). This result links differences observed at the behavioral and benchmark levels to differences in how intermediate representations encode shape information across scale.

Methods

Reproducibility and implementation details.

For complete reproducibility, we provide the full set of architectural specifications, training hyperparameters, augmentation policies, and evaluation protocols in Supplementary Section S1.1.

Our goal was to test how different modeling approaches construct invariant visual representations when models are scaled to contemporary performance regimes. To this end, we compared modernized HMAX-family architectures to standard convolutional neural networks under matched evaluation protocols that probe object recognition performance, out-of-distribution generalization across scale, and alignment with neural and physiological data from the primate ventral stream.

Model architectures

We evaluated a family of progressively modernized HMAX architectures alongside standard convolutional baselines (e.g., AlexNet- and ResNet-style models). The HMAX-family models preserve the core organizational principles of the original framework: explicit multi-scale processing, alternating stages of feature selectivity and pooling, and progressive construction of tolerance across a hierarchical feedforward pathway.

To operate in a modern performance regime, HMAX architectures were updated with learnable convolutional filters, residual connections, and end-to-end optimization. Fixed Gabor filters were replaced with trainable convolutional kernels, and deeper hierarchies were enabled via residual connections while preserving the simple/complex (S/C) structure. Explicit scale channels were implemented using dense sampling of image scale, and scale information was integrated hierarchically through differentiable

pooling operations rather than hard max selection. These design choices allow the model to remain trainable at scale while retaining a ventral-stream–like organization.

Standard convolutional networks served as performance-matched baselines. All models were trained and evaluated using identical datasets and task definitions unless otherwise noted.

Training procedures

Models were trained using supervised learning for object categorization, with cross-entropy loss optimized via stochastic gradient descent–based methods. To isolate architectural contributions to scale tolerance, HMAX-family models were trained without extensive scale jitter, whereas baseline convolutional networks were trained using conventional augmentation practices unless explicitly stated. This contrast is informative because standard CNNs do not include an explicit scale axis with local pooling; any size tolerance must be discovered through optimization and the examples provided during training.

For HMAX architectures, an additional hierarchical scale-alignment objective was applied during training to encourage consistency of representations across adjacent scales throughout the hierarchy. This regularization promotes gradual tolerance building in a manner consistent with ventral-stream processing. Full architectural specifications, optimization parameters, and training schedules are reported in the Supplementary Materials.

Scale generalization protocols

To assess intrinsic scale tolerance, we adopted an off-reference evaluation paradigm. Models were trained at a single reference scale and tested across a range of retinal scales spanning approximately one octave. Performance was quantified both at the reference scale and on held-out scales to distinguish in-distribution accuracy from true extrapolative generalization.

This protocol was applied consistently across datasets, including ImageNet-scale natural images and controlled benchmark stimuli. By decoupling training and test scales, this evaluation isolates mechanisms of scale tolerance from effects driven by data augmentation.

Behavioral evaluation: one-shot generalization

To probe human-like generalization, we evaluated models on a one-shot discrimination task using unfamiliar Hangul characters, following established psychophysical procedures. Models were exposed to a single example of a target character and required to discriminate it from visually similar distractors under varying scale conditions.

Model decisions were based on similarity in feature representations extracted from the penultimate layer. Performance was assessed separately for in-distribution and out-of-distribution scale conditions and compared to behavioral data from human observers. Sensitivity was quantified using signal detection metrics where appropriate.

Neural evaluation: Brain-Score and scale-split analysis

Neural alignment was assessed using the Brain-Score framework, which quantifies the correspondence between model representations and neural responses recorded from macaque inferotemporal cortex. For each model, linear mappings from model activations to neural responses were learned using standard cross-validated procedures, and predictivity was measured as noise-ceiling–normalized correlation.

To probe generalization beyond the fitting regime, we implemented a scale-split evaluation. Linear mappings were fit using stimuli presented at a reference scale and evaluated on neural responses to stimuli at unseen scales. This procedure tests whether neural predictivity reflects robust, scale-tolerant representations or scale-specific correlations.

Physiological validation in area V4

To test whether intermediate model representations reproduce known physiological signatures of ventral-stream processing, we replicated a classic paradigm examining scale-normalized curvature tuning in macaque area V4 [4]. Model units were probed with the same class of parametric shape stimuli across multiple retinal sizes.

For each responsive unit, tuning curves were computed as a function of curvature at each scale, and invariance was quantified by measuring the stability of tuning preferences across scale. Model-derived distributions were compared directly to published V4 data using identical metrics and statistical tests, enabling a mechanistic comparison between artificial and biological representations.

Discussion

The present study revisits a biologically grounded model of the ventral visual stream to test which computational principles remain explanatory when models are scaled to contemporary performance regimes. By systematically modernizing the HMAX framework while preserving its core organizational commitments, we show that different modeling approaches can achieve similar levels of task performance and neural predictivity while diverging in how invariant visual representations are constructed. Together, these results clarify how invariant representations depend on the form of computation rather than task optimization alone.

Architectural differences and the form of invariance

A central finding of this work is that performance-matched models can differ markedly in the *form* of invariance they implement. Modern convolutional networks can recognize trained categories across a range of sizes, often aided by scale augmentation, but this tolerance is not enforced by a local pooling circuit across a scale axis and can transfer poorly to novel categories [1, 27]. In contrast, modernized HMAX-family models exhibit robust generalization across scale without reliance on large numbers of scale-augmented examples. This distinction is evident across natural-image benchmarks, one-shot behavioral tasks, and neural evaluations, indicating that differences

in tolerance-building mechanisms lead to qualitative differences in representation that are not captured by aggregate performance metrics.

Importantly, this conclusion does not contradict evidence that CNNs can support substantial invariance under appropriate training conditions. Psychophysical comparisons show that both humans and CNNs can exhibit robust online translation tolerance following extreme displacements, provided that the networks are trained with suitable data and architectural choices [31]. However, such invariance is typically acquired through exposure to specific transformations and does not appear to generalize broadly across untrained regimes, nor does it necessarily align with the invariance profiles observed in human perception [30]. The present results extend this literature by showing that models with similar task accuracy and neural predictivity can nevertheless differ in how invariance is constructed and generalized.

More broadly, these dissociations motivate incorporating *topographic* constraints as mechanistic priors in models of visual cortex. The ventral stream is not only a hierarchy of response properties, but also a hierarchy of spatial organization: retinotopy and early feature maps in V1 are complemented by reliable large-scale structure in higher cortex, including clustered category-selective regions and semantic gradients in IT/VTC. Recent work shows that augmenting task-optimized networks with spatial objectives that approximate wiring-length minimization or enforce response smoothness across a cortical sheet can reproduce key aspects of this organization and, in some cases, improves behavioral alignment [32, 33]. Complementary approaches derive domain topography from explicit connectivity constraints (e.g., distance-dependent wiring and sign constraints) while learning high-level visual representations [34]. Together, these results suggest that predictive benchmarks based solely on stimulus–response alignment are underconstrained: adding topographic constraints can shrink the space of admissible solutions and yield models that are more mechanistically interpretable, complementing the scale-space pooling commitments studied here.

These results support the view that invariance in the ventral stream is not merely a statistical consequence of exposure to transformed exemplars, but reflects computational structure that builds tolerance progressively across hierarchical stages [17, 35]. Explicit multi-scale processing and hierarchical pooling provide one plausible mechanism for constructing such representations, consistent with longstanding theories of ventral-stream organization.

Implications for neural predictivity benchmarks

Our findings refine the interpretation of neural predictivity benchmarks such as Brain-Score [25]. While modern deep networks often achieve strong correspondence with neural responses under standard evaluation protocols [23, 24], this alignment does not uniquely constrain underlying computation. By introducing a scale-split evaluation, we show that models with similar Brain-Score values can diverge substantially when tested outside the regime used to fit neural mappings. In this setting, models that preserve scale tolerance maintain neural predictivity, whereas standard convolutional networks do not.

These results do not undermine the utility of neural predictivity benchmarks; rather, they highlight the importance of probing generalization beyond the fitting

regime. Neural predictivity provides an important constraint, but it is insufficient on its own to adjudicate between competing mechanistic accounts of visual representation. Incorporating targeted perturbations that probe specific invariances may therefore be critical for evaluating candidate models as theories of cortical computation.

Prediction, explanation, and levels of analysis

The dissociations observed here bear on a long-standing distinction between predictive adequacy and mechanistic explanation. Models optimized for task performance can achieve strong neural alignment while implementing qualitatively different computational strategies. This separation echoes classical distinctions between levels of analysis in cognitive science, in which multiple algorithmic or mechanistic realizations can satisfy the same computational goal.

Recent advances in deep learning raise the possibility that increasingly flexible and scalable models may eventually discover representations resembling those shaped by biological evolution. However, the present results show that models with comparable predictive accuracy and neural predictivity can diverge systematically in how invariance is constructed, and that these differences become apparent only under targeted tests that probe generalization beyond the training distribution. Thus, even if data-driven optimization can recover brain-like behavior in some regimes, mechanistic convergence cannot be assumed.

From this perspective, biologically grounded models should not be viewed as architectural prescriptions, but as experimental instruments that enable controlled tests of competing mechanistic hypotheses. Scale tolerance provides one such test case. More broadly, these findings underscore the need to complement predictive benchmarks with hypothesis-driven probes if computational models are to move from prediction toward explanation [36].

Behavioral and physiological convergence

The convergence of behavioral and physiological results strengthens the interpretation of scale tolerance as a biologically meaningful diagnostic. Modernized HMAX models reproduce human-like generalization in one-shot discrimination tasks that require extrapolation across scale [1], a regime in which standard convolutional networks fail despite comparable in-distribution accuracy. At the physiological level, intermediate model representations reproduce normalized curvature tuning in macaque area V4 [4], a hallmark signature of scale-invariant shape encoding. The absence of this signature in contemporary convolutional architectures underscores the value of mechanistic tests that go beyond correlational alignment.

Taken together, these findings suggest that scale tolerance provides a particularly informative window into ventral-stream computation. More broadly, they illustrate how targeted probes of invariant coding can reveal differences in representational structure that remain hidden under standard task and neural benchmarks.

Limitations and future directions

Several limitations of the present work point to important directions for future research. First, the models studied here are purely feedforward, whereas biological vision involves extensive recurrent and feedback interactions that contribute to attention, contextual modulation, and temporal integration [35]. Incorporating such dynamics may further refine invariant representations and improve alignment with neural responses. Second, while scale tolerance provides a clear and tractable test case, other forms of invariance—such as tolerance to viewpoint, illumination, and occlusion—remain to be examined within the same framework. Finally, the present study focuses on visual object recognition; extending these approaches to multimodal and task-dependent settings will be necessary to assess the generality of the proposed principles.

Conclusions

By revisiting a biologically grounded ventral-stream model under modern performance constraints, this work demonstrates that predictive success and neural alignment do not uniquely specify underlying computation. The results show that invariant visual representations depend on how tolerance is constructed within a model, not solely on task optimization. More generally, they underscore the importance of targeted mechanistic tests in evaluating computational models as explanations of cortical processing.

References

[1] Han, Y., Roig, G., Geiger, G. & Poggio, T. Scale and translation-invariance for novel objects in human vision. *Sci Rep* **10**, 1411 (2020).

[2] Logothetis, N. K., Pauls, J. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Current Biology* **5**, 552–563 (1995). URL https://doi.org/10.1016/S0960-9822(95)00108-4.

[3] Li, N. & DiCarlo, J. J. Neuronal learning of invariant object representation in the ventral visual stream is not dependent on reward. *Journal of Neuroscience* **32**, 6611–6620 (2012). URL https://www.jneurosci.org/content/32/19/6611.

[4] El-Shamayleh, Y. & Pasupathy, A. Contour curvature as an invariant code for objects in visual area v4. *Journal of Neuroscience* **36**, 5532–5543 (2016). URL https://www.jneurosci.org/content/36/20/5532.

[5] Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology* (1962).

[6] Hubel, D. H. & Wiesel, T. N. Ferrier lecture-functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **198**, 1–59 (1977).

[7] Tootell, R. B., Silverman, M. S., Hamilton, S., Switkes, E. & De Valois, R. Functional anatomy of macaque striate cortex. v. spatial frequency. *Journal of Neuroscience* **8**, 1610–1624 (1988).

[8] Nauhaus, I., Nielsen, K. J., Disney, A. A. & Callaway, E. M. Orthogonal micro-organization of orientation and spatial frequency in primate primary visual cortex. *Nature neuroscience* **15**, 1683–1690 (2012).

[9] Lu, H. D. & Roe, A. W. Optical imaging of contrast response in macaque monkey v1 and v2. *Cerebral Cortex* **17**, 2675–2695 (2007).

[10] Lu, Y. *et al.* Revealing detail along the visual hierarchy: neural clustering preserves acuity from v1 to v4. *Neuron* **98**, 417–428 (2018).

[11] Zhang, Y., Schriver, K. E., Hu, J. M. & Roe, A. W. Spatial frequency representation in v2 and v4 of macaque monkey. *Elife* **12**, e81794 (2023).

[12] Witkin, A. P. *Scale-space filtering*, 1019–1022 (1983).

[13] Koenderink, J. J. The structure of images. *Biological cybernetics* **50**, 363–370 (1984).

[14] Lindeberg, T. *Scale-Space Theory in Computer Vision* (Springer, Berlin, Heidelberg, 1994).

[15] Lindeberg, T. Feature detection with automatic scale selection. *International journal of computer vision* **30**, 79–116 (1998).

[16] Lindeberg, T. A computational theory of visual receptive fields. *Biological cybernetics* **107**, 589–635 (2013).

[17] Logothetis, N. K., Pauls, J. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Current biology* **5**, 552–563 (1995).

[18] Logothetis, N. K. & Sheinberg, D. L. Visual object recognition. *Annual review of neuroscience* **19**, 577–621 (1996).

[19] Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nature neuroscience* **2**, 1019–1025 (1999).

[20] Serre, T. & Riesenhuber, M. Realistic modeling of simple and complex cell tuning in the hmaxmodel, and implications for invariant object recognition in cortex (2004).

[21] Lampl, I., Ferster, D., Poggio, T. & Riesenhuber, M. Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex. *Journal of neurophysiology* **92**, 2704–2713 (2004).

[22] Serre, T. Deep learning: the good, the bad, and the ugly. *Annual review of vision science* **5**, 399–426 (2019).

[23] Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences* **111**, 8619–8624 (2014).

[24] Cadieu, C. F. *et al.* Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology* **10**, e1003963 (2014).

[25] Schrimpf, M. *et al.* Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* **108**, 413–423 (2020).

[26] Goodfellow, I., Lee, H., Le, Q., Saxe, A. & Ng, A. Measuring invariances in deep networks. *Advances in neural information processing systems* **22** (2009).

[27] Azulay, A. & Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research* **20**, 1–25 (2019). URL http://jmlr.org/papers/v20/19-519.html.

[28] Biscione, V. & Bowers, J. S. Convolutional neural networks are not invariant to translation, but they can learn to be. *Journal of Machine Learning Research* **22**, 1–28 (2021).

[29] Graziani, M., Lompech, T., Müller, H. & Andrearczyk, V. *Evaluation and comparison of cnn visual explanations for histopathology*, 8–9 (2021).

[30] Cui, Y. *et al.* Study on representation invariances of cnns and human visual information processing based on data augmentation. *Brain Sciences* **10**, 602 (2020).

[31] Blything, R., Biscione, V., Vankov, I. I., Ludwig, C. J. & Bowers, J. S. The human visual system and cnns can both support robust online translation tolerance following extreme displacements. *Journal of Vision* **21**, 9–9 (2021).

[32] Margalit, E. *et al.* A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron* **112**, 2435–2451 (2024).

[33] Lu, Z. *et al.* End-to-end topographic networks as models of cortical map formation and human visual behaviour. *Nature Human Behaviour* **9**, 1975–1991 (2025).

[34] Blauch, N. M., Behrmann, M. & Plaut, D. C. A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences* **119**, e2112566119 (2022).

[35] DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).

[36] Serre, T. & Pavlick, E. From prediction to understanding: Will ai foundation models transform brain science? *Neuron* **113**, 3504–3508 (2025).

[37] Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* **115**, 211–252 (2015).

[38] Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience* **35**, 13402–13418 (2015).

[39] El-Shamayleh, Y. & Pasupathy, A. Contour curvature as an invariant code for objects in visual area v4. *Journal of Neuroscience* **36**, 5532–5543 (2016).

[40] He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition*, 770–778 (2016).

[41] Araujo, A., Norris, W. & Sim, J. Computing receptive fields of convolutional neural networks. *Distill* **4**, e21 (2019).

Supplementary Information

Beyond neural predictivity: Scale tolerance as a mechanistic test of ventral-stream models

Iván Rodríguez[†], Xizheng Yu[†], Nishka Pant, Jacob Mulliken, Thomas Serre[*]

[1]Department of Cognitive & Psychological Sciences, Brown University, Providence, RI, USA
[†]These authors contributed equally to this work.
[*]Correspondence: thomas_serre@brown.edu

Contents

S1 Supplementary Methods

S1.1 Reproducibility checklist

Code, environment, and hardware

- Code repository URL: Private for now.
- Exact commit hash for experiments reported in this paper: Private for now.
- Framework + version: Python 3.10.15; PyTorch 2.5.1+`cu121`; CUDA 12.1.
- Experiments were run on the Brown Oscar cluster using 8 NVIDIA 24 GB GPUs (A5000 or Quadro RTX-class). CPU and RAM were not separately recorded.
- Total training time per model approximately 8 days with the hardware listed above.
- Random seeds: ImageNet training used seed `42`. Hangul and Pasupathy analysis scripts set seed `1`. Results are from single run rather than averages across multiple seeds.

Datasets and preprocessing

- ImageNet-1k (ILSVRC2012) [37], using the standard training and validation splits. In the code, these are loaded through the `torch/imagenet` dataset interface with split names `train` and `validation`.
- Brain-Score benchmark(s) used `MajajHong2015public.IT-pls`, i.e. macaque IT neural responses from the Majaj and Hong stimulus set [38], evaluated in a scale-generalization setting.
- Hangul stimuli followed the one-shot discrimination paradigm of Han et al. [1], using 27 target/distractor pairs of visually similar characters.
- Pasupathy/El-Shamayleh stimuli followed the normalized-curvature contour set of El-Shamayleh and Pasupathy [39] and these stimuli are loaded as pre-rendered image files.

Input pipeline details (ImageNet).

- Inputs were normalized with ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. For pretrained baseline models, resize/crop settings followed the model-specific `timm` input configuration. For HMAX, training used a fixed input canvas and did not rely on model-specific resize/crop defaults.
- Evaluation used the standard model input size (322×322 for HMAX; 224×224 for the pretrained AlexNet, ResNet-18, and ViT-B/16 baselines), with the same ImageNet normalization.
- Scale evaluation preprocessing followed the ImageNet scale protocol described below (Supplementary S1.1).

Model architectures (exact specification)

HMAX

- Image-pyramid scales were sampled at quarter-octave intervals (multiplicative spacing $2^{1/4}$) around the base input size.

- For the backbone, with input size 322×322 and `ip_scale_bands` $= 3$, the internal image pyramid is scaled to factors of $\{0.708, 0.842, 1.0, 1.189\}$ relative to 322. In addition, during training only, the outer adjacent-scale wrapper samples from

$$\{0.49, 0.59, 0.707, 0.841, 1.0, 1.189, 1.414, 1.681, 2.0\}$$

  and uses adjacent pairs (s_i, s_{i+1}) for the consistency loss; at test time the wrapper uses the unscaled image.
- Channel widths per block/layer (S1/S2/S2b/S3 etc.): S1: $3 \to 48 \to 48 \to 96$. C1 output width: 96 channels per pooled scale-pair feature map. S2: $96 \to 128 \to 256$. C2 output width: 256 channels per pooled scale-pair feature map. S2b (bypass): four branches, each $96 \to 256$, concatenated to 1024 channels; then 1×1 projection $1024 \to 256$. S3: $256 \to 256 \to 384 \to 384 \to 256$. Classifier: $18432 \to 4096 \to 4096 \to 1000$.
- A high-resolution bypass stream branches from C1, is processed through S2b/C2b, and is concatenated with the main stream before classification.
- Kernel sizes, strides, padding, activation, normalization: All convolutional stages follow the standard ResNet basic-block design [40]: each residual block uses two 3×3 convolutions with BatchNorm and ReLU, and a 1×1 projection shortcut when the spatial resolution or channel dimension changes. In our model, S1 contains 3 residual blocks, S2 contains 2 residual blocks, S3 contains 4 residual blocks, and the bypass branch S2b has 4 parallel streams containing 1, 2, 3, and 4 residual blocks, respectively.
- Definition of the scale scorer k_φ (architecture, sharing across layers): The scale scorer maps each feature map to a scalar by adaptive global average pooling to 1×1 followed by a bias-free linear projection,

$$k_\varphi(x) = w^\top \mathrm{AvgPool}_{\mathrm{global}}(x).$$

  Within each pooling layer, the scorer is shared across candidate scales/scale pairs; different layers use separate scorer instances.
- Definition of the "feature adapter" used before adjacent-scale pooling: Before adjacent-scale pooling, feature maps are resized to a common spatial resolution and passed through a two-layer convolutional adapter

$$\mathrm{Conv}(C, C, k_1) \to \mathrm{ReLU} \to \mathrm{Conv}(C, C, k_2) \to \mathrm{ReLU}.$$

- Where (which layers) pairwise scale pooling is applied, and when global pooling is applied: Pairwise adjacent-scale pooling is applied at C1 and C2. In the bypass path, global pooling across scales is applied after S2b (the C2b scorer). In the main path, global pooling is applied after S3. For the default configuration (`ip_scale_bands`=3), this final main path global pooling is standard cross-scale max pooling rather than scorer-based pooling.

Baselines (AlexNet / ResNet-18 / ViT-B/16).

- Source implementation: AlexNet and ResNet-18 were both instantiated through the local `timm` model registry, using standard pretrained weights distributed by the official PyTorch/torchvision model zoo. ViT-B/16 was likewise loaded from `timm` using the standard pretrained `orig_in21k_ft_in1k` variant. No architectural modifications were made to these baseline models.
- Parameter counts reported in Table S1: We report total parameter counts computed as
$$\sum_{p\in\theta} \text{numel}(p),$$
matching the counting used in the codebase.

Training hyperparameters (must be explicit)

Objective and regularization.

- Classification loss: Standard multi-class cross-entropy:
$$\mathcal{L}_{\text{cls}} = \text{CE}(y, \hat{y}).$$
- Hierarchical learnable scale selection and contrastive alignment loss: For scorer-based scale pooling, let $F_\sigma^{(l)}(x)$ denote the feature map at layer l and scale σ, and let
$$s_\sigma^{(l)}(x) = k_\varphi\left(F_\sigma^{(l)}(x)\right), \qquad \alpha_\sigma^{(l)}(x) = \frac{\exp(s_\sigma^{(l)}(x))}{\sum_{\sigma'}\exp(s_{\sigma'}^{(l)}(x))}.$$
The pooled representation is
$$\phi_l(x) = \sum_{\sigma\in\Sigma_l} \alpha_\sigma^{(l)}(x)\, F_\sigma^{(l)}(x).$$
In the HMAX model, this scorer-based pooling is used in intermediate scale-pooling modules, whereas the final main-path global scale pooling is standard cross-scale max pooling.
For adjacent scales (σ_i, σ_{i+1}), the consistency loss is
$$\mathcal{L}_{\text{cl}} = \sum_{l\in\{C1,C2,\text{Out},\text{bypass}\}} \lambda_l \left\|\phi_l(T_{\sigma_i}x) - \phi_l(T_{\sigma_{i+1}}x)\right\|_1,$$
with
$$\lambda_{C1} = 1, \qquad \lambda_{C2} = 1, \qquad \lambda_{\text{Out}} = 0.1, \qquad \lambda_{\text{bypass}} = 1.$$
The full training objective is
$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{cl}}\,\mathcal{L}_{\text{cl}}, \qquad \lambda_{\text{cl}} = 0.1.$$

- Additional regularization: The classifier used dropout with rate 0.5 before the first two fully connected layers. No mixup, cutmix, label smoothing, or EMA were used in this training path.

Optimizer and schedule.

- SGD with momentum 0.9.
- Batch size = 16 per GPU (8 GPUs; effective total batch size = 128); 90 epochs; no gradient clipping.
- Initial learning rate = 0.01; warmup = 0 epochs; step schedule with decay every 30 epochs and multiplicative decay factor 0.1; no explicit minimum learning rate.
- Weight decay 5×10^{-4}.

Augmentation policy (critical for scale-invariance claims).

- Scale jitter range used for *each* model (HMAX and baselines): HMAX was trained with scale range $[1.0, 1.0]$, i.e. no scale jitter. The baseline models (AlexNet, ResNet-18, and ViT-B/16) were used as pretrained models; we therefore refer to their original papers and standard pretraining pipelines for augmentation details, which include scale jittering, color jittering, and a broader range of data augmentations.
- Other augmentations (color jitter, flips, RandAugment, etc.): For HMAX, training used fixed-scale preprocessing (`scale`=[1,1], `ratio`=[1,1]) and horizontal flip ($p = 0.5$); thus, no multi-scale jitter was used. No color jitter, RandAugment/AutoAugment, mixup, cutmix, or random erasing were used. For the pretrained baseline models, augmentations follow their original pretraining setups.
- HMAX was trained without multi-scale jitter; the exact augmentation window was $[1.0, 1.0]$.

Evaluation protocols (exact implementation)

ImageNet scale protocol.

- The seven evaluation scales were $\Sigma = \{160, 192, 227, 270, 322, 382, 454\}$. Each scale denotes the resized length of the smaller image dimension before mapping to the model input size. Images were resized while preserving aspect ratio; if the resized image was smaller than the model input size, it was center-padded with a black background, and if it was larger, it was center-cropped.
- Let s_{ref} denote the reference scale and $\text{Acc}(s)$ the top-1 accuracy at scale $s \in \Sigma$. The off-reference score was defined as

$$\text{Off} - \text{ReferenceScore} = \frac{1}{|\Sigma| - 1} \sum_{s \in \Sigma,\ s \neq s_{\text{ref}}} \text{Accuracy}(s),$$

  i.e., the mean top-1 accuracy across all non-reference scales.
- Evaluation used the standard ImageNet-1k validation split (50,000 images). Results were single-crop/single-view, with one resized-and-centered transform per image and no multi-crop test-time augmentation.

Multi-label ImageNet scale protocol.

- The multi-label ImageNet scale evaluation used the same seven-scale protocol and the same resize/pad/crop procedure as the standard ImageNet scale evaluation. The only difference was the metric: instead of standard top-1 accuracy, we reported mean per-class multi-label accuracy, where a prediction was counted as correct if the model's top-1 class matched any valid label for the image, and accuracies were then averaged over classes.

Hangul one-shot protocol.

- For mapping degrees to pixels in our model stimuli, we used a conversion of 26 pixels $= 1^\circ$ of visual angle. In the evaluation, character sizes of 13, 52, and 130 pixels were used to instantiate the $30'$, 2°, and 5° conditions, respectively, which corresponds to an conversion of 26 pixels per degree of visual angle.
- The five evaluated size conditions were $30'/2^\circ$, $2^\circ/30'$, $30'/5^\circ$, $5^\circ/30'$, and $30'/30'$. The $30'/30'$ condition was treated as in-distribution, and the remaining four conditions were averaged as cross-scale out-of-distribution performance.
- Features were extracted from an internal model layer and compared after flattening, with no additional feature normalization. Similarity was measured by the Pearson correlation between flattened activation vectors. For layers that returned multiple feature maps (e.g., multi-scale HMAX outputs), the implementation used the first feature map only. In the final postprocessed analyses, the best resulted layers selected were `model_backbone.s2` for HMAX, `features.5` for AlexNet, `layer3.1.conv1` for ResNet-18, and `blocks.7.norm2` for ViT-B/16.
- The Pearson-correlation threshold was determined separately for each model, layer, and size condition. The code recreates 1000 random resampling splits; for each split, a threshold is chosen from the observed correlation values to maximize a threshold-based discrimination criterion, and performance is then summarized across the 1000 splits.
- For each condition, hit rate and false-alarm rate were computed from thresholded Pearson correlations over 27 same-identity pairs and 27 distractor pairs, respectively. We report a d′-like discrimination score,

$$d'_{\text{proxy}} = \text{HitRate} - \text{FalseAlarmRate},$$

  which serves as a threshold-based sensitivity measure for this one-shot matching setup.

Brain-Score protocol.

- The benchmark used was `MajajHong2015public.IT-pls`, i.e. macaque IT neural responses from the Majaj and Hong stimulus set [38], evaluated in a scale-generalization setting.
- Neural predictivity was computed with partial least squares regression using 25 components and no additional scaling or regularization. The implementation used a fixed train/test split rather than internal cross-validation.
- The evaluated model layers were fixed a priori as IT-like candidate layers.

- Scale generalization was evaluated over six stimulus-scale bins,

$$\{(0.75, 0.85),\ (0.85, 0.95),\ (1, 1),\ (1.05, 1.15),\ (1.15, 1.25),\ (1.25, 1.35)\}.$$

  The reference bin was the unit-scale bin $(1, 1)$. For each model and layer, the linear mapping was fit on the reference-scale training set and tested separately on each scale bin in the held-out test set.

Pasupathy / V4 normalized-curvature replication.

- Stimuli were evaluated at relative sizes 0.4×, 0.6×, 0.8×, and 1.0×. For each analyzed layer, images were center-cropped to the analytically computed receptive-field (RF) size of the central unit and then padded back to the full model input size before scale manipulation. Receptive-field (RF) sizes were computed analytically for each model layer using the standard recursive convolution/pooling formulas with an upper bound clipped to the model input size [41]. Exact layer-wise RF sizes for all analyzed models are reported in Table S2, S3, S4. For ViT-B/16, the patch embedding layer has a local receptive field of 16 px, but because each transformer block performs global self-attention over all patch tokens, all subsequent block outputs were assigned full-image support (224 px).
- Responsive units were selected by Z-score filtering. For each neuron, responses were standardized across the full Pasupathy stimulus set, and the neuron was included if its maximum Z-scored response satisfied $Z \geq 2.0$.
- For the V4 comparison, scale-invariance scores were binned into 11 bins of width 0.4, with edges

$$\{-2.2, -1.8, -1.4, -1.0, -0.6, -0.2, 0.2, 0.6, 1.0, 1.4, 1.8, 2.2\}.$$

  V4 counts were obtained by scaling the published bar heights to $N = 80$. Group differences were assessed with a standard χ^2 contingency test after excluding bins with zero counts in both groups, and with a Monte Carlo independence test using 20,000 simulations under a pooled multinomial null.
- Scale invariance score definition: For each responsive neuron, a linear slope was fit to its mean response as a function of stimulus scale at the neuron's preferred orientation; the reported layer-level score is then

$$\text{InvScore} = \text{median}_i\big(|\beta_i|\big),$$

  where β_i is the fitted scale slope for neuron i. Lower values indicate greater scale invariance.

S2 Supplementary Results

S2.1 Supplementary tables

Model	**Parameters (M)**
HMAX	128.9
AlexNet	61.1
ResNet-18	11.69
ViT-B/16	86.56

Table S1: Comparison of model parameters.

Table S2: Analytical receptive field sizes and effective strides for AlexNet.

Layer	RF size	Effective stride
features.0	11	4
features.2	19	8
features.3	51	8
features.5	67	16
features.6	99	16
features.8	131	16
features.10	163	16
features.12	195	32
features	195	32
classifier	195	32

Table S3: Analytical receptive field sizes and effective strides for ResNet-18.

Layer	RF size	Effective stride
conv1	7	2
maxpool	11	4
layer1	43	4
layer2	99	8
layer3	211	16
layer4	224	32
global_pool	224	32

Table S4: Analytical receptive field sizes and effective strides for HMAX.

Layer	RF size	Effective stride
s1	47	16
c1.pool1	79	32
c1.pool2	95	48
c1	79	32
s2	322	32
c2.pool1	322	64
c2.pool2	322	64
c2	322	64
s2b.s2b_seqs.0	207	32
s2b	322	32
c2b_score.pool1	322	64
c2b_score.pool2	322	96
c2b_score	322	64
c2b_seq	322	64
s3	322	64
global_pool.pool1	322	128
global_pool.pool2	322	192
global_pool	322	128

S2.2 Supplementary figures

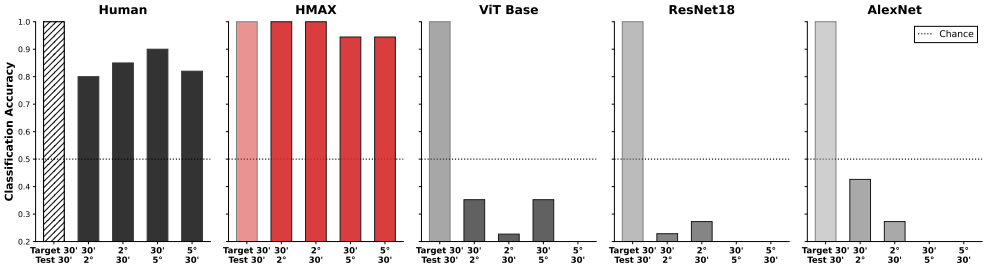


Fig. S1: Hangul one-shot learning sensitivity analysis.

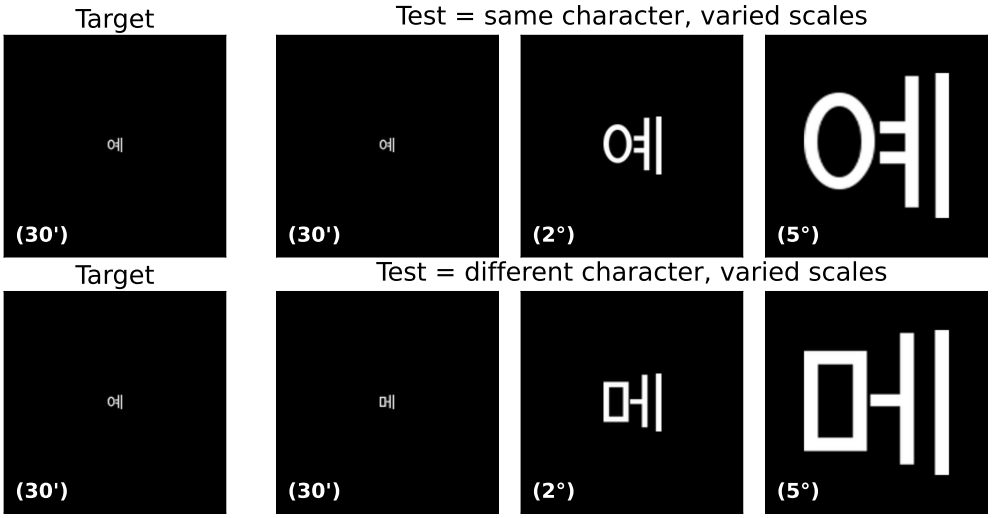


Fig. S2: Example Hangul one-shot stimuli used in the evaluation. The left column shows the target character, and the remaining columns show test characters at the three evaluated sizes: 30′, 2°, and 5°. The top row illustrates same-identity matches across scale, and the bottom row illustrates different-identity comparisons across scale.

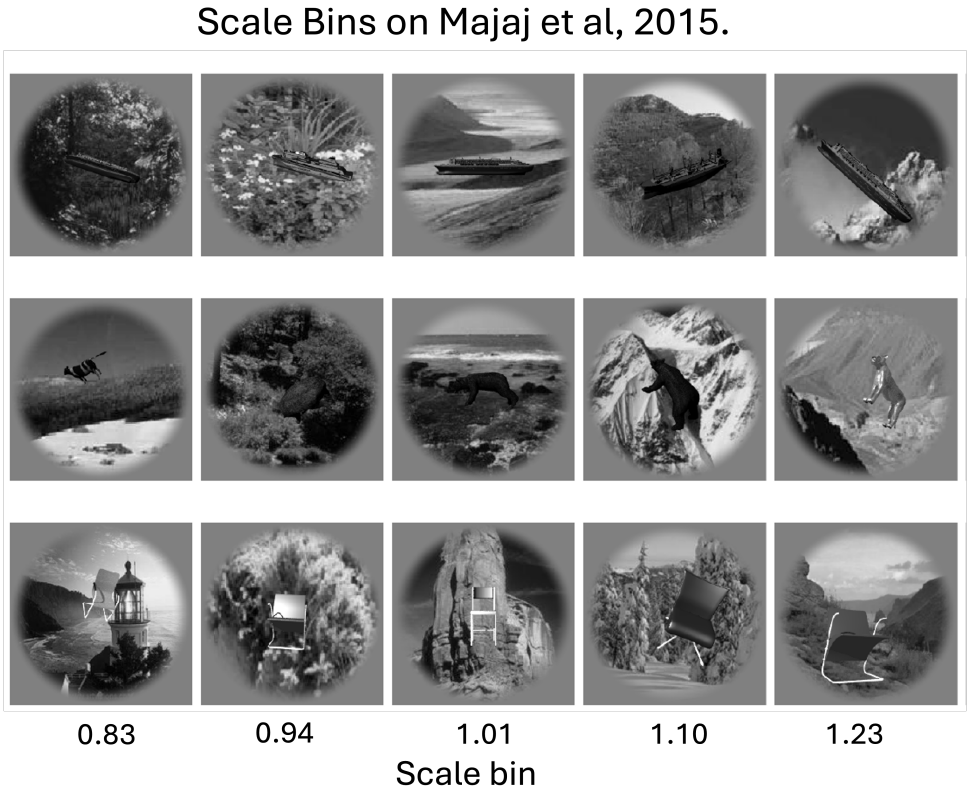


Fig. S3: Example stimuli from the Brain-Score benchmark dataset.