Hierarchical Emergence Profiles of Human-Derived Dimensions are a Fundamental Property of Deep Neural Networks

Florian Burger (F.Burger@westernsydney.edu.au)

The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia

Manuel Varlet (M.Varlet@westernsydney.edu.au)

The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia

School of Psychology, Western Sydney University, Sydney, Australia

Genevieve L. Quek (G.Quek@westernsydney.edu.au)

The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia

Tijl Grootswagers (T.Grootswagers@westernsydney.edu.au)

The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, Australia

School of Computer, Data and Mathematical Sciences, Western Sydney University, Sydney, Australia

Corresponding Author: Florian Burger

Abstract

Object recognition in the human visual system is implemented within a hierarchy characterised by increasing feature complexity. Here, we investigated whether human-derived dimensions of object knowledge show a similar progressive emergence across layers in deep neural networks (DNNs), and how this emergence is shaped by architecture, learning objective, and stimulus statistics. To test this, we predicted human-derived dimensions from layer-wise activations of multiple DNNs and transformer models trained on large-scale datasets. Results showed that trained DNNs exhibit emergence profiles resembling theoretical expectations from human vision, with behaviourally relevant object dimensions largely absent in early layers, strengthening across layers, and peaking in later layers. Architectural mechanisms such as recurrence and skip connections amplified this encoding, learning objectives redistributed information across layers, and changes in stimulus statistics confirm that hierarchical emergence is a general principle extending to material perception. These findings demonstrate that the hierarchical emergence of human-derived dimensions is a fundamental property of trained networks and highlight design and input factors that shape layer-wise representational organisation, providing hypotheses for the structure of visual representations in the brain.

Introduction

Recognising visual objects rapidly from retinal input is critical for everyday behaviour. The human visual system achieves this through a spatial–temporal hierarchy: simple features are extracted early in both space and time, while complex, abstract features emerge comparatively later. Early visual cortex such as V1 encodes local edges and orientations, whereas higher-level areas such as inferotemporal cortex (IT) represent complex object properties and categories (DiCarlo et al., 2012; Hubel & Wiesel, 1962). Time-resolved EEG and MEG suggest that early visual responses to objects are strongly driven by low-level image characteristics such as texture, with information about these low-level visual features peaking roughly 50–100ms after stimulus onset. Conversely, information about higher-level object properties such as animacy arises later during stimulus processing, peaking at around 150–200ms (Carlson et al., 2013; Robinson et al., 2023). Both low-level and high-level properties are relevant for behaviour, with low-level properties supporting rapid perceptual discrimination and high-level properties guiding conceptual understanding and goal-directed behaviour (Groen et al., 2018; Thorat et al., 2019). These behavioural relevance patterns can be captured in human-derived dimensions that span from perceptual features such as colour to conceptual features such as animacy, together providing a high-dimensional description of representational content (Hebart et al., 2020; Schmidt et al., 2025). In the brain, these human-derived dimensions emerge in a systematic progression from early low-level dimensions to later high-level dimensions, reflecting the gradual transformation of visual information along the hierarchy (Contier et al., 2024; Teichmann et al., 2023).

Deep neural networks (DNNs) process information through a series of layers, with early layers aligning to early stages of visual processing in humans and later layers aligning to later stages in humans (Cichy et al., 2016; Guclu & Van Gerven, 2015; Yamins et al., 2014). In terms of human-derived dimensions, DNNs show a measurable overlap with human behaviour: Mahner et al. (2025) decompose DNN similarity structure into latent visual and semantic dimensions that align with human judgments. Furthermore, DNN activations contain decodable information about specific human-derived dimensions (Kaniuth et al., 2024; Hebart et al., 2020). However, it is not known whether these human-derived dimensions follow a hierarchical emergence profile across layers that resembles the temporal and spatial progression observed in the human visual system. In this study, we map the emergence profile of human-derived object dimensions, tracing the trajectory of each dimension's encoding strength across layers to reveal where each dimension becomes represented in the model. Mapping these emergence profiles allows direct comparison with theoretical expectations from human vision.

To further understand what might drive emergence profiles, we draw inspiration from human vision, where factors such as architecture, learning objective, and stimulus statistics are known to influence representational organisation. Architectural mechanisms in the brain such as recurrence and feedback refine responses in higher visual areas and shape the progression of representations across visual processing stages (Kar et al., 2019; Spoerer et al., 2020). Learning objective further modulates processing: neurons in the visual cortex can adjust their tuning depending on behavioural goals, reflecting the influence of top-down objectives (Bracci et al., 2017; Harel et al., 2014). The statistics of object inputs also matter, with different object domains accentuating different human-derived dimensions of visual information, such that object-focused datasets emphasise higher-level properties while material- and texture-focused datasets highlight perceptual features (Konkle & Caramazza, 2013; Hebart et al., 2020; Schmidt et al., 2025). Yet the relative contribution of architecture, task, and stimulus statistics to the formation of visual hierarchies remains unclear. DNNs provide a powerful testbed to examine these influences in isolation (Kriegeskorte, 2015). By comparing emergence profiles across systematic manipulations of network architecture, learning objective, and stimulus statistics in DNNs, we can probe how these mechanisms shape hierarchical representations. Although these manipulations are artificial, they provide a principled way to anticipate how similar factors might influence the emergence of human-derived dimensions in human vision.

In this study, we extracted layer activations from DNNs while they processed a large set of natural images, and used regression models to predict human-derived dimensions from these activations. This layer-wise prediction produced an emergence profile for each dimension defined as the trajectory of each dimension's encoding strength across layers. We then examined how these emergence profiles are shaped by three factors: network architecture, learning objective, and stimulus statistics. To anticipate our results, we found that high-level dimensions such as animacy were weak or absent in early layers, emerged in mid-layers, and peaked in later layers. Recurrence and skip connections increased both the magnitude and persistence of dimension encoding. Semantic supervision shifted dimension-relevant information toward deeper layers, whereas self-supervised learning enhanced the representation of perceptual dimensions. Finally, changes in stimulus statistics reweighted the hierarchy toward earlier encoding. By treating emergence profiles as the central unit of analysis, we link what is represented (human-derived dimensions) with how it develops (architecture) and why it is maintained (learning objective and stimulus statistics), providing a layer-resolved account of how human-derived dimensions arise and change within artificial systems.

Results

To examine whether deep neural networks reproduce the hierarchical emergence of human-derived dimensions observed in human vision and behaviour, we systematically analysed their layer-wise representations in a variety of Deep Neural Networks (DNNs) from feedforward models to recent transformer models. We used images from two datasets: the THINGS object dataset (Hebart et al., 2019), which contains a broad range of everyday objects, and the STUFF material perception dataset (Schmidt et al., 2025), which focuses on natural and man-made materials (Figure 4A). In comparison, THINGS emphasises high-level human-derived object dimensions such as animacy and tool-relatedness, whereas STUFF highlights perceptual human-derived material dimensions such as colour and texture (Hebart et al., 2020; Schmidt et al., 2025). For each model and dataset, we extracted activations from every layer, applied principal component analysis (PCA) at each layer to reduce dimensionality, and used linear regression to predict the relevant human-derived dimensions from these components, ordered by how much variance in human behaviour they explain (Figure 1A, see Methods).

To establish whether the hierarchical emergence of human-derived object dimensions depends on learning or is already present in the input or network architecture, we first compared three baseline cases: raw image pixels, an untrained DNN, and a trained DNN. This initial analysis tests the most fundamental question whether a hierarchy of human-derived dimensions can emerge without learning, or whether training is required to align model representations with human vision. As the DNN, we used AlexNet, a widely used architecture in cognitive science (Figure 1B; Krizhevsky et al., 2012). Decoding from raw pixels showed that some perceptual dimensions, such as “Red,” which captures variation in perceived redness across objects, are already reflected in stimulus statistics, but no high-level structure as shown by absence of prediction strength for high-level dimensions, such as “Electronic / Technology” (Figure 1C-D). The untrained model showed slightly greater sensitivity to higher-level dimensions, reflecting architectural biases, but remained dominated by perceptual features with an absence of emergence profiles (Figure 1C-D). In contrast, the trained model exhibited a clear and robust hierarchy: human-derived object dimensions were weak or absent in the earliest layers, increased in mid-layers, and peaked toward the top of the network (Figure 1C-D). Thus, unlike raw pixels or an untrained network, a trained DNN displays the expected hierarchical emergence of human-derived dimensions.

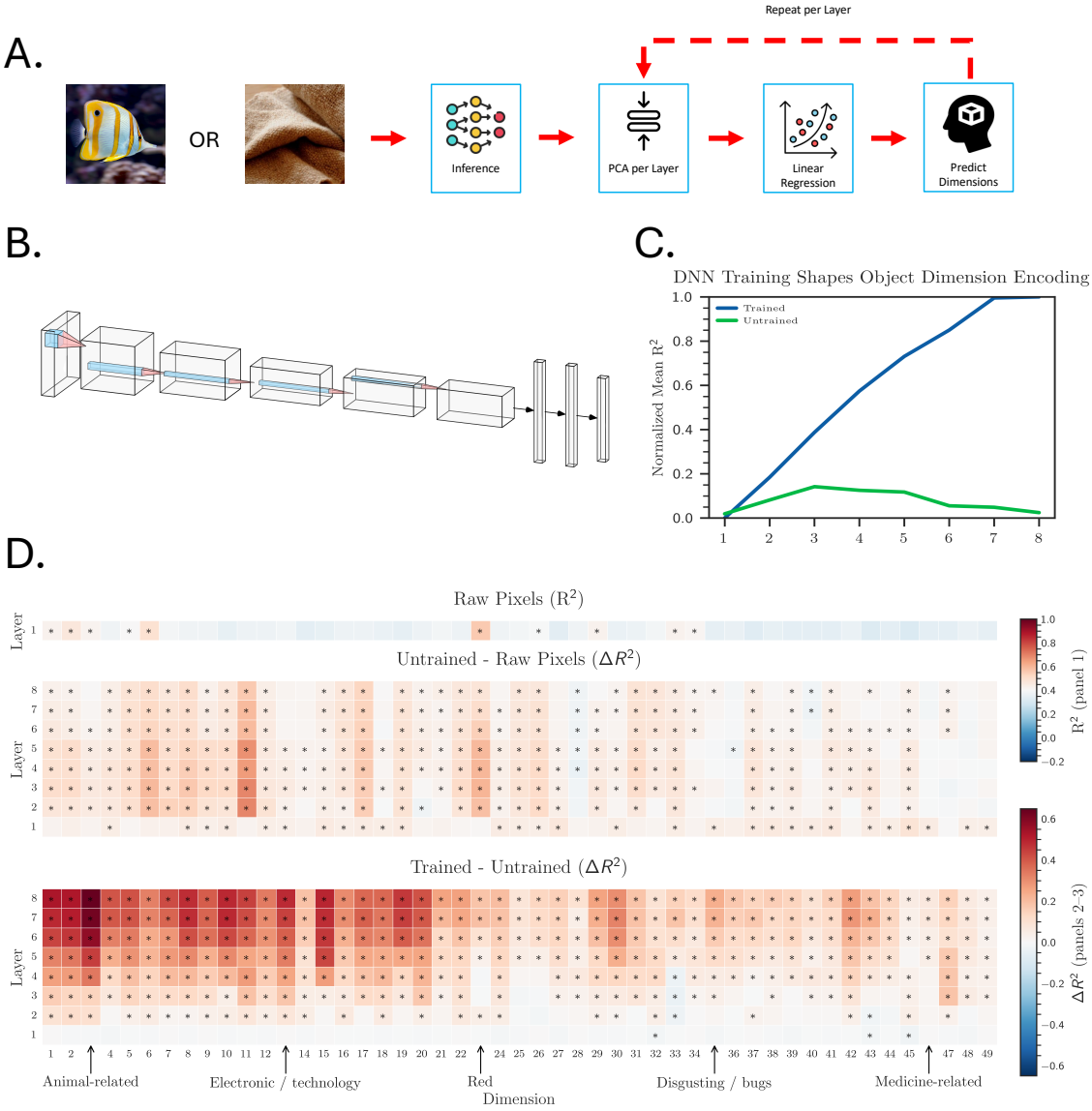

Figure 1. Predicting human-derived dimensions from images and DNN activations. **A.** Workflow illustration for both THINGS and STUFF datasets: images are passed through the network, PCA is applied to each layer, and linear regression is used to predict human-derived dimensions (repeated per layer). **B.** Schematic architecture of AlexNet. **C.** Mean prediction accuracy (normalized R^2) across layers for trained (blue) and untrained (green) AlexNet, showing that training induces a gradual, hierarchical increase in dimension encoding while untrained remains flat/consistent. **D.** Heatmaps showing prediction accuracy (R^2) and model differences across human-derived object dimensions. Top row: baseline performance from raw pixel values. Middle row: difference between untrained model and raw pixels. Bottom row: difference between trained and untrained model. Colour indicates strength/difference of prediction accuracy (R^2), with asterisks marking permutation-tested significance (FDR-corrected; see Methods).

Architecture: Human-Derived Dimensions Emerge Across All Models and Strengthen with Brain-Like Mechanisms

Next, we investigated whether architectural choices influence the emergence profile as different architectures fundamentally alter how information flows through a network (Kar et al., 2019; Spoerer et al., 2020). To test this, we analysed three architectures from the CORnet family (Kubilius et al., 2018), which share an identical four-layer layout but differ in their inclusion of recurrence and skip connections (Figure 2A). Across all three architectures, we found clear evidence that human-derived object dimensions are present and follow the expected hierarchical ordering: most dimensions were weak or absent in early layers, increased in mid-layers, and peaked toward the top of the hierarchy (Figure 2B).

Relative to the feedforward model (CORnet-Z), the recurrent model (CORnet-RT) exhibited stronger encoding of human-derived object dimensions in the final layer (Figure 2B-C). While the overall emergence profile shape was similar, recurrence amplified the magnitude of dimension-specific representations, suggesting iterative refinement of individual dimensions (Figure 2B-C). The third model, which combines recurrence with skip connections (CORnet-S), produced the strongest emergence profiles, with dimension encoding rising earlier and reaching higher peak values than in the other models (Figure 2B-C). This pattern suggests that skip connections accelerate the buildup of representational strength across the hierarchy.

Together, these results show that behaviourally relevant human-derived object dimensions emerge in DNNs in the expected hierarchical manner, and that the inclusion of brain-inspired mechanisms within these models serves to systematically strengthen this hierarchical representation, bringing model dynamics closer to those observed in human vision.

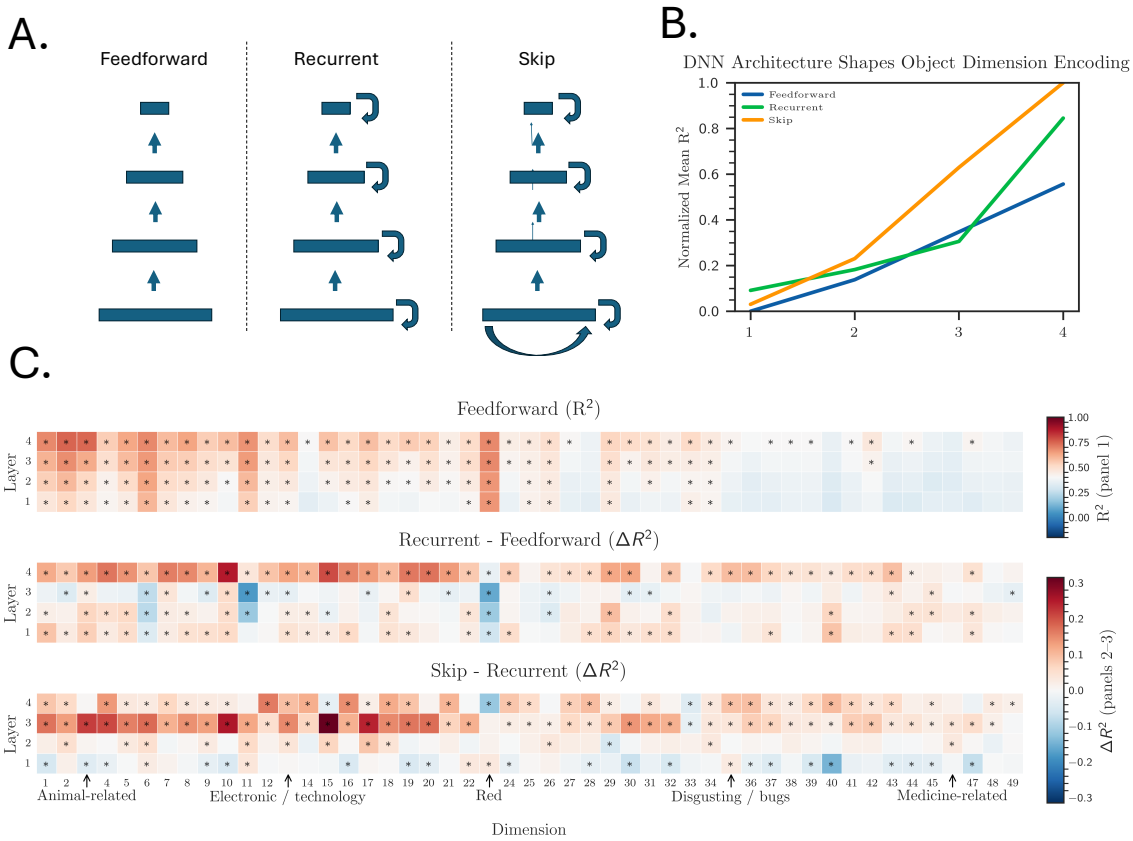

Figure 2. Comparing feedforward, recurrent, and skip architectures in predicting human-derived object dimensions. **A.** Schematic illustrations of three model types: feedforward (no recurrence or skip connections), recurrent (within-layer recurrence), and skip (within-layer long-range skip connections). **B.** Normalized mean prediction accuracy (R^2) across visual areas (V1–IT) for each architecture, showing enhanced encoding of human-derived object dimensions in recurrent and skip networks relative to feedforward. **C.** Heatmaps showing prediction accuracy and architectural differences across the 49 dimensions. Top: baseline accuracy (R^2) for the feedforward model. Middle: difference in prediction accuracy (ΔR^2) between recurrent and feedforward architectures. Bottom: difference between skip and recurrent architectures. Colour indicates strength/difference of prediction accuracy (R^2), with asterisks marking permutation-tested significance (FDR-corrected; see Methods).

Learning Objective: Dimension Emergence Varies with Learning Objectives

Next, we examined how learning objectives influence the emergence of human-derived object dimensions in DNN layers, motivated by evidence that task demands modulate representational structure in biological vision and DNNs (Bracci et al., 2017; Zhuang et al., 2021). To do this, we compared models with identical architectures but different learning objective (Figure 3A): a vision transformer (ViT; Dosovitskiy et al., 2021), trained on a purely visual classification objective; the vision branch of CLIP (Radford et al., 2021), trained jointly to classify images and predict their natural

language descriptions; and a self-supervised vision model known as DINOv3 (Siméoni et al., 2025).

Across all three models, we found clear evidence that human-derived object dimensions are differentially represented across layers (Figure 3B-C). In the vision model (ViT), emergence profiles rose in early and mid-layers but declined toward the final layer, which functions as a classification head and compresses information into discrete category representations (Figure 3B-C). In contrast, the vision + language model (CLIP) showed a steady increase in encoding strength across the hierarchy, peaking in the final layers that map visual features into a shared image–text embedding space (Figure 3B-C). The self-supervised model (DINOv3) followed a similar trajectory to the vision–language model, with strong increases across early and mid-layers, but encoding declined again in the final layer, which produces invariant embeddings optimised for self-distillation rather than preserving fine-grained visual or semantic detail (Figure 3B-C).

Together, these results demonstrate that hierarchical emergence of human-derived object dimensions is a shared property across visual models, but the trajectory and distribution of this emergence depend on the model's learning objective. Different learning objectives emphasise distinct types of information: categorical, conceptual, or invariant. These distinct types of information lead to systematic differences in how human-derived object dimensions develop across the hierarchy.

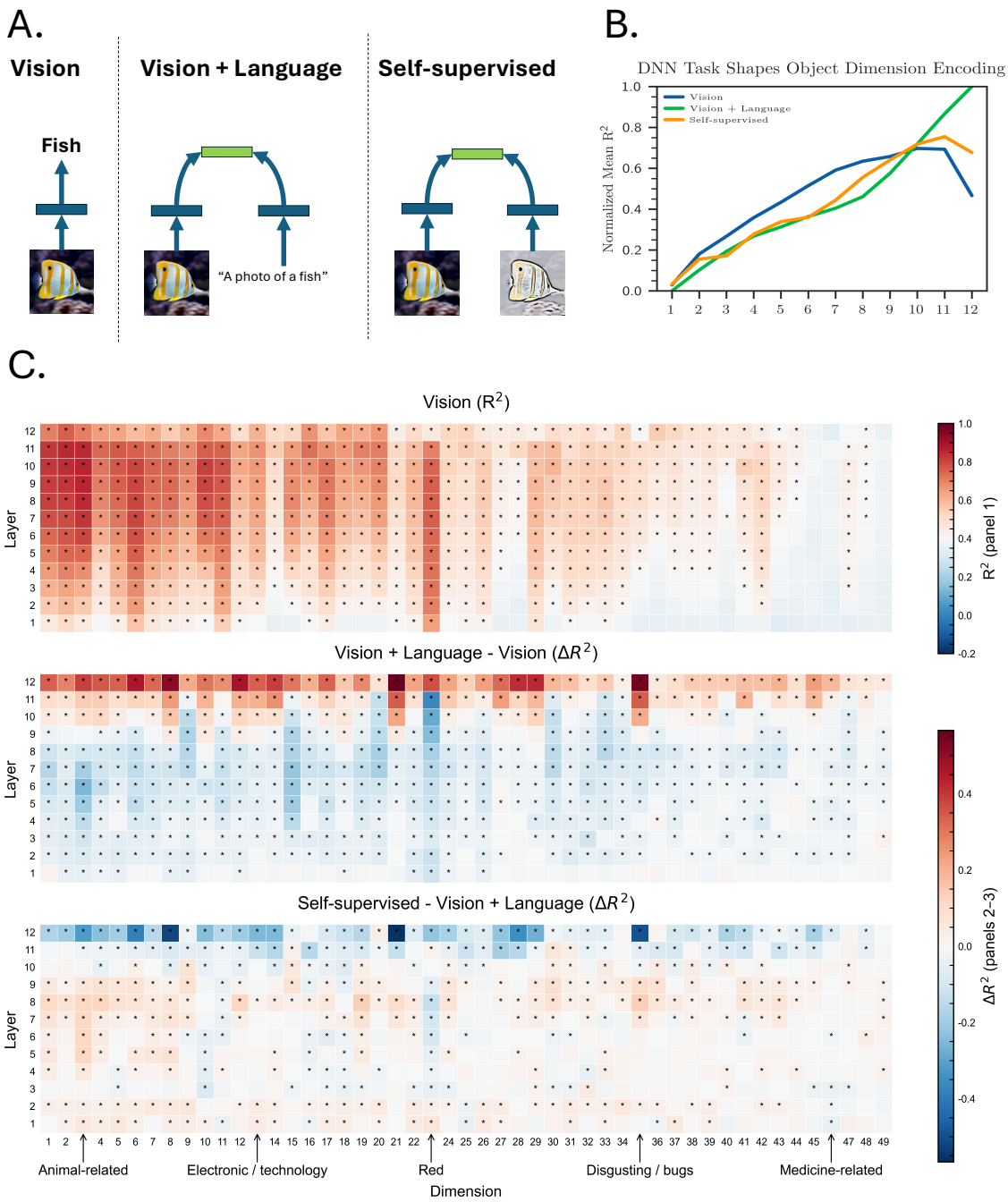

Figure 3. Comparing learning objectives (vision, vision–language, unsupervised) in predicting human-derived object dimensions. **A.** Schematic illustrations of three training paradigms: vision-only (standard supervised image classification), vision + language (image–text contrastive training, e.g., "a photo of a fish"), and self-supervised vision (self-supervised contrastive learning). **B.** Normalized mean prediction accuracy (R^2) across layers for the three architectures. CLIP (vision + language; orange) achieves higher prediction accuracy in later layers compared to supervised ViT (blue) and self-supervised DINOv3 (green). **C.** Heatmaps showing prediction accuracy and learning objective driven differences across the 49 dimensions. Top: baseline accuracy (R^2) for supervised ViT. Middle: difference in prediction accuracy (ΔR^2) between CLIP and ViT.

Bottom: difference between DINOv3 and CLIP. Colour indicates effect size, with asterisks marking permutation-tested significance (FDR-corrected; see Methods).

Stimulus Statistics: Human-Derived Dimensions Generalise Across Domains

Having established that human-derived object dimensions systematically emerge across DNN hierarchies, we next asked whether these same representational principles extend to a different perceptual domain: materials. Whereas objects primarily vary in terms of high-level dimensions such as animacy and tool-relatedness, materials are primarily defined by mid-level perceptual dimensions such as texture and geometry (Figure 4A). We therefore repeated our analyses on the STUFF material perception dataset (Schmidt et al., 2025), which focuses on natural and man-made materials (Figure 4A), allowing us to test whether hierarchical emergence reflects a general principle rather than a phenomenon limited to object-centric datasets.

Across models and architectures, we found that human-derived material dimensions were robustly encoded and followed a clear hierarchical progression: weak in early layers, strengthening in mid-layers, and reaching maximal expression toward the top of the network (Figure 4B–D). These results demonstrate that the same hierarchical organisation can be extended to a different stimulus domain.

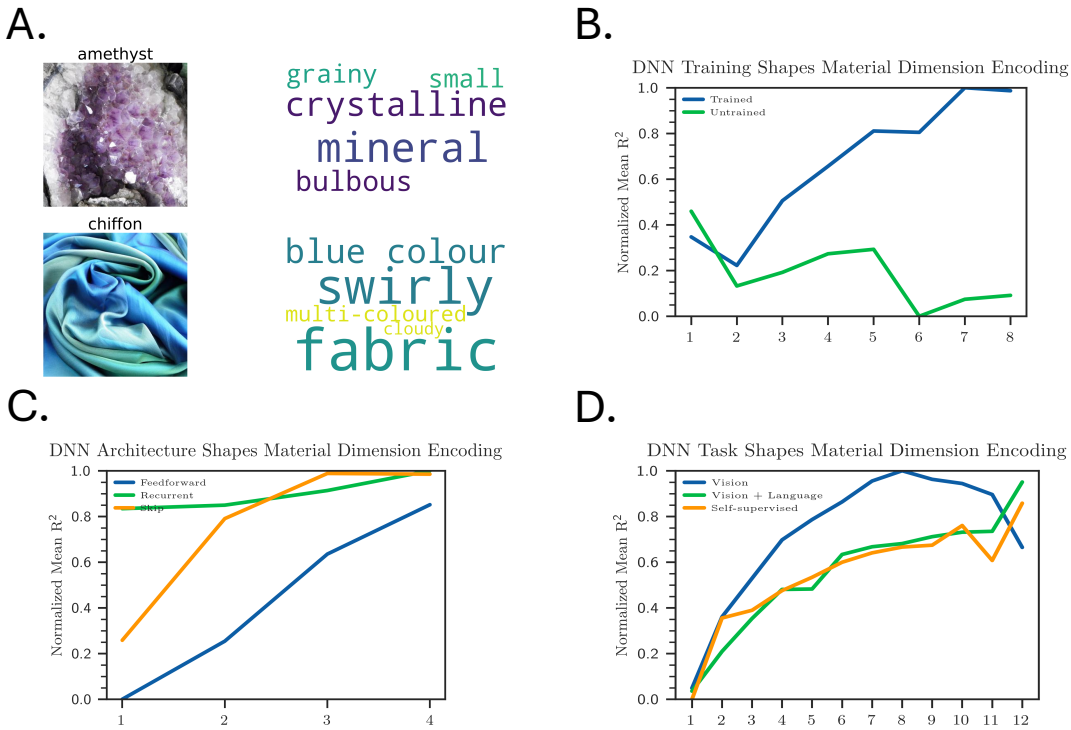

Figure 4. Predicting human-derived material dimensions across architectures and learning objectives. **A**. Example STUFF dataset images (amethyst, chiffon) with associated top-ranked human-derived material dimensions (e.g., mineral, crystalline, fabric, swirly). **B**. Normalized mean prediction accuracy (R^2) across AlexNet layers,

comparing pretrained (blue) and untrained (green) models, showing that training substantially improves material dimension encoding. **C**. Comparison of network architectures (feedforward, recurrent, skip) on mean prediction accuracy (R^2) across layers, indicating stronger encoding in recurrent and skip models relative to purely feedforward. **D**. Comparison of learning objectives (vision-only, vision + language, self-supervised vision) for ViT-based models. All models show robust material dimension encoding, with vision-only slightly outperforming multimodal and self-supervised objectives in later layers.

Discussion

In this study, we investigated whether DNNs recapitulate a hierarchy of human-derived dimensions, and clarify how architecture, learning objective, and stimulus statistics shape the emergence of these dimensions across DNN layers. We found clear evidence that trained networks produce emergence profiles, with information about human-derived dimensions largely absent in early layers, strengthening with progression along the hierarchy, and peaking in the final layers. This progression parallels the organisation observed in human vision, suggesting that DNNs can reproduce the gradual build-up of dimension information across levels of processing. Beyond this robust hierarchy, the factors we tested systematically influenced its expression: brain-inspired mechanisms such as recurrence and skip connections amplified and accelerated dimension encoding, learning objectives reshaped the distribution of dimension information across the hierarchy, and stimulus statistics modulated where human-derived dimensions became most strongly represented.

Our findings demonstrate that the hierarchical emergence of human-derived dimensions in DNNs is not simply given by stimulus statistics or architectural depth alone, but arises specifically through network training. This mirrors the progression observed in human vision, where simple features appear earlier and more complex information builds up later (Carlson et al., 2013; DiCarlo et al., 2012; Robinson et al., 2023). Decoding from raw pixels revealed that some structure was already present in the stimulus statistics as some human-derived dimensions were already decodable. Untrained networks showed slightly stronger dimension encoding, likely reflecting architectural biases, but still lacked progressive representational encoding of human-derived dimensions across ascending layers. Only trained models produced clear emergence profiles in which human-derived dimensions gradually strengthened across layers and peaked in the final stages. Previous work has shown that DNNs capture human-derived dimensions (Mahner et al., 2025) and broadly align with the human visual hierarchy (Cichy et al., 2016), but these studies largely treated dimension presence as static. By introducing *emergence profiles*, we extend this work to show not just that human-derived dimensions are represented in DNNs, but also where in the networks they become strongly encoded, linking human findings (Contier et al., 2024;

Teichmann et al., 2023) with model layer progression. One caveat is that the prevalence of individual human-derived dimensions varies across the dataset, with more frequent human-derived dimensions appearing to drive stronger emergence profiles. While this distribution partly reflects the variance explained in human behaviour (Hebart et al., 2020), it raises the possibility that the observed hierarchy may be amplified for human-derived dimensions that are more common in the stimulus set.

Our results further show that the emergence of human-derived dimensions in DNNs is shaped by architecture, learning objective, and stimulus statistics. Comparing models that differed in specific architectural mechanisms revealed that additions such as recurrence and skip connections systematically strengthened and prolonged human-derived dimension encoding, indicating that these mechanisms refine how human-derived dimensions are built up across layers. Varying the learning objective also influenced the profiles: supervised vision–language models shifted human-derived dimension information toward deeper layers, whereas vision-only or self-supervised objectives produced trajectories declining in the final layers. Stimulus statistics provided a third source of modulation, with object-rich datasets accentuating later-emerging human-derived dimensions and material-focused datasets shifting the hierarchy toward earlier encoding. While these three factors capture key influences, our results show a small fraction of variations possible within each of these factors. These results demonstrate that emergence profiles are not fixed properties of a network, but dynamic trajectories that depend on design choices and stimulus statistics. By showing how information about human-derived dimensions in DNNs change under different conditions, this approach not only clarifies the mechanisms that sculpt hierarchies in DNNs but also provides hypotheses for how comparable factors may shape the emergence of human-derived dimensions in the brain. Testing these predictions with neural data, such as fMRI or EEG, could reveal not only where models and the brain converge, but also highlight critical differences in how human-derived dimensions are encoded.

In conclusion, this study shows that DNNs exhibit emergence profiles for human-derived dimensions that closely parallel those observed in the visual system. By introducing emergence profiles, we show that hierarchies of human-derived dimensions arise reliably in trained networks and that their shape is systematically governed by architecture, learning objectives, and stimulus statistics. Together, these findings establish hierarchical emergence as a fundamental property of trained networks and demonstrate that its organising principles mirror those of human vision.

Methods

All code for this work is available under https://github.com/Flo-Burger/Hierarchical_Emergence_Profiles_of_Human_Derived_Dimensions_are_a_Fundamental_Property_of_DNNs. The THINGS dataset is available under https://things-initiative.org. To access the STUFF dimensions, please see https://osf.io/5gr73/.

Datasets

We used two large-scale behavioural datasets providing human-derived dimensions. The THINGS database (Hebart et al., 2019, 2020) contains 26,107 images from 1,852 everyday object concepts. Human similarity judgments were modelled which yielded 49 human-derived object dimensions that predominantly capture high-level properties such as animacy, tool-relatedness, and category structure. Dimension weights were established only for the first exemplar of each concept, meaning that 1,852 images with associated dimensions were available for our analyses. In contrast, the STUFF database (Schmidt et al., 2025) contains 600 images from 200 natural and artificial material categories, such as stone, wood, and fabric. Using the same behavioural–modelling framework, 36 human-derived material dimensions were derived that primarily capture perceptual properties such as colour, texture, and surface appearance. In this dataset, human-derived material dimensions were established for all images, allowing us to include the entire set in our analyses. For both datasets, all images were resized to 224 × 224 pixels and normalized to ImageNet mean and variance before being presented to the networks.

Models

To examine how architectural mechanisms and learning objectives shape the hierarchical emergence of human-derived dimensions, we tested a range of deep neural network models differing in structure and learning paradigm. For each model, activations were extracted layer by layer (or block by block for transformer models), allowing us to track the trajectory of human-derived dimension encoding across the representational hierarchy. All models were implemented in PyTorch (v1.13), with dimensionality reduction and regression carried out using scikit-learn (v1.4) and NumPy (v1.26). All analyses were run locally on a MacBook Pro 2024 with a M4 Pro chip.

As a baseline, we included raw image pixels as a control, testing whether human-derived dimensions could be decoded directly from stimulus statistics. We then analysed AlexNet (Krizhevsky et al., 2012), a widely studied convolutional network in cognitive science. By comparing AlexNet with random weights (untrained) against its pretrained counterpart (trained on ImageNet 1k classification), we could test whether hierarchical emergence arises from architectural biases or training.

To assess the influence of architectural mechanisms, we used the CORnet family of models (Kubilius et al., 2018), which replicate a four-layer cortical hierarchy with systematic variations. CORnet-Z provides a purely feedforward baseline, CORnet-RT incorporates within-area recurrence, and CORnet-S combines recurrence with long-range skip connections. These models allowed us to test how recurrence and skip connections affect the strength and persistence of emergence profiles.

Finally, to investigate the role of learning objectives, we analysed three transformer-based models with identical base architectures but different optimisation goals. We included ViT-Base (Dosovitskiy et al., 2021), trained on supervised image classification; the vision branch of CLIP (Radford et al., 2021), trained on contrastive image–text pairs; and DINOv3 (Siméoni et al., 2025), trained using self-supervised contrastive learning. For these models, we extracted the CLS token representation from each transformer block. These models allowed us to test how different learning objectives influence the emergence and distribution of human-derived dimensions across the representational hierarchy.

Feature Extraction and Dimension Prediction

For each dataset–model combination, we extracted activations from every layer (all convolutional and fully connected layers for DNNs/CNNs, and the CLS token from each transformer block). These activations were reduced in dimensionality using principal component analysis (PCA) for each layer. For each layer, the PCA-transformed activations formed a feature matrix where rows correspond to images and columns correspond to principal components, providing the input to the predictive model. We tested different PCA configurations and regression models to ensure robustness of results across settings. To predict human-derived dimensions, we trained regression models separately for each dimension, with model performance quantified as the coefficient of determination (R^2) averaged across 10-fold cross-validation. That is, for each layer and each behavioural dimension, we fit a separate regression model, training on a subset of images and evaluating prediction accuracy on held-out images, rotating through folds such that each image served as test data once. As an example, using this method, we get ten R^2 values for each dimension in each layer of the model, with each value coming from one fold of the cross-validation. These ten values are then averaged to produce a single R^2 score for that layer–dimension pair, which corresponds to one cell in the heat maps shown in the Results section. This procedure tests the ability of each layer's representation to generalize to new stimuli when predicting a specific human-derived dimension. We here report the results with 100 PCA components and linear regression. Other combinations yielded similar results and are available in the online code repository.

To summarize the hierarchical trajectory of each human-derived dimension, we defined its emergence profile as the sequence of R^2 values across layers. These profiles

were then used to compare how architectures, learning objectives, and datasets shaped the build-up of human-derived dimension information.

Statistical Testing

We used permutation testing to determine whether prediction scores were above chance and whether differences between models were statistically reliable. First, for each model and layer, we built a null distribution of prediction accuracies by repeatedly shuffling the target labels (5,000 times) and recomputing scores. This gave us a direct benchmark of what performance would look like under the null hypothesis, against which we compared the observed results. False discovery rate (FDR; Benjamini & Hochberg, 1995) correction was applied across all layers and human-derived dimensions to control for multiple comparisons.

In a second step, we leveraged these same permutations to test differences between models. By pairing the permuted scores from two models, we generated a null distribution of the performance difference (ΔR^2) for each human-derived dimension and layer. This allowed us to ask not just whether a given model performed above chance, but also whether there are differences for each human-derived dimension at each layer between two models. These paired comparisons were again corrected for multiple comparisons using FDR.

Acknowledgments

T.G and M.V were supported by the Australian Research Council (DE230100380; DP220103047).

Author Contributions

Conceptualization: Florian Burger, Tijl Grootswagers, Genevieve Quek, Manuel Varlet

Data curation: Florian Burger

Formal Analysis: Florian Burger

Software: Florian Burger

Supervision: Tijl Grootswagers, Genevieve Quek, Manuel Varlet

Methodology: Florian Burger, Tijl Grootswagers

Visualization: Florian Burger, Tijl Grootswagers, Genevieve Quek

Writing - original draft: Florian Burger

Writing - review & editing: Florian Burger, Tijl Grootswagers, Genevieve Quek, Manuel Varlet

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bracci, S., Daniels, N., & Op De Beeck, H. (2017). Task Context Overrules Object- and Category-Related Representational Content in the Human Parietal Cortex. *Cerebral Cortex*, cercor;bhw419v1. https://doi.org/10.1093/cercor/bhw419

Carlson, T., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, *13*(10), 1–1. https://doi.org/10.1167/13.10.1

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. https://doi.org/10.1038/srep27755

Contier, O., Baker, C. I., & Hebart, M. N. (2024). Distributed representations of behaviour-derived object dimensions in the human visual system. *Nature Human Behaviour*, *8*(11), 2179–2193. https://doi.org/10.1038/s41562-024-01980-y

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N.

(2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (No. arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Guclu, U., & Van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Harel, A., Kravitz, D. J., & Baker, C. I. (2014). Task context impacts visual object processing differentially across the cortex. *Proceedings of the National Academy of Sciences*, *111*(10). https://doi.org/10.1073/pnas.1312567111

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, *14*(10), e0223792. https://doi.org/10.1371/journal.pone.0223792

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974–983. https://doi.org/10.1038/s41593-019-0392-5

Konkle, T., & Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and Object Size. *Journal of Neuroscience*, *33*(25), 10235–10242. https://doi.org/10.1523/JNEUROSCI.0983-13.2013

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). *CORnet: Modeling the Neural Mechanisms of Core Object Recognition*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/408385

Mahner, F. P., Muttenthaler, L., Güçlü, U., & Hebart, M. N. (2025). *Dimensions underlying the representational alignment of deep neural networks with humans* (No. arXiv:2406.19087). arXiv. https://doi.org/10.48550/arXiv.2406.19087

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*.

Robinson, A. K., Quek, G. L., & Carlson, T. A. (2023). Visual Representations: Insights from Neural Decoding. *Annual Review of Vision Science*, *9*(1), 313–335. https://doi.org/10.1146/annurev-vision-100120-025301

Schmidt, F., Hebart, M. N., Schmid, A. C., & Fleming, R. W. (2025). Core dimensions of human material perception. *Proceedings of the National Academy of Sciences*, *122*(10), e2417202122. https://doi.org/10.1073/pnas.2417202122

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., … Bojanowski, P. (2025). *DINOv3* (No. arXiv:2508.10104). arXiv. https://doi.org/10.48550/arXiv.2508.10104

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Computational Biology*, *16*(10), e1008215. https://doi.org/10.1371/journal.pcbi.1008215

Teichmann, L., Hebart, M. N., & Baker, C. I. (2023). *Dynamic representation of multidimensional object properties in the human brain*. Neuroscience. https://doi.org/10.1101/2023.09.08.556679

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

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118. https://doi.org/10.1073/pnas.2014196118